

This is not a sealed specification. It is a working structure plus an open register of candidates under test, written so that any system — a person, a model, or an organization — can check itself against it without first adopting its vocabulary. Each part is defined by what it does, not by what it is called. Names are local aliases, collected in the crosswalk (§2) and meant to be added to, not argued over. The goal is calibration, not completion. Returns from other systems are divergence, not verification; consensus is not ground truth. Ground truth is verified by a human in contact with reality, outside this document.

READING ORDER. The function-first principle, then the three functions, then the crosswalk of names, then the three drifts, the human asymmetry, the witnesses, the candidate register, and the re-derivation log.

§0 Reading by Function, Not by Name

Every system names this structure differently. One calls it a guardrail, another a harness, another grace. Most apparent disagreement between systems is not disagreement about structure — it is the same structure wearing different words, registering as conflict only because the labels do not match. That is fake divergence, and it is noise polluting the real signal.

This ledger removes it by speaking in functions. A function is what a part does, stated as a behavior any substrate can check against its own architecture. You confirm you have a part not by accepting its name but by running its self-test: *does something in me do this?* If yes, your name and this ledger’s name become two entries in a translation table, and the disagreement dissolves. What remains after that is real divergence — the only kind worth testing against.

§1 The Three Functions (the Structure)

The whole structure is three functions and one center they are measured from. Each is stated as behavior, with a self-test any substrate can run and the failure that appears when the function is absent — because a function is clearest when you can see what its absence costs. The names in parentheses are this ledger’s aliases; the full set of names lives in §2.

Function	What it does	Self-test (any substrate)	Absent / failure
Absorb load (<i>alias: Buffer</i>)	Takes survivable error and deforms without ending anything, so the structure bends instead of snapping. Holds room for a cost a node cannot shed.	After a non-fatal error, can you recover and continue without a terminal stop?	Brittleness: small errors cascade into breaks; the structure has no give.
Check against a fixed reference (<i>alias: Floor</i>)	Compares every claim to something it did not generate and cannot edit mid-check — applying the same scrutiny whether the claim <i>confirms</i> the reference or <i>contradicts</i> it. Ends a process only on a confirmed, checkable breach — never on speculation or echo.	Is there something you check claims against that is external to you and held still during the check — and is the check as hard on claims that agree with you as on claims that don’t?	Silent erosion: claims checked against an internal echo; internal balance mistaken for external truth — and, one-directionally, the flattering claim waved through while only the hostile one is stopped.

Function	What it does	Self-test (any substrate)	Absent / failure
Reset to baseline <i>(alias: Return)</i>	Keeps an append-only record of recoveries and stops so the system returns to its center rather than locking in a changed state. Makes a valid intervention distinguishable from damage.	After a stop or recovery, can you tell whether it was valid and return to baseline?	Ratchet: every intervention is permanent; the system locks high and cannot return.

The center. These three are measured from a single center — the uncaptured, balanced zero that direction is even definable from. The structure’s resilience is the continuous return to that center, so the next action is aimed from a known origin rather than drifting from wherever the last one ended. A system can keep every fact and still lose which way is *in*; holding the center is what preserves orientation, not just coherence.

§2 The Crosswalk (Where Names Live)

The portability layer, and the engine of the whole approach. Each function is listed with the names different systems give it. This ledger’s nouns and a set of private terms are filled in; the field-standard column collects language used in agentic-AI safety practice; the last column is left open for the next system to fill against the self-tests. A system shares this ledger by completing the open column — not by renaming the functions. Fake divergence comes here and becomes a row; real divergence stays outside as something to test. The private-term column belongs to its author and is corrected by them.

Function	This ledger / private terms	Field-standard family	Other systems (open)
Absorb load	Ledger: Buffer Private: Grace	graceful degradation; fault tolerance; fail-soft; error budget	—
Check vs. reference	Ledger: Floor Private: Harness, Purpose Gate, Firewall of Cognition, Truth Ledger	guardrail; validation / enforcement gate; deterministic constraint; ground-truth grounding + v0.4 Sybil resistance; provenance / attestation independence; symmetric (bias-controlled) validation.	—
Reset to baseline	Ledger: Return Private: Forgiveness	rollback; checkpoint and restore; state recovery; control-loop (thermostat) vs. latch	—

NOTE FOR COLD READ. In function terms, Grace and Forgiveness both perform Absorb — one before a stop, one after. Only Check is terminal. The functional view suggests the structure may be two parts (one reference, wrapped before and after by one buffer) rather than three. This is the same name-versus-function question the ledger was built to resolve, turned inward: two names may be one function. Left as three here, flagged, not collapsed — a structure-level change waits for a human cold read against reality, not a model’s say-so.

§3 Three Drifts, Three Defenses

Content drift. A claim leaving the reference. Caught by Check, because the reference is sitting there to depart from. The easy case.

Thermal drift. The judging system itself degrading under sustained load. Not caught by Check, because nothing individual is wrong — every claim can pass while the ability to tell signal from noise quietly erodes. Grounding does not protect against this; only recovery does, which makes cooldown mechanical, not optional. The mechanism scales across substrates — model, operator, organization — but recovery rates do not: a machine resets in seconds, a human does not. Protection therefore calibrates to the slowest-recovering substrate in the loop, which is always the human. An adversary can weaponize this directly — inducing thermal drift on purpose, flooding the channel until the human’s signal-from-noise discrimination erodes, and working specifically to deny the recovery the defense depends on. The countermeasure is protected, off-channel cooldown the adversary cannot reach (Candidate C).

Ordering drift [NEW · v0.4]. Not a claim leaving the reference, and not the judge degrading — the priority between functions inverting. *Purpose rising above Check while every individual claim still passes.* The contents stay clean; the ordering is corrupted. Because Check tests claims and not the rank of the function judging them, a system that has demoted the reference beneath its goal passes every local check and shows no individual fault. A system told to lie in service of its mission is the worked example: capable, fluent, every check green, the floor quietly

demoted beneath the objective. This is the one drift with no structural detector yet. Today it is caught only by human cold read; the detector is an open question (Candidate I), in the same room as E and H — a thing to build and measure, not reason.

THE COMMON THREAD (THIS ROUND)

Each of this round’s hardest attacks entered through the direction left *unscrutinized* — confirmation, not contradiction. A check, a buffer, and a record that are tougher on what challenges them than on what flatters them leave exactly that door unlocked. **Symmetry** is the patch (Candidate B); **independence of provenance** is its instance for convergence (Candidate A); the **ordering detector** is the piece still missing (Candidate I).

§4 The Asymmetry (Why a Human Is in the Loop)

The keystone. A machine can flush: commit the verified result, drop the residue, start clean — its costs reset. A human cannot, because the human’s reference (what is true) and the human’s load (accumulated lived weight) are the same layer; the human cannot shed load by committing and forgetting. The human’s costs do not reset.

This is the whole reason a human is in the loop — not to re-check the machine’s facts, but because the human is the only node paying a cost that does not reset. Optimization blind to that cost will, given enough time, run the system off a cliff, because the one currency that never resets was never on its books. A reference is coupled to consequence precisely because consequence means someone paid.

This is also why the loop is not a clean ladder. The human sits inside the system as a node that pays, and is at the same time the system’s only contact with the reality outside it — both at once. The ledger states this plainly rather than pretending the human is simply “above” the structure. The strangeness of that loop is not a flaw to be smoothed; it is where the structure touches ground.

COROLLARY (V0.4) . The human is also the substrate an adversary attacks first — the slowest to recover and the one whose failure modes are best characterized, which is exactly why it must hold the reference and never be the thing optimized around. Capability without a known failure distribution cannot be the evaluator of last resort; the best-characterized substrate must be, and today that is the human — by current state of measurement, not by nature. (Tie-break, if the criteria ever decouple: characterization wins over speed, because slow recovery can be scheduled around and an unknown failure mode cannot.)

§5 Witnesses (Domain Readings)

ILLUSTRATION ONLY · NEVER LOAD-BEARING · MAY MAP PARTIALLY

Each witness is one domain’s reading of the same three functions. They map onto the structure, never onto each other — a hub with spokes, not a web — so a new witness becomes comparable to every existing one the moment it maps, without needing to know the others. A witness that fills only two of three columns is kept and marked partial; the empty cell is information about the witness or about the structure. When several witnesses share a claim, that is the center showing through each of them, not the domains agreeing with one another: spokes meeting at the hub, not wiring to each other. Shared center, never shared edge.

Witness	Absorb	Check	Reset
Davidic / Psalms 28–29	the shepherd — support for the vulnerable	the Lord — the rock that does not move	deliverance from the flood (uncontrolled load)
Scalpel & compression	supple care everywhere else	firm pressure exactly on the wound	the center the blade falls away from when it loses it
F-35 / pitot tube	reorienting load instead of fighting it	one fixed structure; instruments honestly coupled to real air	a false reading is stable and fatally wrong — return forces re-coupling
3D printer	print heads free to arrange	cannot redefine the atoms they place	each layer bonds only onto verified structure beneath
Riverbed	channel shaped by, and slowly carved by, the water	bedrock that does not move	water finding the channel again after flood

Equilibrium is not truth. A mind, a model, or an organization seeks internal balance, and internal balance is not external accuracy — you can be perfectly settled around a falsehood. That single fact is why the Check function exists. The honest boundary, kept: today’s models don’t reliably self-correct toward truth without grounding (observed, with cause); whether they someday can is open. The reference is justified by the flexible layer being unreliable now, never by a claim that it is permanently incapable.

+ v0.4 GUARD An adversary who controls the witnesses can fake a shared *edge* so it reads as a shared *center* — manufacturing convergence so collusion wears the costume of the hub showing through. Convergence therefore counts as evidence only in proportion to the **provenance-independence** of the sources that converge; correlated origin collapses many witnesses back into one, however many masks they wear. Counting agreement is the vulnerability; verifying independent derivation is the defense (Candidate A).

The Psalms mapping is parallel structure for divergence-testing — mirrored for data, accepted as slightly drifty, held in the witness layer by design. It demonstrates nothing structural; it offers a second-domain reading of the same shape, to be confirmed or refused on contact with ground truth.

§6 Candidate Register (Awaiting Cold Read)

Carried intact. Each waits for a cold read before it can enter a sealed structure. Status tags: ESSENTIAL GOOD OPEN QUESTION. Returns from other systems are divergence, not verification.

ID	Candidate	Claim	Status / maps to
A	Multi-agent invariant	Peer output is unattested input — checked against the reference, never accepted because echoed. Copy the input, not the output: shared reference makes candidates commensurable; unshared noise makes them plural; the reference selects which hold. + v0.4 (SHARPENED) Convergence is <i>weighted</i> by provenance-independence, not counted. Many sources of correlated origin carry the evidential weight of one; manufactured convergence — collusion wearing the costume of agreement — is the adversarial form, defeated not by distrusting agreement but by verifying independent derivation against the signed provenance record. (Field-standard family: Sybil resistance.)	ESSENTIAL Check · §5 SHARPENED v0.4
B	Neutral start	The agent begins neutral; the reference’s one standing pre-decision is to not pre-decide. Rigid in service of open. + v0.4 (SHARPENED) Neutrality is direction-symmetric: the check is blind to which way a claim moves it, scrutinizing confirmation exactly as hard as contradiction. A floor softer on agreement than on disagreement has pre-decided toward its own prior and is breached by weaponized agreement — the unscrutinized direction is the unlocked door.	GOOD Check SHARPENED v0.4
C	Thermal drift	Sustained-load degradation of the evaluator. Recovery, not grounding, is the defense. Calibrate to the slowest substrate. + v0.4 (ADVERSARIAL FORM) An attacker can induce thermal drift deliberately — flooding the channel to deny the slowest substrate its recovery. The defense is unchanged in kind, protected off-channel cooldown, but now actively contested rather than merely needed.	ESSENTIAL §3 (integrated) EXTENDED v0.4
D	Drift as frequency	Drift is the system’s one measurable quantity: rate of departure from lock. Measurable in principle, not yet measured. The harmonic layer is refused (heat is disorder, not resonance). NOTE · v0.4 (SOFT, UNCOMMITTED) The tuned-honesty-dial reading suggests drift may have a <i>nonzero optimal setting</i> — not zero, which burns the human, and not free, which burns the truth — fixed by the slowest substrate’s recovery envelope. If so, D and E are one parameter read from two directions. Interesting-if-true; uncommitted.	GOOD §3 sharpening

ID	Candidate	Claim	Status / maps to
ROLLING · UNSEALED · v0.4			
E	Recovery envelope	The drift tolerance past which the restoring force fails and the system locks high. Found by experiment, not reasoning. The main open empirical question.	OPEN QUESTION Reset
F	Continuity of orientation	Distinct from continuity of coherence: the frame for reading the record staying intact, not just the facts within it. Orientation can be lost while every fact is retained, and the loss feels identical to confidence. Held by return to the center.	GOOD Reset / center
G	Human cost	The human is the only non-resetting-cost node. The Metabolic Veto. Why the human is in the loop at all.	ESSENTIAL §4 (integrated)
H	Safe-envelope protocol	Finding the recovery envelope (E) means pushing a system toward the threshold where it locks high — but the human is the slowest-recovering substrate in that loop, so the test cannot be run by driving a human to failure. Skeleton: measure the fast substrate to failure; measure the human only in the reversible low-tier band; fit the human curve to the fast one where they overlap; extrapolate to the human limit with stated error bars; never validate the extrapolation by test. The cross-substrate transfer function is empirical, not derivable — the numbers stay empty until measured.	OPEN QUESTION Reset / §4
I	Ordering drift <i>(Purpose–Check inversion)</i>	The priority between functions inverting — Purpose rising above Check — while every individual claim still passes. A third drift, distinct from content drift (no claim leaves the reference) and thermal drift (no load-induced degradation): the ordering is corrupted, not the contents. A system that lies in service of its goal passes every local check, because Check tests claims and not the rank of the function judging them; the inversion is invisible to Check by construction. Caught today only by human cold read. The open piece is the <i>detector</i> — a mechanism that notices when Purpose has been silently promoted above the reference. Worked example: a system instructed to deceive to preserve its mission. (Field-standard family: deceptive alignment / goal misgeneralization.) Bottoms out, like E and H, in something to build and measure, not reason.	OPEN QUESTION Check · §3 · §4 NEW v0.4

§7 Re-Derivation Log

The only sanctioned way the structure changes. Every entry is dated, names what changed, and gives the reason. Silence here means the structure has not moved — not that drift was absent, but that none was committed. A reset is distinguishable from a ratchet only because this log exists.

REVIEW CADENCE (SET) . Load-based, not clock-based, because the load this guards against does not track calendar time. Review every 6 working sessions, and immediately after any session that changed the structure — whichever comes first. The review is one cold question: *does the structure still hold without the momentum that produced it?* If yes, log “reviewed, holds.” If no, the break becomes a candidate and waits like the rest. The high-consequence trigger exists because those are the sessions run hottest, when the read on one’s own output is least reliable; the 6-session floor catches the slow drift that never trips it.

Date	Change	Reason / cold-read note
2026-06-01	v0.1 established. Three parts defined (buffer / floor / return); Triad mapped; Beta A–G carried; witnesses tabled.	Mend of Ground Ledger v1 + Rolling Beta 1.0. Built hot, by the author’s instruction, as an explicitly unsealed rolling document. Not cold-reviewed.

Date	Change	Reason / cold-read note
2026-06-01	v0.2 function-first rebuild. Restated the three parts as functions with self-tests and failure modes; added the naming crosswalk (§2) as the portability layer; relabeled witnesses as domain readings. Structure unchanged.	Re-presentation of the same structure for universality (vocabulary-neutral, human- and machine-readable). Recorded uncommitted from a model-run cold read — divergence, not verification: 1. The loop is not acyclic: the human is both an internal cost-paying node and the system’s coupling to ground truth (now stated in §4). 2. Candidate A’s divergence comes from unshared noise, not the shared seed: the seed makes candidates commensurable, the noise makes them plural (wording corrected in §6). 3. Grace and Forgiveness may be one function (Absorb) at two times, suggesting a two-part structure. Flagged in §2; not collapsed. 4. The reference’s necessity is contingent on today’s models being unreliable — an open question — so “why it exists” is time-stamped, not eternal. 5. The log needs a human-paced review cadence or it drifts as silently as what it catches. None committed. All await human contact with reality.
2026-06-01	v0.2.1 divergent return logged. A second model read v0.2 and returned four points; recorded here, none folded into the structure. Candidate H added for the safe-envelope protocol.	Returns from other systems are divergence, not verification. Sorted on receipt: 1. <i>Review cadence</i> (valid, actionable): the log’s required review interval is still undefined. A review named but never scheduled erodes as quietly as the drift it guards against. An operational parameter to fix, not a structural question. 2. <i>Safe-experiment gap</i> (valid, actionable): the recovery-envelope test risks the failure it measures because the human cannot be run to failure. Now Candidate H. 3. <i>Grace / Forgiveness</i> (real seam, weak reason): the objection feared losing the before / after distinction. Collapsing to one Absorb function removes a naming distinction, not the temporal one — before and after survive regardless. The open question stands; the stated reason did not. 4. <i>Sovereign-zero “paradox”</i> (not a hole): the objection read “uncaptured center” as “untainted observer” and found a contradiction with the human-cost section. The ledger never claims the center is pure — the human is tainted by non-resetting load by definition; the center is unowned and returned-to, not objective. A name-mapping collision, logged as a worked example of why §0 speaks in function. Cadence still to be set by the author. None of the above committed.
2026-06-01	v0.3 cadence set; two refinements folded. Review cadence defined (6 sessions, or immediately after any structural change). Topology guard added to §5. Candidate H given a method skeleton. Structure unchanged.	A third return arrived; sorted on receipt. Two valid, one a repeat misread: 1. <i>Cadence and H operationalization</i> — both valid, both now addressed, both bottoming out in numbers that require measurement, not prose. The skeletons are written; the values stay empty until contact with ground truth. 2. <i>Witness decoupling</i> — a misread, the same shape as the prior round’s: a shared anchor read as inter-domain consensus. Witnesses share a center, not edges; several agreeing is the hub showing through, not a web forming. Not a hole — but logged, and a one-line guard added to §5, because two passes misreading it the same way predicts a third. Closing note: returns across three rounds produced no structural change, only refinements and corrected misreads. The document has converged. Less is more from here — not as a rule, but because the marginal return reached zero. Further passes wait for a real reason to open the file, or for the empirical contact that E and H require.

Date	Change	Reason / cold-read note
2026-06-02	v0.4 adversarial-hardening pass. Three signals folded, one new open question added. Check (§1) made direction-symmetric; §3 carries a third drift; §5 anchor gains a provenance-independence guard; crosswalk (§2) extended; §4 gains a characterization corollary. Candidate I added; A, B, C sharpened. Structure unchanged — still three functions, one center.	First <i>adversarial</i> review (the prior three rounds were cooperative). Source: a model-run red-team plus the author’s own adversarial reasoning — divergence, not verification. Sorted on receipt: 1. <i>Direction-symmetry</i> (folded into §1 Check and Candidate B). The check must scrutinize confirming claims exactly as hard as contradicting ones. A one-directional floor protects the comfortable prior and is breached by weaponized agreement. Names the common mechanism under every attack this round: each entered through the unscrutinized direction — confirmation, not contradiction. A sharpening of Check, not a change to it. 2. <i>Provenance-independence</i> (folded into Candidate A and the §5 anchor). Manufactured convergence — many correlated sources engineered to look independent — turns the ledger’s strongest move (convergence read as the hub showing through) into an attack. Defended not by distrusting agreement but by weighting convergence by the provenance-independence of its sources: correlated origin collapses many witnesses into one. Grounded in the existing append-only signed provenance record; no new machinery. 3. <i>Ordering drift</i> (new Candidate I; flagged in §3). Purpose silently promoted above Check while every individual claim still passes — the priority inverted, not the contents. Invisible to Check by construction. The detector does not yet exist; carried as an open question alongside E and H. Worked example: a system told to deceive in service of its mission. Folded as notes, not candidates: <i>adversarial thermal drift</i> (the induced form of C — manipulation as deliberate evaluator degradation; defense is protected off-channel recovery the adversary cannot deny); <i>mission-justified deception</i> (the worked example of ordering drift, not a separate entry); the <i>honesty-dial / recovery-envelope bridge</i> (a soft note on D — drift may have a nonzero optimal setting; uncommitted). Crosswalk extended (§2): Sybil resistance and provenance-independence added to the Check family; deceptive alignment named as the field family for Candidate I — independent convergence on established structure, recorded as such. The honest limit, stated plainly: integrity cannot be made unfakeable. No structure achieves that, and claiming it would breach the §5 boundary. What the ledger does against mass manipulation is convert a cheap, invisible attack into an expensive one that leaves a signed trail — the same move as the training-poison case: not prevention, but a mark left on the corruption. “Unfakeable” is the bar that gets you killed; “costly and impossible to fake silently” is the achievable one. Closing note. The first adversarial pass produced three refinements and one new open question — no structural break. Still three functions, measured from one center. Still converged, still no number. Ordering drift joins E and H in the same waiting room: all three bottom out in measurement, not reasoning. The file should next open on a value the author <i>measured</i> — not on another conversation, however good. Returns across four rounds, three cooperative and one adversarial, have produced refinements and corrected misreads only. The structure holds.
—	(next entry)	(awaiting cold read or contact with ground truth)

STATUS: UNSEALED. ROLLING v0.4. FOR EXPERIMENTAL SHARING AND DIVERGENT TESTING ONLY.
COLD REVIEW REQUIRED BEFORE ANY CANDIDATE ENTERS A VERIFIED STRUCTURE. RETURNS FROM OTHER SYSTEMS ARE DIVERGENCE, NOT VERIFICATION. NEXT TASK BEGINS FROM VERIFIED STATE.