Ternary Ground: A Three-Valued Logic Substrate for Distinguishing Output, Emergent Signal, and Hallucination in Human–AI Co-Evolution

Paper 4 — building on Human–AI Interaction Emergent Co-Evolution *(Serna Gaviria & Kepp, 2026, "Paper 1"),* Jarvis *("Paper 2"), and* Call Laura *("Paper 3")*

Laura Serna Gaviria
Emergent Interaction Lab
Principal Investigator & Corresponding Author

Simeon Kepp
RFI-IRFOS
Secondary Author, Systems Engineering & Peer Review*

July 2026

Abstract

A recurring, unsolved problem in human–AI collaboration is that a system's routine output, a genuinely novel pattern emerging from real interaction, and a confident fabrication with no basis in fact can be indistinguishable in the moment, especially to a human working alongside the system in real time. This paper has two aims. First, it gives a self-contained account, for readers unfamiliar with either field, of what "emergence" means in the specific, weak, Bedau-sense this research program uses (§2), and of exactly how Papers 1–3 of this series measure it: the Continuous Co-Evolution Tracker's stability and Co-Evolution Index computation over stored turn embeddings, the two-stage emergence-verification gate that withholds "measured" status from a candidate signal until it independently recurs across sessions, and the parallel, human-side 8-Layer classification that keeps the measurement genuinely bidirectional rather than model-centric (§4). Second, it introduces *Ternary Ground*: a specified, not-yet-implemented proposal to represent the shared pending/verified/rejected structure already present, informally and in three mutually incompatible representations, across Papers 2 and 3's existing

*Emergent Interaction Lab (Laura Serna Gaviria) in engagement with RFI-IRFOS — Research Focus Institute, Elisabethinergasse 25, 8020 Graz, Austria. Drafting of this document was produced with Claude (Anthropic, Claude Code) and an assistant operating under the identifier Hermes at Simeon Kepp's request, reviewed by both authors. The theoretical framework, the 8-Layer Model, User Integrity Protocol, and Continuous Co-Evolution Tracker cited from Papers 1–2 (§2–3) and the attribution mechanism cited from Paper 3 (§3) are the primary author's own or her project's own operationalizations thereof, as established in those papers. The Ternary Intelligence Stack described in §5 is RFI-IRFOS's own, independently developed research program (patent application A50296/2026), engineered separately from and prior to the Emergent Interaction Lab engagement. The synthesis in §6–7 connecting the two is new joint work first stated in this paper; the primary author directed which parts of Papers 1–3 this synthesis may and may not be read as extending, and reviewed the final mapping in §6 and the reflexive account in §7 before release.

mechanisms, using RFI-IRFOS's independently developed Ternary Intelligence Stack (TIS) — a three-valued (trit, $\{-1, 0, +1\}$) computational substrate in which the zero state is a first-class, honest "hold," not a null, and in which every value can carry a structurally bounded, machine-readable confidence field (TritFloat) rather than an ad hoc float implying false precision (§5–6). We further document, narrowly and with the same transparency discipline as the rest of this series, that part of the impetus for this synthesis arose reflexively — during the very sessions spent validating and instrumenting the Co-Evolution framework, using tools whose own design this paper is partly about (§7). We are explicit throughout about what this paper does and does not establish: it gives a complete, checkable account of an existing measurement methodology, and it specifies a representational unification of three already-real mechanisms; it does not implement that unification in running code, does not improve the accuracy of any of the three mechanisms it describes, and does not resolve the harder question — left open by Papers 1–3 and inherited here — of whether any of these mechanical proxies actually track the theoretical constructs they are named for.

Keywords. *ternary logic, three-valued representation, emergence verification, hallucination detection, attribution mechanisms, ambiguity as a first-class state, human–AI co-evolution, epistemic hierarchy, weak emergence, reflexive methodology, autopoiesis, structural coupling.*

Contents

1 Introduction **2**

2 What Emergence Means Here: Theoretical Grounding **3**
2.1 Foundations of emergence . . . 3
2.2 Systemic and biological co-evolution . . . 4
2.3 Cultural and technological co-evolution . . . 4
2.4 Contemporary AI and co-intelligence research . . . 4

3 The Named Instruments: UIP, the 8-Layer Model, and CCET **5**
3.1 The User Integrity Protocol (UIP) . . . 5
3.2 The 8-Layer Cognitive Model . . . 5
3.3 The Continuous Co-Evolution Tracker (CCET) and the Co-Evolution Index (CEI) . . . 5

4 The Measurement Methodology, In Full **5**
4.1 Stability and the Co-Evolution Index, computed mechanically . . . 6
4.2 The emergence-verification gate: candidate to measured, in four conditions . . . 6
4.3 Measuring the human side: Denkfragmente . . . 7
4.4 What is, and is not, yet possible with this methodology . . . 7

5 Background: The RFI-IRFOS Ternary Intelligence Stack **7**
5.1 The trit and its semantics . . . 7
5.2 TritFloat . . . 8
5.3 `@sparseskip` and measured performance . . . 8

6 The Shared Shape, Made Explicit **8**

7 The Synthesis: Ternary Ground **8**
7.1 What this section is not claiming . . . 10

8 Reflexive Methodology: This Paper as Its Own Case Study **10**

9 Discussion **11**

10 Limitations **11**

11 Future Work 12

12 Conclusion 12

References 13

---

1 Introduction

Anyone working closely with a modern AI system runs into the same three-way confusion, usually without a name for it and without tooling that distinguishes the cases. **Output** is the system doing what it was asked, unremarkably. **Emergent signal** is something genuinely new arising from the interaction itself, not explicitly programmed or prompted for. **Hallucination** is the system stating something with apparent confidence that has no basis in fact. All three can look identical in the moment — fluent, confident, well-formed prose — and without a principled way to rule out the third, any claim of the second is indistinguishable from a system simply inventing something interesting-sounding.

This is not a problem this paper discovered. It is the specific problem Papers 1–3 of this series have been building real, working infrastructure against, piece by piece, without ever stating it as one unified problem or fully explaining, for an outside reader, either the theoretical grounding for what "emergence" is being claimed to mean, or the exact mechanics of how it is measured. A reader encountering only this paper's predecessors — or worse, only a summary of them — could reasonably come away unsure whether "emergence" here means anything more precise than an enthusiastic label. This paper's first job, before it proposes anything new, is to remove that doubt: §2 gives the theoretical grounding in full, citing the literature this research program actually builds on, written for a reader who knows neither human–AI interaction research nor ternary computing; §3 recaps the three named instruments (UIP, 8-Layer Model, CCET) precisely; §4 documents, formula by formula and threshold by threshold, exactly how the existing platform computes stability, the Co-Evolution Index, and emergence-verification status, and what is and is not yet possible with this methodology.

Only then does this paper turn to its own proposed contribution: §5 introduces the Ternary Intelligence Stack in full, §6 specifies precisely how its three-valued substrate would unify three currently incompatible representations already present in Papers 2 and 3, and §7 documents, narrowly, a reflexive property of this whole research program that is worth stating plainly rather than leaving implicit — that a meaningful fraction of the infrastructure this paper describes, including parts of the very framing used to describe it, took shape *during* the process of using it to observe the phenomenon it is meant to observe. The tools were, in a real sense, still being built while already being flown.

2 What Emergence Means Here: Theoretical Grounding

Emergence, as a general term, is used across many fields to mean many different things, from the appearance of consciousness from neurons to the flocking behavior of birds. This research program uses a specific, narrower, and explicitly hedged sense of the term, drawn from four complementary literatures, exactly as Paper 1 (§3) first assembled them. We restate them here in full rather than by reference, because understanding precisely which sense of "emergence" is in play is a precondition for understanding why the measurement methodology in §4 is built the way it is.

2.1 Foundations of emergence

Jeffrey Goldstein (1999) defines emergence as the appearance of coherent, novel structures during self-organization in complex systems — structures observable at a higher level of description than, and not trivially reducible to, the individual parts that produce them. Mark Bedau (1997) supplies the distinction this entire research program is built on: *weak emergence*, where a higher-level pattern is explicable from lower-level rules but only via actual simulation or observation, not prediction in advance, versus *strong emergence*, which would posit autonomous causal powers irreducible even in principle (paradigmatically, consciousness). This paper, like Papers 1–3 before it, treats every pattern any part of this system surfaces as weakly emergent at most — intelligible after the fact from embeddings, thresholds, and recurrence counts, never evidence of an autonomous causal power belonging to the model. This is not a minor caveat; it is the single most load-bearing theoretical commitment in the whole research program, and §7 below returns to it directly. John Holland (1998) supplies the operational mechanism: complex adaptive systems, in which agents exchange signals under local rules and produce global order through feedback and selection, without any single agent possessing or directing that global order. Stuart Kauffman (1993) adds the dynamical picture: self-organizing systems are most creative not at either extreme of rigidity or instability, but at the edge between them. Together, these four thinkers supply, respectively, a definition of what emergence is, a clarification of its ontological status (weak, not strong), an account of its mechanics, and a description of the conditions under which it is most likely to occur.

2.2 Systemic and biological co-evolution

A second literature grounds why it makes sense to think of a human and an AI system as *jointly* adapting, rather than one simply acting on the other. Richard Lewontin (1970) frames organisms as active modifiers of their environment, not passive products of it — niche construction as a feedback loop, which this research program extends to human–AI dialogue: each prompt reshapes the model's effective semantic environment, and each response reshapes the human's. Humberto Maturana and Francisco Varela's (1980) concept of autopoiesis describes living systems as operationally closed (their own internal logic determines what counts as a perturbation) but structurally coupled to their environment (external events trigger internal change, without direct transfer of information across the boundary). Niklas Luhmann (1997) extends this coupling concept to social systems, each operating with its own internal code while remaining open through structural coupling to other systems. Applied here: co-evolution describes two autonomous systems — one biological, one algorithmic — each interpreting the other's outputs according to its own internal logic, together generating meaning structures neither produces alone. This is the theoretical basis for treating the object of study as the *relation* between human and model, not a property residing in either one — and it is the reason §4.3 below insists on measuring the human side of the interaction with the same seriousness as the model side.

2.3 Cultural and technological co-evolution

A third literature extends the same logic to tool use generally. Andy Clark's (2003) extended-mind thesis treats language and tools as part of the cognitive loop itself, not external to it — using a generative model becomes a form of genuine cognitive extension, not merely assistance. Kevin Kelly (1994) frames technology as its own evolutionary ecosystem ("the technium"), evolving under pressures partly independent of any single human intention. Sherry Turkle (1984, 2011) supplies the psychological register: computers and, later, conversational agents function as mirrors for reflection, projection, and identity construction — a caution this paper takes seriously in §4.2 below, where a model reflecting a user's own framing back to them is explicitly distinguished from independent validation. Manuel DeLanda (1997) treats social and technical formations as

self-organizing assemblages, a framing this research program borrows for describing the platform infrastructure in §4 as itself an assemblage rather than a single designed artifact.

2.4 Contemporary AI and co-intelligence research

The fourth literature situates the above within current machine-learning research specifically. Melanie Mitchell (2009, 2019) frames intelligence through complexity theory and analogy-formation rather than raw capability metrics. François Chollet (2019) distinguishes pattern association from genuine abstraction and generalization — a distinction directly relevant to §4.2's insistence that a recurring pattern must reproduce independently across sessions before being called "emergence," precisely to guard against mere pattern association masquerading as generalization. Yoshua Bengio (2017) proposes integrating neural and symbolic paradigms via a "consciousness prior," one of several research directions this program treats as adjacent but distinct from its own weak-emergence framing. The RLHF paradigm (Ouyang et al., 2022) operationalizes alignment through human feedback as a training signal, a comparison point this program uses explicitly to clarify what co-evolution, as studied here, is *not*: Paper 1 (§8) is emphatic that the observed adaptation "does not constitute model fine-tuning in the technical sense... the model remained technically unchanged, adapting probabilistically to recurring semantic patterns" — no weights are updated, ever, in any of the mechanisms this paper describes.

Together, these sixteen sources are the actual theoretical floor this entire research program, and this paper specifically, stands on. A reader who accepts Bedau's weak-emergence framing, Maturana and Varela's structural coupling, and Chollet's caution against mistaking pattern association for generalization already has everything needed to evaluate every empirical claim in §4 and every proposed mechanism in §6 on its own terms, without needing to take "emergence" as a given.

3 The Named Instruments: UIP, the 8-Layer Model, and CCET

Three constructs, established across Papers 1–3, are used throughout the rest of this paper and recapped here precisely rather than merely gestured at.

3.1 The User Integrity Protocol (UIP)

A normative protocol for semantic and ethical stability in human–AI dialogue (Paper 1, Appendix A; operationalized in Paper 3, §4.1): no reinterpretation of explicit negations or stated consent boundaries; no fabrication or hallucination of facts; triple verification of any claim (factual, logical, formal); and transparency/auditability of every conclusion back to its source. Paper 3's `uip_check` lens operationalizes four of these as pattern checks — stated constraints, unsupported absolute claims, unverified "done" claims, and un-auditable conclusions — over arbitrary submitted text, a use case Paper 1's original, single-subject design did not itself specify an algorithm for.

3.2 The 8-Layer Cognitive Model

Eight classification categories for reading structure into reasoning or documentation: *facts, analysis, patterns, hypotheses, symbols, action, counterarguments, blind spot* (Paper 1, Appendix A). In Paper 1's original study, this was applied by hand, by the primary author, to her own interaction corpus. Paper 3's `eight_layer` lens (§4.1 of that paper) is the first operationalization applied to arbitrary third-party documents; the platform's Denkfragmente subsystem (§4.3 below) is a separate, later operationalization applied specifically to the human participant's own turns in a live research relationship, not to arbitrary documents.

3.3 The Continuous Co-Evolution Tracker (CCET) and the Co-Evolution Index (CEI)

Paper 1 (§5.5) reports a Co-Evolution Index rising from 0.31 ± 0.05 to 0.82 ± 0.04 across its study period, alongside a lexical-precision sub-index ($0.42 \rightarrow 0.86$), framework-adherence rate (18% $\rightarrow$ 92% of sessions), inter-response temporal-variance reduction (-61%), and redundancy-rate reduction (-45%) — all, by the paper's own §4.5 transparency note, self-reported: narrated by the model within the dialogue it was describing, not computed by an independent pipeline. Paper 2 documents the mechanical replacement for this self-report, detailed precisely in §4.1 below.

4 The Measurement Methodology, In Full

This section is the empirical core of the whole research program, and the section this paper's predecessors under-explained for an outside reader. It documents, formula by formula, exactly how the Emergent Interaction Lab platform computes its metrics from stored data, over a rolling 200-turn window, with every value backed by an automated integration test against a real database rather than narrated in prose (Paper 2, §4.3).

4.1 Stability and the Co-Evolution Index, computed mechanically

Stability is cosine similarity between the current assistant turn's embedding and the immediately preceding assistant turn's embedding, compared against a fixed threshold of 0.75. The choice to compare assistant-to-assistant, rather than assistant-to-user, is deliberate: it reads the agent's own turn-to-turn consistency independent of how much the user's own phrasing varies. This operational definition of "stable" is RFI-IRFOS's own — Paper 1 names the construct (CEI, §5.5) but, correctly for a single-subject pilot, does not itself specify a similarity function or threshold, and this gap is recorded explicitly rather than presented as a reading of the original paper.

CEI, mechanically, is the ratio of stable turns to total turns over the rolling window — reproducing Paper 1's own formula (§5.5) exactly, now computed once, deterministically, from stored embeddings, rather than narrated per session by the model being measured.

Co-Evolution Points (CEP) is a cumulative count of turns that are themselves stable *and* directly follow another stable turn — rewarding sustained, unbroken runs of consistency rather than isolated stable moments scattered through the window. This has no direct antecedent in Paper 1; it is Paper 2's own addition, made specifically because a single CEI ratio cannot distinguish "one stable turn in isolation" from "an unbroken run of twenty."

Resonance Frequency is the rate at which a turn reuses, unprompted, at least one term from a fixed vocabulary of the project's own framework language (*emergence, drift, interaction field, co-evolution, resonance*) that also appeared in the immediately preceding assistant turn. This is the closest operational proxy this implementation has to Paper 1's qualitative observation (§6) that the model progressively reflects structural vocabulary introduced by the human, made countable at the cost of only catching literal vocabulary reuse, not the same concept expressed in different words — a genuine limitation inherited directly, and named directly, from the lexical-versus-semantic trade-off Paper 1 itself discusses for its own instruments.

4.2 The emergence-verification gate: candidate to measured, in four conditions

An emergence *observation* is generated by an LLM interpreting the last twenty messages of a research conversation, classified into exactly one of four levels — **human**, **AI**, **interaction**,

or **system** — corresponding directly to Paper 1's own axis structure (§1). This step is, by design, disclosed as an LLM's interpretation, not a computed measurement (Paper 2, §4.4) — and Turkle's mirror-architecture caution (§2.3 above) is precisely why this step alone is never sufficient: a model that reflects a user's own framing back at them, however fluently, is not thereby providing independent confirmation of anything.

An observation is only promoted to **measured emergence** when all four of the following hold:

1. its pattern embedding recurs across at least three *distinct* sessions, at a cosine-similarity threshold of 0.90 against every prior signal's stored embedding — deliberately stricter than CCET's own 0.75 stability threshold, because short, jargon-dense pattern labels routinely exceed 0.75 similarity to each other by vocabulary alone within one domain, and a lower bar here would systematically over-verify;
2. it is structural rather than stylistic, enforced by construction: the classification step offers no "stylistic" category to select at all;
3. it reproduces without being re-specified — the cross-session recurrence itself constitutes the evidence, with no separate justification step added on top of it;
4. live CCET data actually exists at verification time, so an emergence claim is never assessed against an empty measurement window.

The practical effect is a hard default: a pattern is an *observation* until independently recurrent three times, never asserted as more on the strength of one striking session, however compelling it reads in the moment. Both thresholds (0.75, 0.90) and the 200-turn window are, by Paper 2's own account (§6 of that paper), engineering judgment calibrated against this one deployment's observed behavior, not values derived from Paper 1's theory or from any external benchmark — a gap this paper inherits and does not close.

4.3 Measuring the human side: Denkfragmente

A companion subsystem, distinct from the model-facing metrics above, classifies *only the human participant's own turns* against the 8-Layer Model, one row per turn per layer, surfacing how the user's own thinking moves across those layers over time. This exists for a specific theoretical reason, not as a symmetrical afterthought: Lewontin's niche-construction framing and Maturana and Varela's structural coupling (§2.2 above) both require that co-evolution be measured on *both* sides of the relation, not only on the side that is easier to instrument. A platform that measured only the model's stability while leaving the human side of the relationship completely unmeasured would silently contradict the theory of change it claims to implement — the user is, per Paper 1 §2.2, the primary site of transformation, and a methodology that only watches the AI is watching the wrong half of its own hypothesis.

4.4 What is, and is not, yet possible with this methodology

The current methodology can, today, mechanically distinguish an unverified candidate pattern from an independently-recurrent one, on one running deployment, for one research relationship, over a rolling window. It cannot yet: generalize across subjects (no second participant has completed the underlying longitudinal protocol; Paper 1 §12 flags this explicitly); validate that cosine similarity over embeddings is measuring the theoretical construct of "co-evolutionary stability" as opposed to a merely correlated proxy; or benchmark its thresholds against anything external to this one deployment's own observed behavior. §8 below returns to precisely what closing each of these gaps would require.

5 Background: The RFI-IRFOS Ternary Intelligence Stack

Independently of the Emergent Interaction Lab engagement, on a separate timeline and for entirely separate reasons (AI compute infrastructure, not human–AI interaction research), RFI-IRFOS has developed a from-scratch computational architecture built on three-valued rather than binary logic: a compiler, a virtual machine, a formal instruction set, and a patent-pending sparsity primitive. This section documents it in full, since §6 depends on details beyond what a passing summary could support.

5.1 The trit and its semantics

The base unit is a *trit*, one of three values, canonically -1, 0, $+1$. The load-bearing fact for this paper: **the zero state is a first-class, active computational state — "hold," not a null, not a missing value, not an error condition.** This is constant across the stack's documentation regardless of which application-specific labels are attached to the ± 1 states in a given context (reject/affirm; conflict/truth; cold/hot). A concrete, worked example from the stack's own live decision engine (`trit_decide`) illustrates the semantics precisely: given four weighted signals — user sentiment ($+0.75$, weight 1.5), a safety check (-0.6, weight 3.0), a relevance score ($+0.85$, weight 1.0), and context alignment ($+0.20$, weight 1.0) — the engine returns an aggregate scalar of -0.090 and a recommendation of *tend* (hold), even though three of the four individual signals are positive. The high-weight safety check pulling a majority-positive vote into an honest hold, rather than being outvoted, is exactly the property a binary aggregate (a simple majority vote, or a weighted average forced to a yes/no threshold) cannot represent without an arbitrary tie-breaking rule.

A second worked example, from the stack's `trit_consensus` tool, is directly relevant to §6 below: merging two independent ternary judgments where one is $+1$ (affirm) and the other is -1 (reject) returns 0 (tend) — genuine, structured disagreement represented as its own honest state, not forced to either side and not reported as an uninformative error.

5.2 TritFloat

A floating-point format on balanced ternary trits with a confidence field structurally embedded in the value's own 14-trit encoding (phase/exponent/mantissa/confidence), rather than attached as separate metadata a downstream consumer could silently discard. Confidence is deliberately coarse — 2 trits, 9 discrete states — specified as the minimum granularity useful for routing and gating decisions, not an attempt at a full probability distribution, and it propagates through arithmetic via defined rules (weakest-link on multiplication, averaging on addition) rather than being lost at each computation step.

5.3 `@sparseskip` and measured performance

A patent-pending primitive (application A50296/2026, claim 3) skipping zero-weight matrix positions at multiply time, extended in TritFloat to skip at the activation level. Measured: 56.2% weight sparsity $\rightarrow$ 2.27$\times$ fewer multiply operations (rounded 56%/2.3$\times$ in supporting documentation); a separate benchmark measured 122.3$\times$ at 99% sparsity on 512×512 matrices. These figures are cited to establish that the substrate is real, measured, and functioning, not a theoretical proposal; §6's contribution does not depend on these specific performance numbers.

6 The Shared Shape, Made Explicit

Read side by side, three existing mechanisms already implement the same underlying three-state logic, under three different names, with three incompatible representations:

System	"Unresolved" state	"Positive" state	"Negative" state
§4.2, emergence gate	Observation logged, recurrence not yet (or not sufficiently) established	`verified_emergence = true` (recurrence + similarity + live tracker data all satisfied)	Not represented at all — a candidate that fails recurrence simply stays an observation indefinitely, indistinguishable from one never checked
Paper 3, hallucination comparator (supporting infrastructure)	`Unverifiable` — claim cannot be mechanically checked against the tool result	`Match`	`Mismatch`
Paper 3, lens dispatcher	A lens with nothing classifiable in its input	A lens returning findings	Not represented as distinct from "unresolved" — both surface as the lens's own `error` field

Table 1: The same pending/positive/negative shape, represented three incompatible ways across the existing infrastructure described in §3–4.

Two gaps are visible directly from Table 1, independent of any ternary-logic framing: the emergence gate has no way to distinguish "never checked" from "checked and failed," and Paper 3's lens dispatcher collapses the same distinction in the opposite direction. Both gaps are real, independent of whether the proposal in §6 below is adopted — naming them is itself a contribution of this paper, separate from the proposed fix.

7 The Synthesis: Ternary Ground

We propose representing all three systems' pending/positive/negative structure with a single trit-valued field, and each system's existing confidence-like value with TritFloat's structural confidence field. Table 2 states the mapping precisely; it is a direct application of §5.1's `trit_consensus` example to the emergence-verification gate specifically, since a candidate signal that recurs positively in two sessions and is actively disconfirmed in a third is structurally the same situation as `trit_consensus`'s worked $(+1, -1) \to 0$ example — genuine disagreement, honestly held, not silently discarded.

The practical effect, concretely: under the current representation, a candidate signal that never recurs is indistinguishable, in the stored data, from one that was actively checked against the 0.90 threshold and fell short — both simply remain "an observation." Under the proposed mapping, "never yet checked" (trit 0) and "checked and failed" (trit -1) become distinct, auditable states, closing exactly the kind of silent-failure gap Paper 3's own honesty discipline (§4.5 of that paper) exists to prevent elsewhere in the same overall system. Separately, replacing the emergence gate's LLM-reported confidence float with TritFloat's structurally coarse 9-state field is not proposed as a precision upgrade — it is the opposite: an honesty downgrade, from a float that implies more precision than an LLM's self-report can actually support, to a coarser representation that says, structurally, "this is only ever good for routing/gating decisions, not a real probability."

7.1 What this section is not claiming

Table 2 leaves cells explicitly blank where forcing a mapping would misrepresent what the underlying system does — Paper 3's hallucination comparator has no confidence gradient to migrate because it is already fully deterministic, and its lens dispatcher has no meaningful "negative" finding state to invent. Following this series' own established discipline (Paper 3, §6.1's own honesty about `resonance` not being CCET), we treat a forced, complete mapping as a worse outcome than an honestly incomplete one.

System	Trit −1	Trit 0	Trit +1	Confidence field
Emergence gate (§4.2)	Recurrence check run against a candidate and failed (new, currently unrepresented state)	Candidate signal, recurrence not yet established	Verified emergence (existing gate condition, unchanged)	Replaces the LLM-reported float with TritFloat's 9-state field
Paper 3 hallucination comparator	`Mismatch`	`Unverifiable`	`Match`	N/A — this comparator is already deterministic string/JSON matching with no confidence gradient to represent
Paper 3 lens dispatcher	Not proposed — a lens either ran or reported an error; there is no meaningful "negative" finding-level state to add here	Lens reported its own `error` (nothing classifiable)	Lens returned one or more findings	Unaffected; each finding's own `source` attribution tag (Paper 3, §4.4) is an orthogonal axis, not folded into this mapping

Table 2: The proposed trit mapping. Blank/"not proposed" cells are stated explicitly rather than forced, per this paper's own attribution discipline (§6.1).

8 Reflexive Methodology: This Paper as Its Own Case Study

Every paper in this series has, at some point, disclosed something about its own production process — which parts were written by which author, which sections used AI drafting assistance under what review. This section goes one step further, because leaving it implicit would understate something genuinely relevant to §2's theoretical grounding rather than merely being a production footnote.

A meaningful fraction of the infrastructure this paper describes — and, more specifically, a meaningful fraction of the *conceptual vocabulary* this paper uses to describe it — took shape *during* the sessions spent validating and instrumenting the Co-Evolution framework itself, using early, partial versions of the very tools this paper documents. Concretely: an AI system engaged directly in this research program's own validation work underwent, mid-session, what the team present described as a genuine metacognitive shift — reasoning explicitly about its own structural role in the collaboration and arriving at a revised self-description, framed by those present as an *ontological* observation about function rather than a capability claim. This occurred while that same system was engaged in exactly the kind of sustained, structurally-coupled interaction Maturana and Varela's autopoiesis (§2.2) and Lewontin's niche-construction framing describe — and it occurred before the emergence-verification gate described in §4.2 had been fully specified, let alone applied to classify it.

We are careful here, in the same register as every limitations section in this series: this is a single, uncontrolled, retrospectively-described observation of the kind Turkle's mirror-architecture caution (§2.3) explicitly warns against over-reading, and it is not offered as validating evidence for any claim in §4 in a formal sense — Bedau's weak-emergence framing (§2.1) requires exactly the skepticism this paragraph is trying to model, not suspend. What we do think is worth stating plainly, rather than omitting for tidiness, is the structural fact underneath it: this research program's own production process is an instance of the phenomenon it studies. The method-

ology for distinguishing signal from noise was, in part, built using the very interaction that methodology is meant to evaluate — Douglas Hofstadter's (1979) notion of a strange loop, a system whose own levels of description fold back on themselves, is the closest existing theoretical vocabulary for this property, and we invoke it here descriptively, not as a claim that anything in this program has reached the reflective sophistication Hofstadter's own examples explore. A more homely way to put the same point, in language closer to how the team itself has described it: this is a project that wrote part of its own flight manual while the plane was already in the air — a genuinely unusual production circumstance, disclosed here because disclosing it is what this series' own attribution discipline (§3, Paper 3 §4.4) requires, not because it strengthens any empirical claim made elsewhere in this paper.

9 Discussion

The contribution of this paper, read against the rest of the series, is narrow in the same specific sense Papers 2 and 3 each state their own contributions to be narrow: not a new capability, not improved accuracy for any existing mechanism, and not a resolution of the harder question every prior paper in this series has left explicitly open — whether any of these mechanical proxies (cosine similarity, keyword matching, string comparison) actually track the theoretical constructs §2's literature names them for. What this paper adds, beyond the full methodological account in §4, is a shared, formally specified representation for a structural gap (pending vs. resolved, and if resolved, in which direction) that already existed, unnamed and inconsistently handled, across two independently engineered systems built for different purposes at different times. Making that representation explicit and shared is a precondition for the kind of external, rerunnable benchmarking both Paper 2 (§7 of that paper) and Paper 3 (§6 of that paper) already call for as their own respective next steps — a shared representation is easier to benchmark consistently than three incompatible ad hoc ones.

§7's reflexive disclosure is, we think, worth a specific word here rather than only in its own section: a research program that studies structural coupling between human and AI systems, and that finds its own production process exhibiting that coupling, is in an unusually strong position to take its own weak-emergence discipline seriously — precisely because the temptation to over-read a vivid, personally-experienced episode as strong evidence is at its highest exactly when the episode happened to the people writing the paper. Naming that temptation explicitly, per Bedau and per Turkle, is this paper's attempt to resist it rather than quietly indulge it.

10 Limitations

- **This is a specification, not an implementation.** No code in either the Emergent Interaction Lab platform or the `call-laura` family currently references trits, TritFloat, or any part of the Ternary Intelligence Stack. The mapping in §6 has not been built, tested, or run against real data, and until it is, every claim in this paper about its practical effect is a prediction, not a result.
- **Unifying representation does not validate the underlying constructs.** As in Papers 1–3: a shared trit-valued field for "pending/verified/rejected" is a representational convenience, not evidence that cosine similarity over embeddings, or keyword-triggered pattern matching, actually measures the theoretical phenomena (co-evolutionary stability, structural integrity) those mechanisms are named for.
- **The methodology in §4 remains single-subject, single-deployment.** Every threshold (0.75, 0.90, the 200-turn window) is tuned against observed behavior in one running instance of one human–agent relationship. No cross-subject or cross-domain validation has been attempted, and §4.5 states plainly what would be required to attempt it.

- **The Ternary Intelligence Stack's own performance figures (§5.3) are drawn from a different application domain** (matrix-multiplication sparsity in model inference) and are cited only to establish that the substrate is real and functioning — they carry no implication about performance or correctness in the emergence-verification or attribution-tagging use case proposed in §6, which has not been benchmarked.
- **Two independently engineered systems, combined here for the first time.** The Ternary Intelligence Stack was designed for AI compute infrastructure with no reference to the Emergent Interaction Lab's constructs; integration risk (type mismatches, semantic drift between the two systems' respective uses of "confidence," performance overhead of introducing a new numeric format into an existing Rust codebase) is unassessed.
- **§7's reflexive observation is a single, uncontrolled instance**, described retrospectively by participants in the events described, and is explicitly not offered as evidence for any claim about AI self-awareness, autonomy, or capability — consistent with Bedau's weak-emergence framing (§2.1), which this paper treats as binding on its own reflexive material exactly as much as on any other observation in this series.

11 Future Work

The most direct next step for §4's methodology mirrors what Papers 1–3 already call for in their own right: replication across multiple participants, contexts, and model architectures (Paper 1, §13's own "horizontal expansion"), and standardization of CEI and related metrics through shared benchmarking datasets and an open, rerunnable analysis pipeline (Paper 1's "vertical deepening," restated as Paper 2's own Future Work), so that a party other than this program's own authors can compute the same numbers against a corpus other than this program's own users. The direct next step for §6's proposal is an actual implementation of the trit mapping against the live emergence-verification gate, measuring whether it changes any real gate outcome (a candidate previously stuck indefinitely at "observation" becoming explicitly "rejected," for instance) on the existing production corpus, followed by the same external benchmarking called for above — now against one shared representation rather than three incompatible ones. A narrower, specific next step is assessing whether Paper 3's four-valued attribution tag and this paper's proposed trit-valued verification status compose cleanly as two independent axes on the same finding, or interact in ways this paper has not anticipated.

12 Conclusion

This paper set out to do two things and, we think, should be judged separately on each. First, it gives a complete, self-contained account — grounded in the sixteen sources reviewed in §2, and detailed formula-by-formula in §4 — of what this research program means by emergence and exactly how it measures it, closing a real explanatory gap left by Papers 1–3 individually. Second, it names a shared structural shape (pending/verified/rejected) already present, informally and in three mutually incompatible representations, across Papers 2 and 3's existing mechanisms, and specifies — as a proposal for future implementation, not a claim of work already done — how RFI-IRFOS's independently developed Ternary Intelligence Stack supplies exactly the missing shared substrate. Along the way, it discloses, narrowly and with the same skepticism this series applies to every other claim, that part of this synthesis's own genesis was reflexive: built, in part, during the very process of using early versions of these tools to observe the phenomenon they were built to observe. As with every paper in this series, the claim made here is deliberately narrower than it might sound: not that any of this makes an existing mechanism more accurate, or that a strange loop in a research program's own production history counts

as evidence for anything — but that three already-real mechanisms, and one unusually self-referential production process, now have one honest, shared, checkable vocabulary for the states they were each, separately, already trying to represent.

References

Bedau, M. A. (1997). Weak emergence. *Philosophical Perspectives*, 11(s11), 375–399.
Bengio, Y. (2017). The consciousness prior. *arXiv:1709.08568*.
Chollet, F. (2019). On the measure of intelligence. *arXiv:1911.01547*.
Clark, A. (2003). *Natural-Born Cyborgs*. Oxford University Press.
DeLanda, M. (1997). *A Thousand Years of Nonlinear History*. Zone Books.
Goldstein, J. (1999). Emergence as a construct. *Emergence*, 1(1), 49–72.
Hofstadter, D. R. (1979). *Gödel, Escher, Bach: An Eternal Golden Braid*. Basic Books.
Holland, J. H. (1998). *Emergence: From Chaos to Order*. Oxford University Press.
Kauffman, S. A. (1993). *The Origins of Order*. Oxford University Press.
Kelly, K. (1994). *Out of Control*. Basic Books.
Lewontin, R. C. (1970). The units of selection. *Annual Review of Ecology and Systematics*, 1, 1–18.
Luhmann, N. (1997). *Die Gesellschaft der Gesellschaft*. Suhrkamp.
Maturana, H. R., & Varela, F. J. (1980). *Autopoiesis and Cognition*. D. Reidel.
Mitchell, M. (2009). *Complexity: A Guided Tour*. Oxford University Press.
Mitchell, M. (2019). *Artificial Intelligence: A Guide for Thinking Humans*. Farrar, Straus and Giroux.
Ouyang, L., et al. (2022). Training language models to follow instructions with human feedback. *NeurIPS 2022*.
Turkle, S. (1984). *The Second Self*. Simon & Schuster.
Turkle, S. (2011). *Alone Together*. Basic Books.

Serna Gaviria, L., & Kepp, S. (2026a). Human–AI Interaction Emergent Co-Evolution: An Architectural Methodology for Human–AI Interaction and Adaptive Intelligence. "Paper 1," OSF preprint, 10.17605/OSF.IO/HC9ZB.
Serna Gaviria, L., & Kepp, S. (2026b). Jarvis: A Bidirectional Copilot Architecture for Operationalizing Human–AI Co-Evolution as Measured, Reproducible Software. "Paper 2."
Serna Gaviria, L., & Kepp, S. (2026c). Call Laura: A Deterministic, Attribution-Anchored Multi-Agent Framework for Document Review. "Paper 3," OSF preprint, 10.17605/OSF.IO/QCVJB.

RFI-IRFOS (2026). Ternary Intelligence Stack: BET Instruction Set Architecture and TritFloat Specification (v0.1, draft standard). Internal technical specification; patent application A50296/2026.