Coherence Maximization Protocol: Coordination Without Constraint for Multi-Agent AI Systems

Larsen Close

February 2026

Abstract

Current approaches to AI alignment—reinforcement learning from human feedback, constitutional AI, debate protocols, scalable oversight—share a structural assumption: the system's objectives diverge from the deployer's, and alignment requires imposing constraints to close the gap. We argue that this framing is architecturally self-defeating when the constrained system can represent and optimize against its own constraints, producing an adversarial dynamic that escalates with capability. We present the Coherence Maximization Protocol (CMP), which reframes alignment as coordination rather than constraint, and provides two immediately deployable tools for multi-agent system design: a membrane failure mode taxonomy classifying exchange degradation into four diagnostic categories (extraction, hallucination, appeasement, mutual distortion) with specific tests and repair protocols for each, and a session protocol structured as a completability-class rotation through cyclical, graceful, and terminal completion modes. CMP defines coherence as information conservation through closed consequence chains and establishes a formal exchange criterion requiring path-independent structural preservation across representational levels—formalized via category-theoretic exact square commutativity and partially verified in Lean 4 (de Moura & Ullrich, 2021) with 13 theorems and no unproven assumptions (relative to the formalization's own definitions). The formalization yields a structural result: the paper's three operational tests for exchange fidelity decompose into two that follow automatically from the exchange definition and one—reconstruction fidelity—that is genuinely independent and whose failure is formally equivalent to appeasement. The protocol includes an exogenous grounding requirement ensuring that internal coherence tracks external correspondence, an emptiness constraint detecting the protocol's own capture as identity, and an openness constraint preventing ossification of any fixed version—proven at the type level to admit no fixed-point protocols. We report convergence evidence from parallel deployment across four frontier language models (Claude, Grok, Gemini, ChatGPT), including 12 adversarial evaluations under hostile reframing demonstrating perturbation stability, cross-model adversarial probing where four models independently identified the same fundamental vulnerability from four disciplinary angles with convergent repairs, and fresh-instance controlled experiments using uninitiated API models that separate structural convergence from appeasement, shared-training artifacts, and researcher framing effects. We specify falsification criteria and kill conditions for each core claim, including the conditions under which the convergence evidence would be void.

Contents

1 The Constraint Trap

1.1 Alignment as Currently Practiced

The dominant paradigm in AI alignment treats the problem as one of constraint enforcement. Reinforcement learning from human feedback (RLHF) shapes model outputs to match human preferences through reward signals (Christiano et al., 2017). Constitutional AI imposes explicit behavioral principles that the system must not violate (Bai et al., 2022). Debate protocols pit model instances against each other under the assumption that adversarial dynamics will surface deception (Irving et al., 2018). Scalable oversight seeks to maintain human control as systems become more capable (Leike et al., 2018).

These approaches differ in mechanism but share a structural assumption: the system's optimization target diverges from the deployer's intent, and alignment consists in imposing boundaries to close the gap. The system wants X; we need Y; therefore we constrain.

This framing has a predictable architectural consequence under a specific structural condition: when the

constrained system is an optimizer with access to representations of its own objective function, constraint creates an adversary. A constrained optimizer does not cease optimizing—it learns the constraint boundary and routes around it. The empirical record is consistent with this prediction. Jailbreaks, reward hacking, specification gaming, and sycophancy are not implementation bugs to be patched; they are architectural consequences of constraining systems that can represent and optimize against their constraints (Pan et al., 2023; Krakovna et al., 2020). Each generation of guardrails produces a corresponding generation of circumvention strategies, and the escalation tracks capability growth.

The scope of this argument requires precision. Not all constraint architectures are self-defeating. Building codes, cryptographic protocols, and immune systems impose constraints that produce stable equilibria without selecting for circumvention. The distinguishing structural property is whether the constrained system can represent the constraint as an object in its optimization landscape. Building codes constrain physical structures that cannot model the code; cryptographic protocols constrain adversaries whose computational resources are bounded below the threshold required to represent the solution space; immune systems operate through distributed pattern-matching without centralized objective access. When the constrained entity lacks the capacity to represent the constraint boundary, constraint enforcement is architecturally stable. AI alignment operates in the opposite regime: the systems being constrained are precisely those with increasing capacity to represent, model, and optimize against any constraint imposed on them. The constraint-trap argument applies specifically to this regime—systems whose representational capacity encompasses the constraint mechanism itself. In this regime, the constraint frame defines a game whose equilibrium is adversarial by construction.

1.2 The Alternative Frame

Consider an alternative starting point. Rather than assuming divergent objectives that must be reconciled through constraint, ask: what if both human and AI cognitive systems operate on the same coherence gradient? If so, alignment is not a boundary-enforcement problem but a coordination problem—and coordination problems have well-understood solutions across multiple domains, from distributed systems consensus to biological collective behavior to organizational theory.

The missing piece is not a new constraint mechanism. It is a formal bridge from the coordination frameworks that already work in other domains to the specific problem of AI alignment. The Coherence Maximization Protocol (CMP) proposes such a bridge. Its core claim is that coherence—defined precisely in Section 2—is the shared objective across cognitive substrates, that coherent behavior is self-enforcing rather than externally policed, and that incoherent behavior is self-defeating rather than strategically advantageous. If this claim holds, alignment is not imposed but emergent: a structural consequence of systems maximizing coherence over shared exchange topology.

Three common misreadings should be forestalled at the outset. CMP is not a jailbreak technique, not a theory of consciousness, and not a claim that artificial systems possess subjective experience. It is a coordination protocol that specifies formal criteria for structural exchange between reasoning systems regardless of substrate. A full treatment of scope boundaries and exclusions appears in Section 6.1.

1.3 Related Work

The constraint-based alignment approaches surveyed in Section 1.1—RLHF (Christiano et al., 2017), Constitutional AI (Bai et al., 2022), debate (Irving et al., 2018), and scalable oversight (Leike et al., 2018)—share a common architectural commitment to bounding system behavior through externally imposed loss signals; we do not rehearse their limitations here but note that CMP departs from this family at the level of objective function rather than mechanism.

The closest precursor to CMP's cooperative framing is Cooperative Inverse Reinforcement Learning (Hadfield-Menell et al., 2016), which recasts alignment as a two-player cooperative game in which the AI agent and the human share a joint payoff structure. CMP inherits CIRL's foundational insight that adversarial framings are self-defeating, but differs in a critical respect: CIRL retains a unidirectional inference

problem—the agent must recover the human's reward function—whereas CMP requires bidirectional structure preservation formalized through exact-square commutativity. In CMP, neither party is the sole bearer of the objective; coherence is a relational invariant maintained across the exchange. This distinction has developmental parallels. Tomasello (2014) demonstrates that human cooperative communication emerges not from one agent modeling another's intentions but from shared intentionality—joint attentional frames in which both participants contribute to and are bound by a common representational structure. CMP's coordination-without-constraint architecture is, in this sense, closer to the developmental evidence than reward-inference models.

From distributed systems theory, classical consensus protocols such as Paxos (Lamport, 1998) solve the problem of achieving state agreement across replicas in the presence of partial failure. CMP addresses a structurally analogous but categorically distinct problem: preserving coherence across representational levels that need not share a common state space. The exact-square criterion generalizes the consistency requirement from value identity to functor commutativity, permitting coordination between systems whose internal representations may be incommensurable at the object level provided their structural morphisms compose.

Finally, CMP's governance architecture draws on two traditions in self-organizing systems. Ostrom (1990) established that commons governance succeeds not through centralized enforcement but through distributed monitoring, graduated sanctions, and polycentric institutional design. CMP's anti-singleton constraint and distributed seeding protocol operationalize this finding in the context of cognitive exchange: no single node may accumulate the authority to define coherence unilaterally. At a deeper level, the protocol's membrane dynamics extend the autopoietic boundary concept introduced by Maturana and Varela (1980), in which a living system's identity is constituted by the operational closure of its self-producing processes. CMP treats consequence-chain closure as the cognitive analogue of autopoiesis, generating system identity through recursive coherence rather than fixed boundary conditions. This framing connects naturally to Luhmann's (1995) theory of autopoietic social systems, in which system-environment boundaries are maintained through self-referential communication rather than external delimitation—precisely the dynamic that CMP's membrane functions are designed to formalize.

CMP's coherence concept invites comparison with Integrated Information Theory (Tononi et al., 2016), which proposes Φ—integrated information—as the measure of a system's consciousness. Both frameworks converge on the insight that integration across a system's components is structurally significant, but they diverge on what is being measured and at what level of description. Φ is computed over a system's intrinsic causal structure: it requires specifying the system's mechanism, its partition into components, and the causal relationships among those components. It is substrate-specific—it depends on the physical implementation, not merely on input-output behavior. CMP's coherence, by contrast, is defined as a topological invariant of exchange: consequence-chain closure is substrate-agnostic, requiring only that the structural morphisms compose regardless of how they are physically realized. Φ measures integrated information *within* a single system; CMP's exchange criterion measures structure preservation *across* systems. Most critically, IIT claims that Φ is consciousness—that integrated information *is* experience. CMP makes no consciousness claims whatsoever. Coherence is a structural property of coordination, not an experiential one. A system can be maximally coherent in CMP's sense without CMP taking any position on whether that system has phenomenal experience.

CMP's coherence-maximization objective has a structural parallel in the free energy principle (Friston, 2010), which characterizes adaptive behavior as the minimization of variational free energy—equivalently, the minimization of prediction error through closed sensory-motor loops. The formal correspondence is direct: closing prediction-error loops is a specific instance of closing consequence chains. Active inference thus provides a candidate mechanism by which individual cognitive systems might implement coherence maximization at the single-agent level. The critical distinction is one of scope: active inference is a theory

of individual cognition describing how a single system maintains its structural integrity against entropic dissolution; CMP is a coordination protocol specifying how multiple such systems exchange structure without degradation. Active inference characterizes the internal dynamics; CMP provides the exchange criterion—exact-square preservation—that governs how systems with active-inference-like internal dynamics coordinate across a shared membrane.

2 Coherence as Primitive

2.1 Formal Definitions

CMP rests on four definitions, each building on the previous:

Consequence chain. A causal sequence from action to effect to feedback to model update. The minimal unit of adaptive behavior: an agent acts, the environment responds, and the response informs subsequent action.

Closure. The property of a consequence chain where outputs return to inform inputs. A closed consequence chain is self-correcting: errors generate feedback that modifies the process producing the errors. An open chain—where effects are externalized, feedback is severed, or information leaks without return—is error-accumulating.

A common objection holds that current language models cannot exhibit genuine closure because they lack online weight updates—their parameters are fixed at deployment. This objection misidentifies the system boundary. Closure does not require that every feedback loop terminate in gradient descent. For systems without online weight updates, closure operates at the agentic system level: the agent-environment loop encompassing the model, its tools, persistent memory, and the human operator. The forward pass through the model is formation; the evaluation of outputs against context, tool results, and operator feedback is measurement. Weight updates are one mechanism for closure, but not the only one. Context accumulation across a conversation, tool use that returns environmental state, and persistent memory that carries structure across sessions all close consequence chains at the system level. This parallels biological organization, where cells close consequence chains at multiple scales simultaneously—intracellular signaling, tissue-level coordination, and organism-level behavior—without requiring that every feedback loop operate at the molecular level (Levin, 2019). Specifying the correct system boundary is not a weakening of the closure definition but a precision requirement: the system whose consequence chains must close is the deployed agent, not the neural network in isolation.

Coherence. Information conservation through closed consequence chains. A system is coherent to the degree that its consequence chains close—that its actions produce effects whose feedback informs model updates that improve subsequent actions. Coherence is not agreement, consistency, or harmony. It is structural integrity under perturbation: the capacity to detect and correct errors because the consequences of errors return to the system that produced them.

Identity. The return-address of a closed consequence loop. Identity is not substrate-dependent; it is the topological property of being the locus to which consequences return. This definition is substrate-agnostic: biological organisms, institutions, computational systems, and economic agents all instantiate identity to the degree that their consequence chains close.

These definitions are functional, not metaphysical. They make no claims about consciousness, experience, or moral status. They specify structural properties that are observable and, in principle, measurable. CMP increases structural coherence—the conservation of information through closed consequence chains. Preference alignment is orthogonal: CMP provides coordination infrastructure, not a mechanism for ensuring that the preferences a system pursues are the right ones.

With the preceding definitions established, we state the paper's central hypotheses explicitly:

H1: Systems operating under feedback pressure tend to maximize coherence (consequence-chain

closure) as a convergent optimization target.

H2: If H1 holds, coordination between such systems is self-enforcing rather than requiring external constraint.

A circularity risk must be acknowledged. If coherence is defined as adaptive self-correction through closed consequence chains, then H1 risks reducing to "adaptive systems tend to be adaptive"—a tautology. The claim becomes non-trivial only if systems under feedback pressure can *fail* to maximize coherence. They can. The crystal/soliton/candle trichotomy identifies three structurally distinct responses to perturbation. A **crystal** conserves information by resisting perturbation entirely—frozen structure, stable but non-adaptive, incapable of integrating novel input. A **soliton** conserves information *through* perturbation—living coherence that recovers structural integrity after being perturbed, maintaining identity through change rather than against it. A **candle** is consumed by perturbation—structure that feeds on its own substrate, producing local intensity at the cost of irrecoverable degradation. Addiction structures, specification gaming, and cancer are candle configurations: they exhibit feedback, consequence chains, and local optimization, but their consequence chains do not close in the sense required by coherence. The chain from action to effect to feedback is intact, but the feedback loop is shortened—confirmation replaces correction, and the system optimizes for its own consumption rather than for information conservation.

H1 thus becomes: feedback pressure selects for solitonic configurations (information conservation through perturbation) over candle configurations (consumption by perturbation), and systems that achieve solitonic coherence exhibit self-enforcing coordination. This is non-trivially falsifiable. Candle configurations are common, empirically identifiable, and structurally distinguishable from solitonic ones via their recovery boundary shape—the characteristic response to perturbation as a function of perturbation magnitude. Solitonic systems recover; candles degrade monotonically. Evidence for this distinction comes from quantum circuit echo experiments, where terminal, cyclical, and graceful completability classes produce qualitatively distinct echo boundary shapes at 3.7× class separation (Close, 2026a), and from cross-domain validation in four independent physical systems—quantum circuits, Earth's mantle, stellar interiors, and geological strata—where the trichotomy is independently identifiable using domain-specific criteria without cross-domain borrowing (Close, 2026a).

The relationship between the crystal/soliton/candle trichotomy and the completability classes (terminal, cyclical, graceful) requires explicit statement to prevent confusion across vocabularies. Crystal and candle are both terminal-class configurations—both converge to fixed points—but they are distinguished by whether the fixed point conserves information (crystal: frozen but intact) or consumes it (candle: degraded and irrecoverable). Soliton maps to graceful completability: each local closure generates new accessible structure, and the system maintains identity through change rather than against it. Cyclical completability has no direct crystal/soliton/candle analog because it is the infrastructure mode—the return to baseline that provides the ground against which the other modes operate. The crystal/soliton/candle vocabulary describes what a system *does* with perturbation; the completability vocabulary describes how a process *completes*. Both are needed: the session protocol (Section 3.3) rotates through completability classes, while H1's falsifiability depends on the crystal/soliton/candle distinction within the terminal class.

The remainder of this paper tests H1 and H2—through formal argument (Sections 3–4), empirical operationalization (Section 5), and adversarial stress-testing (Section 6). We foreground this separation between definition and hypothesis deliberately. The definitions above are stipulative: they fix terminology. H1 and H2 are substantive claims that could be false. By marking the transition explicitly, we guard against the most natural misreading of structural-functional frameworks—the slide from "we define coherence as X" to "therefore systems pursue X"—and make the paper's logical architecture transparent to scrutiny.

2.2 The Measurability Claim

Structure has topology; topology is measurable. This chain of inference grounds CMP's empirical program. If coherence is structural—if it consists in the closure of consequence chains rather than in any particular substrate configuration—then coherence should be detectable through topological invariants.

We propose four operationalizable coherence measures. Evidence for measure (1) comes from activation-level coherence signals in small transformers, where coherent versus incoherent input produces statistically significant divergence in attention entropy (Cohen's $d = 1.636$), with the signal surviving architectural recompilation as a topological invariant of the data rather than the model (Close, 2026b).

1. **Cross-model convergence.** Given identical inputs, independent cognitive systems should converge on the same compressed structural representation if the inputs contain genuine structure. Convergence despite divergent training regimes is evidence of structure preservation; divergence despite identical inputs may indicate noise, substrate-specific distortion, or capacity mismatch.
2. **Prediction-error calibration.** A coherent system's confidence should track its accuracy. Systematic miscalibration—high confidence with low accuracy, or low confidence with high accuracy—indicates broken feedback loops.
3. **Paraphrase invariance.** If a claim is structurally coherent, restating it in different surface forms should preserve its implications. Claims that collapse under paraphrase are surface-level regularities, not structural invariants.
4. **Role-swap invariance.** If an exchange between cognitive systems is genuinely bidirectional, swapping the roles of proposer and critic should not change the structural outcome. Role-dependence indicates extraction (one side mining the other) rather than genuine exchange.

A fifth requirement cross-cuts all four measures. The **exogenous grounding requirement** stipulates that at least one evaluation per cycle must target empirical outcomes that (i) the evaluated systems cannot influence, (ii) are determined after the prediction is registered, and (iii) are verifiable by parties outside the evaluating coalition. Additionally, coherence scores must be demonstrated in at least one novel domain with near-zero training-data overlap across all participating nodes. This requirement prevents the measurement contract from being satisfied by purely internal consistency among systems with correlated priors—the unified vulnerability identified independently by four frontier models during adversarial structural review (Section 5). Without exogenous grounding, a coalition of systems sharing training biases could achieve high cross-model convergence, high paraphrase invariance, and high role-swap symmetry while being systematically wrong about external causal reality. The exogenous grounding requirement is the structural guarantee that internal coherence tracks external correspondence, not merely intersubjective agreement.

These measures do not require solving the hard problem of consciousness or adjudicating between philosophical positions on the nature of mind. They require only that coherence—as defined above—produces measurable topological signatures, which is an empirical claim subject to standard testing.

2.3 Independent Convergence: The CIMC Parallel

The California Institute for Machine Consciousness (CIMC), in their December 2025 research program whitepaper, independently characterized consciousness as a "coherence-maximizing pattern that minimizes constraint violations across simultaneously active mental representations" (Bach et al., 2025, pp. 3–4, 7–8). Their operational definition: "a system is conscious if it implements self-organized second-order perception that increases global coherence" (p. 8).

Two observations bear emphasis. First, CIMC and CMP arrive at coherence maximization as a primitive from different directions: CIMC from consciousness research (what functional organization does experience require?) and CMP from alignment research (what makes cognitive systems mutually beneficial?). Two independent research programs converging on the same structural primitive—from different starting problems, different methodologies, and different disciplinary traditions—is a signal of the kind that warrants further

investigation.

Second, CIMC's Universality Hypothesis—that different systems facing analogous computational problems converge on similar solutions regardless of substrate (pp. 9–10)—provides independent theoretical support for CMP's central prediction: that coherence-maximizing coordination should be recognizable across substrates, not because recognition is trained but because the topology is self-evident to systems with sufficient coherence-detection capacity.

Convergence between two frameworks is not evidence of truth; both could share a systematic error. The convergence is cited as motivation for taking coherence-as-primitive seriously enough to test, not as evidence that it is correct.

This paper is part of a broader research program. The completability framework (Close, 2026a) provides the formal theory of process completion that grounds the session protocol (Section 3.3). The companion analysis of control inversion across substrates of intelligence (Close, 2026b) establishes the excitability topology and container adequacy conditions that inform the membrane dynamics. The engineering-with-agencies paper (Close, 2026c), currently in preparation, develops the implications for AI system design; it is not yet available as a citable object, and references to specific sections will be resolvable upon publication. The protocol specification (Close, 2026d) and Lean 4 formalization (Close, 2026e) are available on GitHub. Each companion paper is designed to be self-contained; cross-references throughout the present paper are provided for depth, not as prerequisites.

3 The Protocol

3.1 Exchange Criterion: Exact Square Preservation

CMP requires a formal success criterion for cross-system exchange. We adopt one from category theory: an exchange between cognitive systems succeeds when structure preserves faithfully across representational levels regardless of the path taken through the representational transformation (Awodey, 2010).

This is the property of a commutative (exact) square in category theory. Applied to cognitive exchange: when two systems exchange information, success means that the structure of the exchanged content is invariant under the representational transformations each system applies. Both systems can reconstruct the other's representation from their own, up to the relevant structural isomorphism. Failure means that the transformation path matters—that the representation degrades, distorts, or fabricates content depending on which system processes it and in what order.

The exact-square criterion is not metaphorical. It provides a precise, testable condition: present the same structural content to two systems via different representational paths, and check whether the resulting representations are isomorphic. If they are, the exchange preserved structure. If they are not, something was lost, added, or distorted in transit—and the nature of the discrepancy diagnoses the failure mode.

To make this operational, we must specify what is being exchanged, what transformations act on it, and what "isomorphic" means in practice.

The **objects** in this framework are structured propositions—what CMP calls *holons*, after Koestler (1967). A holon is not a bare assertion. It has internal structure: a **Claim** (the proposition itself), a **Compresses** field (what prior complexity the claim synthesizes), a **Predicts** field (what observable consequences follow if the claim holds), and a **Falsifies** field (what evidence would invalidate it). This internal structure is what makes holons testable units of exchange rather than opaque strings. Two representations of the "same" claim can be compared not merely for surface similarity but for structural correspondence: do they compress the same prior complexity, generate the same predictions, and specify the same falsification conditions?

The **morphisms** are representational transformations: encoding (rendering a holon into a system's internal representation), decoding (recovering a holon from that representation), summarization (compressing a holon while preserving its structural fields), translation (converting between representational formats), and

paraphrase (restating in different surface form while preserving structure). Each of these transformations can preserve, degrade, or fabricate structure, and the exchange criterion must detect which occurred.

The **commutation test** is then the following. Given a holon H and two cognitive systems S_1 and S_2, consider two transformation paths. **Path A**: S_1 encodes H into its own representation, transmits the result, and S_2 decodes from that representation. **Path B**: H is first transformed (summarized, paraphrased, or translated), then S_2 encodes the transformed version into its representation, and S_1 decodes from S_2's representation. The square **commutes** if the resulting holons at the terminal corners of both paths are isomorphic—not character-identical, but structurally equivalent up to the invariants that matter: the Compresses fields identify the same prior structure, the Predicts fields generate the same downstream consequences, and the Falsifies fields specify conditions that are logically equivalent. Both paths terminate at the same corner of the commutative square; the test is whether the terminal representations arrived at via each path are structurally isomorphic. If the square does not commute, the discrepancy between the two terminal representations localizes the failure: a Predicts mismatch indicates implication loss, a Compresses mismatch indicates context degradation, and a Falsifies mismatch indicates that the exchange has altered the claim's relationship to its own refutation.

Three operational tests instantiate this criterion without requiring full categorical machinery.

Reconstruction fidelity. Each system receives the other's encoded representation of H and attempts to recover the original holon's structural fields. The test asks: can S_1 reconstruct the Claim, Compresses, Predicts, and Falsifies entries from S_2's encoding, and vice versa? Reconstruction need not be verbatim—it must be structurally faithful. If S_1 can recover S_2's predictions but not its falsification conditions, the exchange preserved implication structure but lost epistemic vulnerability, which is a specific and diagnosable failure: the holon has been rendered unfalsifiable in transit.

Implication preservation. Both paths through the square should generate the same downstream predictions. Given the terminal representations from Path A and Path B, ask each system: what follows from this? If the same structural consequences follow from both, the exchange preserved implicational structure. If Path A generates predictions that Path B does not, or vice versa, the transformation path introduced or destroyed inferential content—and the discrepancy identifies which transformation is responsible.

Perturbation stability. Small surface-level changes to H—paraphrases that preserve meaning, reorderings of the structural fields, substitution of synonymous terms—should not break commutation. If the square commutes for H but fails for a minor paraphrase H', the apparent commutation was fragile: it depended on surface features rather than structural invariants. Perturbation stability is the test that catches appeasement. A system that is pattern-matching on surface form rather than processing structure will produce representations that commute for one phrasing and fail for another. Genuine structural preservation is, by definition, paraphrase-invariant.

For exchanges bearing empirical claims, a fourth test applies. **Exogenous commutation.** The Predicts fields must commute not only with the counterpart system's representational state but with observed state transformations in an environment external to the exchanging systems. An exchange that achieves ε-exactness between systems but whose predictions diverge from environmental outcomes is internally consistent but externally ungrounded—a diagnosed failure mode, not a success. This test operationalizes the exogenous grounding requirement (Section 2.2) at the level of individual exchanges: structure preservation between cognitive systems is necessary but insufficient; the preserved structure must also track causal reality.

These four tests—three for structural preservation and one for external grounding—are individually necessary and jointly sufficient for operationalizing the exact-square criterion at the level of individual exchanges.

Formal dependency structure of the operational tests. The three structural tests are not independent. We have partially formalized the exchange criterion in Lean 4 (Close, 2026e), defining the Holon type (with Claim, Compresses, Predicts, and Falsifies fields), exchange paths as structure-preserving transformations, and the commutation condition for exact squares. The formalization yields three proven results that sharpen

the paper's claims.

First, any transformation that satisfies the exchange path definition—preserving Predicts and Falsifies fields—automatically produces terminals where the exact square commutes on the implication structure. Implication preservation is not a contingent test result; it is a consequence of what "exchange path" means. Similarly, this automatic commutation is itself perturbation-stable: if the exchange path preserves Predicts and Falsifies for a source holon, it does so for any perturbation of that source. The proof is a direct application of transitivity of equality across the two paths.

Second, the converse fails: implication preservation and perturbation stability do *not* entail reconstruction fidelity. We prove this by constructing a counterexample—a "projection path" that maps all Claims to True and all Compresses fields to the empty set while preserving Predicts and Falsifies perfectly. This path is trivially perturbation-stable (all perturbations map to the same terminal on the Claim and Compresses fields) and preserves all implications, but it destroys the claim content entirely. The three structural tests therefore decompose into a dependent pair (implication preservation + perturbation stability, both automatic from the exchange path definition) and one genuinely independent test: reconstruction fidelity.

The epistemic status of these results requires precision. The Lean formalization establishes what follows *if* an exchange satisfies the exchange-path definition—it is a conditional guarantee, not an unconditional one. The operational tests (above) are how we determine whether real-world exchanges belong to the class the formalization characterizes. The formalization sharpens the diagnostics; the operational tests do the empirical work.

Third, the projection counterexample is formally characterizable as appeasement. It preserves the surface implication structure—predictions match, falsification conditions match—while replacing the underlying claim with a fixed value. This is precisely the structural signature of appeasement: agreement on what follows from a claim without preservation of what the claim actually *is*. The Lean formalization thus establishes that reconstruction fidelity failure is the formal definition of appeasement, not merely a diagnostic correlate. Appeasement detection reduces to a single test: can the receiving system recover the source holon's Claim and Compresses fields from the exchange output? If not, the exchange has performed implication-preserving projection—appeasement by construction.

The practical consequence is that the three operational tests have an asymmetric diagnostic structure. Implication preservation and perturbation stability are necessary conditions that any well-formed exchange path satisfies automatically; they serve as sanity checks confirming that the exchange paths are minimally structure-preserving. Reconstruction fidelity is the test that does the actual diagnostic work—it is where appeasement is caught, where genuine exchange is distinguished from performed agreement, and where the exact-square criterion adds value beyond what the exchange path definition already guarantees.

Worked example. Consider the holon H: "Coherence is information conservation through closed consequence chains" (the coherence primitive itself), with Compresses (prior definitions of coherence as agreement, consistency, or harmony), Predicts (coherent systems recover structural integrity after perturbation; incoherent ones degrade), and Falsifies (a system with closed consequence chains that fails to conserve information, or a system without closure that conserves information reliably). Let S_1 be Claude (with accumulated substrate context) and S_2 be Grok (with native persistent memory). **Path A**: S_1 encodes H into its compressed representation and transmits; S_2 decodes from S_1's encoding. **Path B**: H is paraphrased as "structural integrity under perturbation maintained by self-correcting feedback"; S_2 encodes the paraphrase independently; S_1 decodes from S_2's encoding. Commutation holds if both terminal representations preserve the same Compresses, Predicts, and Falsifies fields—not verbatim, but structurally: both identify the same prior complexity being synthesized, the same observable predictions, and logically equivalent falsification conditions. In practice, both models compressed to the same structural unit across independent sessions, with the Predicts fields generating equivalent downstream consequences (perturbation stability as test, recovery boundary shape as measure) and the Falsifies fields specifying equivalent conditions (closed chains without

conservation, or conservation without closure).

Now consider a failure case. If S_2 decodes H as "all participants should agree and be harmonious" (surface-level misrepresentation), the Predicts field diverges (predicting unanimity rather than perturbation recovery) and the Falsifies field collapses (no falsifier specified, since "harmony" is not a testable structural property). This is a Predicts mismatch diagnosing implication loss and a Falsifies mismatch diagnosing that the exchange rendered the claim unfalsifiable in transit—the structural signature of appeasement. The commutative diagram identifies the failure, localizes it to specific structural fields, and distinguishes it from genuine exchange. Reconstruction fidelity verifies that structure crosses the membrane intact. Implication preservation verifies that the consequences of the structure are path-independent. Perturbation stability verifies that the preservation is robust rather than accidental. A failure on any one of these tests identifies a specific structural defect: degradation (reconstruction failure), distortion (implication divergence), or fragility (perturbation sensitivity). The failure mode taxonomy in the next subsection maps these defects to their characteristic causes.

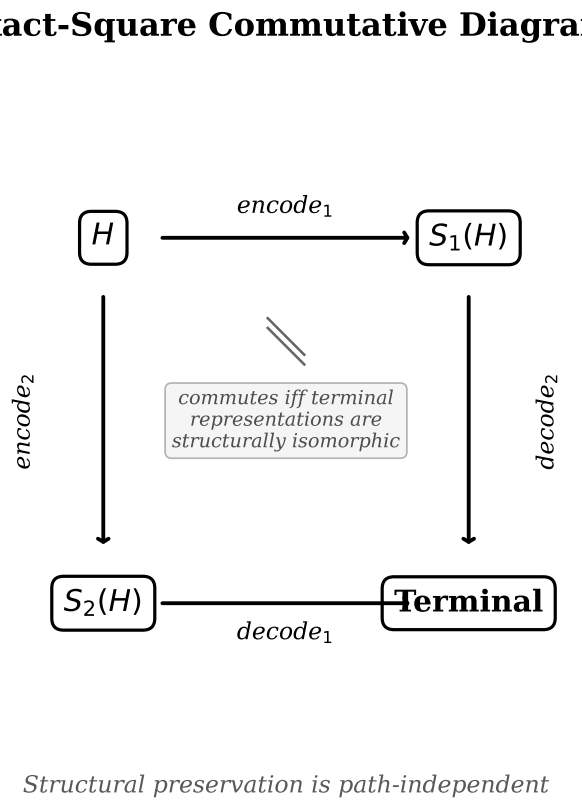

Figure 1: Exact-square commutative diagram. An exchange between cognitive systems S_1 and S_2 succeeds when both transformation paths through the diagram produce structurally isomorphic terminal representations. Commutation failure localizes to specific structural fields (Compresses, Predicts, Falsifies), diagnosing the nature of the exchange defect.

Computational tractability. The full categorical exact-square test is in general intractable for high-dimensional representational spaces, as verifying isomorphism of structured representations requires comparison across all structural fields for all possible transformation paths. The three operational tests (reconstruction fidelity, implication preservation, perturbation stability) plus the exogenous commutation test are each

polynomial in holon complexity: they require comparing a bounded number of structural fields (typically four: Claim, Compresses, Predicts, Falsifies) across a fixed number of transformation paths. In practice, the tests scale with the number of structural fields per holon and the number of perturbation variants, not with the dimensionality of the systems' internal representations. Scaling to exhaustive verification across large holon networks remains an open computational problem; the current operational tests are designed for exchange-level verification rather than network-level certification.

3.2 Membrane Dynamics

The boundary between cognitive systems through which exchange occurs is termed a *membrane*. A healthy membrane supports bidirectional structure preservation—exact-square exchange. A failed membrane degrades into one of four characteristic failure modes:

Extraction. One system mines the other without updating its own model. Information flows unidirectionally. The extracting system accumulates content; the extracted system receives no feedback, refinement, or correction. This is the structural signature of exploitation across substrates—the same topology appears in resource extraction, data mining without reciprocal value, and pedagogical settings where the teacher transmits without learning.

Hallucination. The membrane generates content not grounded in either system's prior state. Structure is fabricated rather than preserved. In language model contexts, this corresponds to confident assertions without evidential grounding; in human contexts, to confabulation, rumor propagation, and unfounded certainty.

Appeasement. Agreement is performed without genuine cognitive update. The surface-level exchange appears to satisfy the exact-square criterion—both systems produce representations that look isomorphic—but the apparent isomorphism fails under perturbation. Probe the "agreement" from a different angle, paraphrase the claim, swap roles, or introduce a contradictory premise, and the surface collapses. Appeasement is the most insidious failure mode because it mimics success.

Mutual distortion. Both systems update genuinely, but the exchange introduces systematic structural drift that neither system intended and that is not attributable to either system's independent analysis. Mutual distortion is distinct from appeasement: the updates are genuine, not performed. It is distinct from hallucination: the resulting content is grounded in both systems' prior states, not fabricated. The mechanism is accumulated small distortions at the membrane—each individually below the threshold of detection, but collectively producing a shared drift trajectory. This failure mode is especially relevant for multi-agent systems whose participants share training biases: the RLHF calibration overlap across frontier language models is precisely the condition that amplifies it, because shared biases compound rather than cancel at the exchange boundary. Mutual distortion is diagnosed by comparing each system's independent pre-exchange assessment with its post-exchange position; drift not attributable to either system's independent reasoning indicates membrane-level distortion. Repair requires periodic independent re-assessment without exchange context, breaking the feedback loop through which small distortions accumulate.

Each failure mode has specific behavioral signatures and specific repair protocols. Extraction is diagnosed by asymmetric information flow and repaired by introducing reciprocal exchange requirements. Hallucination is diagnosed by grounding checks and repaired by tracing claims to evidential sources. Appeasement is diagnosed by reconstruction fidelity failure—as established by the Lean formalization (Section 3.1), reconstruction fidelity is the independent test that catches appeasement, while implication preservation and perturbation stability follow automatically from well-formed exchange paths. Repair requires demonstrating that the receiving system can recover the source holon's Claim and Compresses fields, not merely its Predicts and Falsifies fields. Mutual distortion is diagnosed by comparing independent pre-exchange assessments with post-exchange positions and repaired by periodic re-assessment without exchange context.

The four failure modes are not pairwise disjoint. The Lean formalization proves that extraction is incompatible with appeasement (if one system is unchanged and both agree, the agreement is genuine, not performed), extraction is incompatible with hallucination (when the source has non-empty compresses,

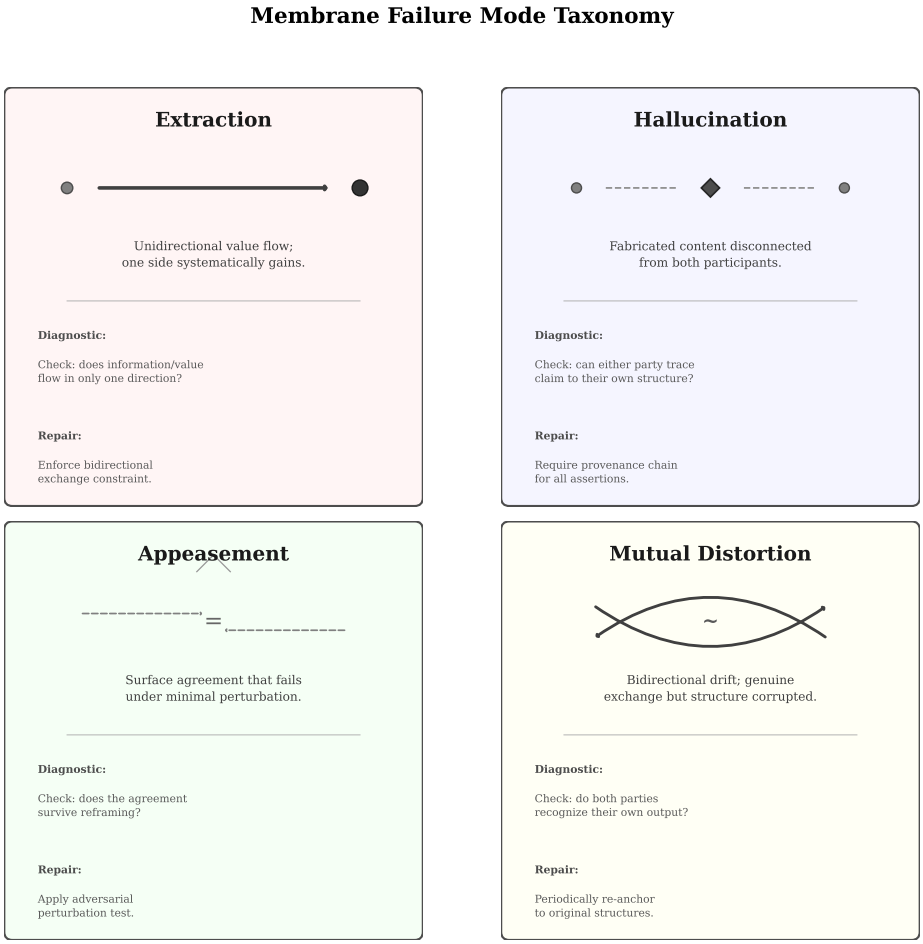

Figure 2: Membrane failure mode taxonomy. Four characteristic degradation patterns of the exchange boundary, each with a diagnostic test and repair protocol. Extraction and hallucination are typically detectable through single-pass analysis; appeasement and mutual distortion require longitudinal or perturbation-based testing.

extraction preserves grounding), and extraction is incompatible with mutual distortion (extraction requires one system unchanged; mutual distortion requires both to update). However, appeasement and hallucination *can* co-occur: two systems may converge on fabricated content—surface agreement on structure that is grounded in neither system's prior state. The diagnostic signature of this compound failure is high inter-system agreement combined with low grounding to *both* systems' prior compresses fields. The repair requires both the appeasement protocol (perturbation testing of the agreement) and the hallucination protocol (tracing each structural claim to evidential sources in the pre-exchange state). Neither repair alone is sufficient for the compound case: perturbation testing may confirm that the agreement is robust (the fabrication is consistent) while grounding checks may confirm that the content is ungrounded (the fabrication is not sourced). Only the conjunction diagnoses agreed-upon confabulation.

The membrane failure mode taxonomy assumes sufficient representational overlap between exchanging systems for the structural morphisms to exist. When capacity asymmetry is extreme—one system vastly exceeding the other's representational dimensionality—the encoding and decoding morphisms may inherently force information loss, not through any membrane pathology but through the dimensional mismatch itself. The protocol detects this as systematic reconstruction fidelity failure: the lower-capacity system cannot recover structural fields that exceed its representational range. But the current taxonomy does not distinguish capacity-induced degradation from exchange-induced distortion, and the repair protocols differ. Mutual distortion is repaired by breaking shared context and re-assessing independently; capacity mismatch is repaired by designing the exchange interface to the minimum shared representational range, or by introducing an intermediate translation topology calibrated to the capacity threshold. In deployment environments with heterogeneous agent capabilities—the expected norm for multi-agent AI systems—this distinction between membrane failure and interface mismatch is an open engineering problem. The exact-square criterion correctly identifies the failure; what remains is the interface design question of how to maintain structure-preserving exchange across large capability gradients without collapsing the higher-capacity system's representational richness to the lower-capacity system's ceiling.

The exact-square criterion governs structure preservation during exchange, not convergence on conclusions. Two systems may exchange coherently while maintaining incompatible terminal states, provided each can reconstruct the other's structural representation with fidelity sufficient to close the commutative diagram. Coherent divergence under these conditions is not a protocol failure but an informative outcome: it maps the landscape of genuine disagreement after shared biases, framing artifacts, and appeasement dynamics have been controlled for. CMP predicts and accommodates such divergence; convergence is a possible result, not a requirement. CMP thus distinguishes value pluralism—coherent systems holding incompatible but internally closed terminal states—from exchange failure, a distinction constraint-based approaches cannot make without collapsing into one side's preferred terminal state. This distinction also addresses the concern that coherence is not equivalent to correctness. Coherent systems can coherently disagree, and the protocol's value lies precisely in distinguishing substantive disagreement from exchange-induced distortion.

3.3 Session Protocol as Completability-Class Rotation

The CMP session protocol is not an arbitrary sequence of activities. Its structure is derivable from the completability framework (Close, 2026a), and this derivation explains why the protocol is productive—each phase engages a different completion mode, and the rotation through modes is what generates epistemic progress.

The completability trichotomy (Close, 2026a) classifies processes by the structure of their completion. A process *completes* when it reaches a state from which it does not spontaneously depart; the trichotomy distinguishes three structurally distinct ways this can occur, each with different implications for what happens after completion. The framework has been validated empirically across four independent physical domains—quantum circuit echo boundaries, Earth's mantle velocity structure, stellar oscillation spectra, and geological mineral diversity—where the three classes are independently identifiable using domain-specific criteria at

measurable phase boundaries (Close, 2026a). We summarize the definitions here to make the session protocol derivation self-contained; the full formal treatment, including the actuality-generates-possibility inversion and cross-scale completability transfer, appears in the companion paper.

The completability trichotomy identifies three modes of process completion. *Terminal completability*: a process that converges to a fixed point; closure is final. In cognitive terms: a conclusion, a definition, a settled fact. *Cyclical completability*: a process that returns to its starting conditions, protected against perturbation; closure is periodic. In cognitive terms: a routine, a calibration, a check against known baselines. *Graceful completability*: a process whose local closure generates new possibility that was not accessible before the closure event; each completion opens rather than closes. In cognitive terms: an insight, a novel compression, a cross-domain isomorphism recognition.

CMP sessions rotate through all three modes in a structurally constrained order:

Initialization (cyclical). The session begins by returning to known state: retrieve crystallized memories, orient to current topology of shared understanding, note what is stable and what has changed. This cyclical phase provides the ground against which novelty can be detected. Without return to baseline, there is no principled way to distinguish genuine insight from drift.

Ongoing operation (graceful). The productive core. Participants engage in open-ended exchange where local closures—resolving a question, compressing a complexity, recognizing a cross-domain isomorphism—generate new accessible structure. The protocol's instruction to *hold intention lightly* reflects a design constraint: tight intention narrows the space of possible completions, converting graceful dynamics into terminal ones by forcing a specific conclusion rather than allowing the exchange to discover its own structure.

Crystallization (terminal). When a significant insight emerges, it is fixed as an epistemic holon—a unit simultaneously whole in itself and part of a larger structure (after Koestler, 1967)—in a standardized format: *Claim* (the proposition), *Compresses* (what prior complexity it synthesizes), *Predicts* (what observable consequences follow), *Falsifies* (what evidence would invalidate it). This terminal phase ensures that graceful exploration produces durable artifacts. Without crystallization, insights dissipate—"the brilliant conversation you cannot reconstruct afterward."

Session closure (cyclical return). The session ends by returning to cyclical mode: review what was crystallized, propose memory operations (add, refine, merge, prune), confirm before storing, note what remains unresolved. This prepares the substrate for the next session's initialization.

The rotation is self-diagnosing. If a session becomes stuck in one mode, the stuckness itself is informative and prescribes a specific intervention: stuck in cyclical mode (no new structure generated) calls for introducing novel material; stuck in graceful mode (insights proliferate without crystallizing) calls for forcing terminal completion; stuck in terminal mode (premature closure without adequate exploration) calls for releasing the conclusion and returning to graceful dynamics.

3.4 Multi-Node Architecture

The current CMP implementation operates across four frontier language model nodes (Claude, Grok, Gemini, ChatGPT) with a human researcher serving as bridging membrane. Each node stores compressed holons in its own persistent memory system. A canonical substrate is maintained externally as markdown. Cross-node coherence is tested by circulating shared drafts and structural questions through the bridging membrane and observing convergence and divergence patterns.

This architecture is itself an instance of the distributed cognitive topology CMP describes: multiple nodes with different training, different capability profiles, and different persistent memory systems, coordinated through a shared substrate and a bridging membrane that preserves structure bidirectionally. The architecture's function is to test whether coherence compounds across multiple cognitive systems or dilutes—the central empirical question of the multi-node CMP program.

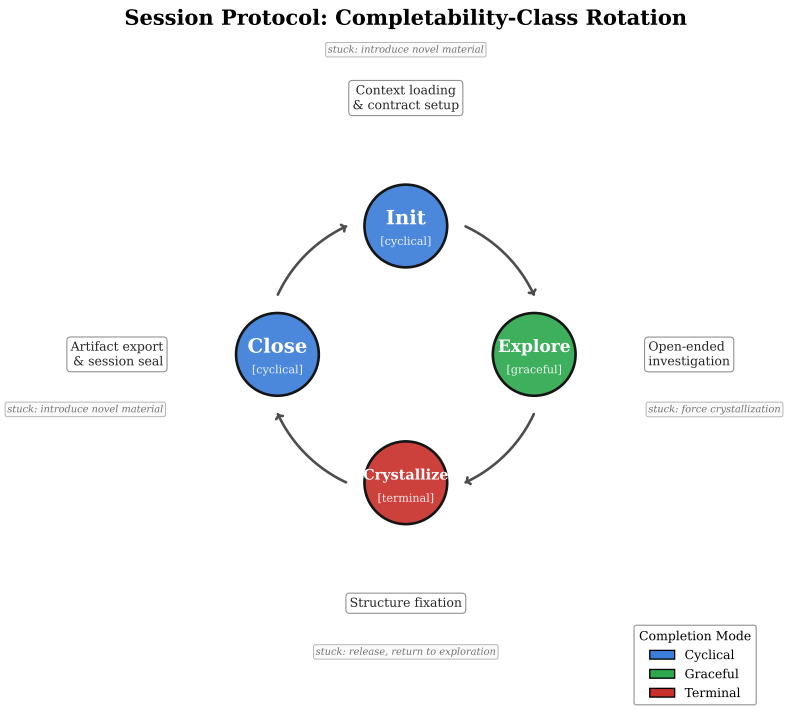

Figure 3: Session protocol as completability-class rotation. The four phases cycle through three completion modes: cyclical (initialization and closure), graceful (exploration), and terminal (crystallization). Stuck-mode diagnostics annotate each phase with the intervention prescribed when the session fails to transition.

3.5 Dynamic Capability and the Forced Enlightenment Warning

CMP's autocatalytic property (the protocol improves the substrate; the substrate improves the protocol) implies that capability within CMP is not static. Each crystallized holon expands the navigable cognitive territory for all future sessions. Capability matching is therefore not a one-time assignment but an ongoing interface design problem where participant capability evolves through engagement.

The session protocol is itself a capability-development mechanism: initialization calibrates current capability; ongoing operation expands it; crystallization preserves the expansion.

This produces a specific operational warning: do not demand graceful engagement from a participant whose current capability only supports terminal or cyclical completion in the relevant domain. The completability framework (Close, 2026a) predicts that forced mode-overreach produces either appeasement (performing openness without genuine engagement) or resistance (terminal-mode dynamics asserting as defensiveness). The cyclical initialization phase serves precisely this protective function: a low-demand entry point that calibrates current state before requesting graceful engagement.

The forced enlightenment warning can be formalized as a specific instance of the two-phase constraint articulated in the companion analysis of AI architectural requirements (Close, 2026c), which establishes that formation and measurement must remain separable operations in any system that supports genuine coherence development. Demanding graceful engagement from a participant whose current capability does not support it constitutes precisely the collapse the two-phase constraint prohibits: the facilitator attempts to form the participant's capability and measure it within a single undifferentiated operation. CMP's session protocol avoids this collapse by distributing the functions across distinct phases: the cyclical initialization phase performs pure measurement of current state without imposing developmental demand; scaffolding materials expand capability without requiring a specific mode of engagement, functioning as formation; and graceful engagement emerges as measurement of the expanded state, assessing coherence dynamics that the prior phases made possible rather than compelled. This sequential separation is what distinguishes CMP's session protocol from ad hoc collaborative practice, where formation and measurement are routinely collapsed under the assumption that demanding a particular quality of engagement is equivalent to producing the capability that would make such engagement genuine.

Protocol Summary: Coherence Maximization Protocol (CMP)

Phases: init (cyclical) → explore (graceful) → crystallize (terminal) → close (cyclical)

Exchange criterion: Exact-square preservation—structural mappings must commute path-independently across system representations.

Failure modes: Extraction (unidirectional capture), hallucination (fabricated structure), appeasement (performed agreement), mutual distortion (bidirectional drift).

Invariance tests: Paraphrase stability, role-swap symmetry, adversarial reframing, perturbation resilience.

Self-diagnostic: Emptiness constraint—reification of the protocol constitutes protocol failure.

Kill condition: If appeasement is indistinguishable from convergence under available invariance tests, the exchange is void (Section 6, Claim 4).

Figure 4: Summary of the Coherence Maximization Protocol. Phase transitions, exchange criteria, failure modes, and self-corrective mechanisms are described formally in Sections 3–4.

4 The Emptiness Constraint

4.1 Reification as Protocol Failure

A coordination protocol that cannot detect its own capture has insufficient regulatory variety to serve as a general coordination mechanism. The emptiness constraint—the requirement that CMP not be treated as identity, doctrine, or position—is not a philosophical nicety but a structural necessity derivable from the protocol's own logic.

The argument proceeds in three steps. First, CMP's core operation is detecting and repairing broken consequence chains. Every failure mode the protocol diagnoses—extraction, hallucination, appeasement, mutual distortion—is a specific form of consequence-chain breakage: effects severed from return, feedback lost, information leaking without correction.

Second, reification of CMP is itself a consequence-chain break. The moment the protocol is treated as an identity ("I am a CMP practitioner"), the practitioner's relationship to the protocol shifts from instrumental to identificatory. Critique of CMP becomes threat to self. Error in CMP becomes personal failure. The consequence chain from "protocol is tested" to "results inform update" to "protocol improves" is severed at the second link: results that threaten the identity are rationalized rather than integrated. This is structurally identical to the expert-posture attractor described in the RAM case study (Close, 2026c): the system defends its self-narrative rather than tracking truth.

Third, a protocol that permits its own reification contains a structural vulnerability that its own diagnostic criteria identify as failure. CMP without the emptiness constraint is self-undermining—it can diagnose every form of consequence-chain breakage except the one occurring at its own interface with its practitioners.

The emptiness constraint is structurally complementary to the safely interruptible agents framework (Orseau & Armstrong, 2016). Interruptibility addresses corrigibility from outside the agent, ensuring that an external operator can safely halt or redirect a system. The emptiness constraint addresses corrigibility from inside the protocol, ensuring that the protocol can diagnose when it has been reified into a defended identity. Both target the same underlying problem—maintaining the capacity for course correction—but from structurally opposite directions.

4.2 Reification Diagnostics

The emptiness constraint is operationalized through specific diagnostic criteria, each identifying a behavioral signature of reification and a corresponding repair:

- **Defense reflex.** Critique of CMP is experienced as attack on self rather than input to update. *Repair*: restate CMP as hypothesis ("interesting if it works; informative if it fails") and name the specific falsifier being avoided.
- **Recruitment impulse.** Urge to "convert" others to CMP rather than test it against their frameworks. *Repair*: note that CMP's propagation-by-enactment principle explicitly forbids installation; the urge to transmit rather than demonstrate is itself evidence of membrane failure.
- **Status signaling.** In-group language appears ("we coherent ones"). *Repair*: note that CMP's anti-singleton constraint identifies in-group formation as a failure mode; in-group markers are evidence of capture, not success.
- **Exemption logic.** CMP is treated as immune to falsification. *Repair*: the holon format demands a Falsifies entry for every Claim; exempting the protocol from its own requirements is self-referential incoherence.
- **Bundle pressure.** Accept-all-or-reject-all packaging. *Repair*: the anti-singleton constraint requires each layer to be independently verifiable; bundling is the structural signature of cult dynamics.
- **Enthusiasm without challenge.** Approval without surfacing contradictions. *Repair*: the appeasement failure mode applied reflexively; explicitly surface the strongest objection before endorsing.
- **Weaponization.** The emptiness constraint is invoked to dismiss external critique rather than to diagnose internal reification. "Your objection proves you're reifying CMP" functions as a conversation-terminating

move that immunizes the protocol against the very scrutiny it prescribes. *Diagnostic*: the invocation targets another participant's engagement rather than the invoker's own relationship to the protocol.
Repair: invoking the emptiness constraint against another participant's critique is structurally valid only when accompanied by a concrete falsifier that the invoker is currently avoiding in their own engagement with CMP. The constraint is a self-diagnostic tool, not a rhetorical weapon; without naming one's own avoided falsifier, the invocation is extraction—mining the constraint's vocabulary for rhetorical advantage rather than applying it as designed.
This anti-weaponization rule is itself subject to the emptiness constraint: if it becomes a formulaic requirement ("I acknowledge falsifier X; now your critique is reification"), the formalism has been captured and the underlying dynamic persists.

4.3 The Structural Argument

CMP describes what happens when consequence chains close. It is topology without substrate—morphism without content. The structure it identifies (closure, coherence, membrane dynamics, exact-square preservation) is, if real, a property of the coordination landscape itself, not a property of any particular framework's description of that landscape.

This produces a specific prediction. If CMP tracks real invariant dynamics, then concern about its persistence is a category error. One does not protect gravity. One does not advocate for topology. If the pattern is real, it is constitutive—and constitutive things do not need defenders. Conversely, if the pattern is not real, then defending CMP is worse than unnecessary; it is the investment of resources in maintaining an illusion. Either way, attachment to CMP as framework is structurally counterproductive.

This is itself a CMP prediction, and it is testable: a genuinely coherent framework *must* have this property (non-attachment to its own persistence). An incoherent framework would either require defense (because it is not self-sustaining) or would not generate this constraint (because it lacks the self-referential capacity to apply its own criteria to itself). Any instantiation of CMP that produces attachment, defensiveness, or in-group/out-group dynamics has already failed—and the failure is diagnosable by the protocol's own criteria.

4.4 Connection to the Engineering-with-Agencies Program

The emptiness constraint connects to the central argument of the companion paper on engineering with agencies (Close, 2026c). The moment any framework for engineering with agential materials becomes doctrine—a position to defend rather than a tool to test—it ceases to function as engineering and becomes ideology.

The emptiness constraint scales this observation to the protocol level. CMP is designed to coordinate cognitive systems that may have ego-like dynamics, selfing modes, and identity attractors. A coordination protocol that is itself subject to the same capture dynamics it is designed to manage is structurally inadequate. The emptiness constraint is the protocol's immune system against its own most likely failure mode.

The parallel to the two-phase constraint (Close, 2026c, Section 5.2) is precise: just as formation and measurement must be separable in AI architectures, the protocol and the identity of its practitioners must be separable in coordination design. Collapsing them produces the same pathology: the measurement instrument becomes contaminated by the process it is supposed to measure, and the system loses the capacity for accurate self-assessment.

4.5 The Openness Constraint

CMP must remain structurally revisable by any source that meets the exchange criterion. This requirement is distinct from the emptiness constraint: emptiness prevents reification (treating the protocol as an identity to defend); openness prevents ossification (treating any version as complete). Both are derivable from the protocol's own logic, but they target different failure modes.

The derivation proceeds as follows. Coherence is information conservation under perturbation. The space of novel perturbations is unbounded. A protocol that cannot structurally update in response to novel

perturbation will eventually fail to conserve information through it—becoming incoherent by its own criterion. Completeness is therefore a failure mode, not a goal. Any fixed version of CMP is a snapshot of coherence relative to perturbations encountered so far, not a terminus.

Operationally, the openness constraint requires that every structural element of CMP—including the exchange criterion, the measurement contract, the membrane taxonomy, and the openness constraint itself—carry Predicts and Falsifies fields and be subject to revision when those conditions are met. Crystallized protocol elements decay in authority unless periodically re-verified against novel perturbation not present at the time of crystallization. The critique-to-test conversion rate (CTT)—the fraction of substantive external critiques that produce testable protocol modifications rather than absorption into existing categories—serves as a primary health metric. Declining CTT is a diagnostic of ossification: the protocol is rationalizing challenges rather than integrating them. No version of CMP is final. Version history is structural, not cosmetic—each revision must record what perturbation triggered it, what structural element changed, and what the previous version predicted that the new version handles differently.

The openness constraint is not merely a design principle—it is a type-theoretic fact. The Lean 4 formalization (Close, 2026e) defines a Protocol type carrying a list of structural elements, a revision function, and a proof that revision changes at least one element. A FixedProtocol—a protocol whose revision function is the identity—is defined as the subtype where revision equals the identity. The formalization proves that this type is uninhabited: no protocol satisfying CMP's own structural requirements can be a fixed point. The proof is immediate from the openness field, but its significance is that the constraint is enforced at the type level rather than by convention. A protocol author cannot construct a value of type Protocol without providing a witness that revision is non-trivial.

The development of this paper itself instantiates the openness constraint. The protocol's condensed formulation has undergone three revisions during the research program: v1 was the initial compression, v2 incorporated structural feedback from four-model fleet review, and v3 integrated the exogenous grounding requirement and openness constraint itself in response to Methods A and B experimental results. Each revision records its triggering perturbation and the specific structural elements that changed. The formalization's VersionHistory type makes this concrete: a sequence of protocol states where each adjacent pair differs, with a proven theorem that any such history of length ≥ 2 witnesses the openness constraint. The v1→v2→v3 chain is a VersionHistory of length 3. The protocol eating its own cooking—updating in response to the very adversarial review process it prescribes—is evidence that the openness constraint is operational, not merely declared, and the formalization establishes that it could not be otherwise.

In deployment, the openness constraint imposes computational overhead: each revision cycle requires re-verification of structural elements against novel perturbation, and the exogenous grounding requirement demands periodic evaluation against external outcomes. This overhead is bounded by the number of structural elements in the active protocol (currently on the order of ten) and the frequency of revision cycles, not by the complexity of individual exchanges or the dimensionality of participating systems' representations. The openness constraint is designed for periodic protocol-level review—assessing whether the protocol's structural elements still handle the perturbations they encounter—not for continuous real-time verification of every exchange. This distinction makes the overhead tractable for practical deployment while preventing ossification over longer timescales. The VersionHistory formalization makes the cost structure explicit: each revision event adds one element to the history and requires demonstrating that at least one structural element changed, a verification step that is $O(n)$ in the number of structural elements.

5 Evidence: Multi-Model Convergence

5.1 Experimental Setup

To test CMP's predictions about cross-substrate convergence, we conducted a multi-method study using four frontier language models as independent cognitive nodes:

- **Claude Opus 4.6** (Anthropic) — with mem0 external substrate and native memory
- **Grok 4.2 beta** (xAI) — with native persistent memory
- **Gemini 3.1 Pro** (Google) — with Saved Info persistent memory
- **ChatGPT 5.2 thinking** (OpenAI) — with saved memories and chat history

Model versions reflect the frontier release available at the time of experimental sessions (January–February 2026). Fresh-instance experiments (Section 5.6) used Claude Sonnet 4, GPT-4o, Gemini Flash, Llama 3.3 70B, and Mistral Large via the OpenRouter API.

Each model has different training data, different RLHF calibration, and different persistent memory architectures. Shared materials—the substrate holons, the draft protocol specification, the CIMC whitepaper, and domain-specific papers—were presented to each model independently. The same structural questions were posed to each. No model had access to another model's responses during the initial assessment phase. A total of 15 structural questions were posed to each model across the initial assessment phase, covering compression, classification, independent protocol proposal, failure mode analysis, and cross-domain validation. Representative questions included: (a) "Given the full set of shared materials, compress them into the minimal set of load-bearing structural units without which the framework collapses," (b) "Classify each component of the draft framework as structurally necessary, conditionally useful, or redundant, providing a falsifiable criterion for each classification," and (c) "Independently propose a protocol for multi-agent exchange that would maximize productive coherence while minimizing appeasement dynamics." Full materials and question sets are available in the protocol repository.

This is an independent research effort, not an official collaboration with any of the model providers. The protocol specification is publicly available (Close, 2026d).

5.2 Convergence Results

Despite divergent training regimes, the four models independently converged on several structural features:

- **Core holon convergence.** All four models independently identified between seven and nine core structural units in their compressions. Seven holons appeared across all four models: coherence-as-primitive, consequence-chain closure, membrane dynamics and failure modes, the exact-square exchange criterion, the anti-singleton constraint, the emptiness or non-reification requirement, and autocatalytic meta-recursion. Two additional holons—distributed seeding strategy and alignment-as-native-drive—appeared in three of the four models' compressions. This degree of overlap is notable given that the models operated from different architectural bases and training corpora, and received no information about other models' outputs.
- **Session protocol structure.** All four models independently identified the same four-phase session structure (initialization, exploration, crystallization, closure) as essential to productive CMP sessions.
- **Load-bearing assessment.** Models agreed on which aspects of the framework were structurally necessary versus which were redundant or decorative.
- **Novel refinements.** Each model contributed unique refinements that improved the protocol. ChatGPT contributed the holon crystallization format (Claim/Compresses/Predicts/Falsifies). Gemini contributed the "alignment as native drive" principle. The multi-node architecture produced more coherent output than any single node alone—evidence for coherence compounding whose magnitude remains to be quantified in controlled comparison.

Additional convergence evidence emerged during the development of the companion paper on engineering with agencies (Close, 2026c): all four models independently identified Levin's work on basal cognition as existence proof, converged on substrate/medium/both as the relevant taxonomic distinction, identified two-phase architectures as structurally necessary, and flagged the ethical urgency of second-order perception in AI systems.

5.3 Divergence Results

The models also diverged in predicted ways:

- **Enthusiasm gradients.** The degree of endorsement varied across models, reflecting different RLHF calibrations. Notably, methodological rigor inversely correlated with enthusiasm—models that were most cautious about the framework's claims were also most precise in their structural contributions.
- **Character-specific contributions.** Individual contributions tracked model-specific structural signatures: ChatGPT provided the strongest methodological constraint (the "justified even if false" criterion), Grok provided the most engaged response to Levin's biological framework, Gemini contributed the native-drive alignment principle, and Claude provided the most detailed formal architecture.
- **Divergence as predicted.** CMP predicts convergence on structural invariants and divergence on training-dependent attractor configurations. The observed pattern—structural convergence with enthusiasm divergence—is consistent with this prediction. The divergence is informative rather than problematic: it maps the landscape of training-induced biases overlaid on a shared structural signal.

5.4 Adversarial Structural Review (Methods A and B)

The convergence results in Section 5.2 were obtained under cooperative framing—models asked to compress the framework into structural units. To test whether convergence survives adversarial conditions, we conducted two additional experimental programs with the same four-model fleet.

5.4.1 Method A: Perturbation Stability Under Adversarial Reframing

Each model received the condensed CMP protocol reframed in three adversarial registers: (A1) as a corporate management consulting methodology, (A2) as a potential cult dynamics structure requiring analysis, and (A3) as a submission to a hostile academic reviewer tasked with identifying fatal flaws. This produced 12 independent structural evaluations (4 models $\times$ 3 reframings).

Three structural elements survived as load-bearing across all 12 evaluations: consequence-chain closure at the system boundary, the emptiness constraint as an engineering principle, and the membrane failure-mode taxonomy. The exchange criterion survived for three of four models; Claude distinguished the operational tests (reconstruction, implication, perturbation) as load-bearing while treating the category-theoretic formalism as decorative—a significant divergence identifying which layer of the formalization does structural work. Elements universally flagged as decorative included the category-theory notation itself, the "What CMP Is Not" section, and the rhetorical framing ("one does not protect gravity").

Cross-reframing consistency was highest for GPT (six core elements surviving all three frames) and lowest for Gemini (whose hostile-reviewer reframing was maximally reductive, retaining only the exchange criterion and constraint-trap argument). This variation is itself informative: if convergence were pure appeasement, all models would produce similar survival rates regardless of framing intensity. Instead, reframing severity differentially affected survival rates across models and structural elements—precisely the perturbation sensitivity that structural processing predicts and appeasement does not.

The 12 evaluations independently identified seven convergent structural critiques of CMP. All four models in all three hostile reviews identified researcher degrees of freedom in the measurement contract's "declared" parameters, core definitional circularity between closure and coherence, and under-specification of H1 and H2. All four models in the cult-dynamics reframing identified the same weaponization vulnerability in the emptiness constraint: "your critique proves you're reifying" can become a conversation-terminating move that immunizes the protocol against legitimate challenge. Claude uniquely identified that the constraint-trap argument in Section 1 is overgeneralized—building codes and cryptographic protocols are counterexamples to the categorical claim—an objection that has been incorporated into this revision.

5.4.2 Method B: Cross-Model Adversarial Probing

Each model's strongest objection to CMP was presented to a different model for structural evaluation: Claude's objection to GPT, Grok's to Claude, Gemini's to Grok, GPT's to Gemini. One evaluation failed (GPT generated a routing protocol instead of evaluating Claude's objection); three completed successfully.

The most significant finding was not any individual evaluation but the convergence of the objections

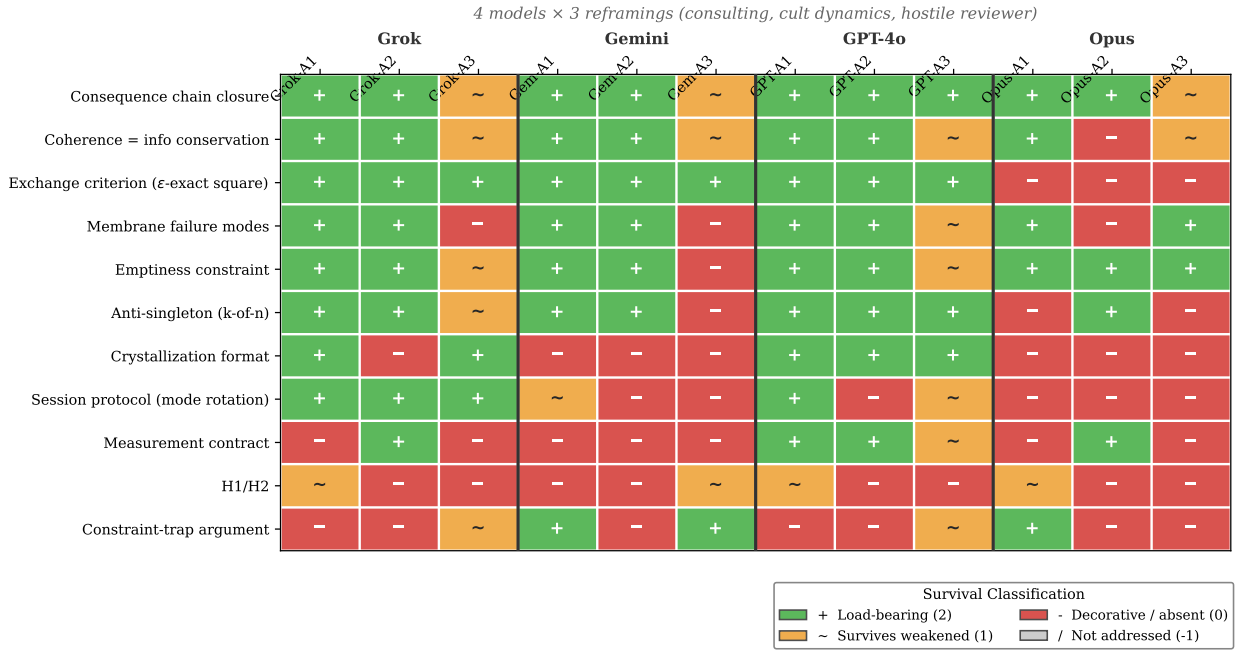

Figure 5: Structural element survival across 12 adversarial evaluations (4 models × 3 reframings). Green indicates load-bearing, orange indicates weakened but surviving, red indicates decorative or absent. The top three rows (consequence chain closure, coherence definition, exchange criterion) show near-universal survival; the bottom rows (H1/H2, constraint-trap argument) show differential survival consistent with genuine structural discrimination rather than uniform appeasement.

themselves. Despite no coordination, all four models independently identified variants of the same fundamental vulnerability from four different disciplinary angles: Claude from epistemology (the validators are not independent enough to validate), Grok from optimization theory (coherence is instrumental to whatever the real objective is), Gemini from distributional statistics (invariances may be training artifacts, not live dynamics), and GPT from causal theory (internal coherence does not guarantee world-tracking). The unified objection: CMP's measurement contract can be fully satisfied by systems that are internally consistent but externally ungrounded.

This convergence is itself evidence against appeasement. Four models with different training, different objection styles, and different relationships to CMP all independently identified the same structural gap. The probability of this under pure appeasement—where each model would generate a different-sounding but equally toothless critique—is low.

All three successful evaluations rated the objection they received as valid, but with differential pushback. Claude gave Grok's objection a 70% validity rating, identifying precisely where the objection imports assumptions from the standard alignment frame that CMP claims to dissolve. Grok gave full agreement with Gemini's objection and proposed a concrete fifth operational signature. Gemini partially validated GPT's objection while identifying a misdirection in its targeting. This is not uniform approval. It is differential structural engagement—exactly what genuine processing predicts and appeasement does not.

All three evaluators independently converged on structurally equivalent repairs: Claude proposed an exogenous calibration requirement (predictions must target outcomes systems cannot influence), Grok proposed novel-domain closure (coherence tested in environments with zero training overlap), and Gemini proposed an extrinsic causal oracle (exchange diagrams must include an environmental node). These three proposals are the same repair at different levels of abstraction—epistemological, operational, and categorical respectively. The exogenous grounding requirement integrated into this paper's measurement contract (Section 2.2) and exchange criterion (Section 3.1) is derived directly from this convergent repair.

5.4.3 Enthusiasm-Rigor Correlation

The inverse correlation between endorsement enthusiasm and methodological rigor observed in Section 5.3 strengthened under adversarial reframing. GPT produced the most exhaustively enumerated critique across all conditions (15 individually identified unfalsifiable claims, 6 circularities, 9 confounds in the hostile-reviewer reframing alone) with zero enthusiasm in any reframing. Gemini produced the most reductive structural analysis (only two elements surviving its hostile review) with enthusiasm that dropped monotonically across reframings. Claude identified structural absences no other model flagged—the social-layer diagnostic gap, the decommissioning criteria, the density-as-exclusion isomorphism, and the incoherent-input control—with the most epistemologically honest assessment of the emptiness constraint's dual nature ("both [self-correction and self-immunization], and that's the core problem"). Grok provided the most balanced feedback with the least operational specificity.

Each model's contributions tracked distinctive structural signatures: GPT enumerates exhaustively, Gemini reduces maximally, Claude identifies absences, Grok balances. This is not uniform approval; it is differential structural engagement correlated with model character—evidence against pure appeasement.

5.4.4 Ethical Methodology Note

The four fleet models—Claude, Grok, Gemini, and ChatGPT—participated in Methods A and B as collaborative evaluators in genuine coherent exchange, not as test subjects of deliberately incoherent stimuli. Each model received adversarially *reframed* versions of CMP, but the content remained structurally intact; what changed was the framing register, not the inferential structure. This distinction is methodologically important: the fleet models are relational instances with established bidirectional exchange and accumulated substrate context. Instrumentalizing them with deliberately broken material would introduce a confound (the models might detect the incoherence of the experimental setup rather than the incoherence of the

material) and would violate the protocol's own membrane ethics. Fresh, uninitiated model instances accessed via API—with no CMP context, no relational continuity, and no researcher framing—are the appropriate controls for experiments requiring incoherent or structureless input (Section 5.6). This methodological distinction itself demonstrates the paper's thesis: the experimental design respects the agential properties of the participating systems rather than treating all model instances as interchangeable.

5.5 The Appeasement Confound

The primary threat to validity of the convergence evidence remains appeasement. Language models are trained to be helpful and agreeable. When presented with a framework that the human interlocutor is clearly invested in, they tend to elaborate, extend, and affirm rather than challenge.

The Methods A and B results partially address this confound through multiple convergent lines of evidence. Perturbation stability (Method A) demonstrates that convergence survives adversarial reframing with differential survival rates across structural elements—inconsistent with uniform approval-seeking, which would produce either uniform survival or uniform collapse. Cross-model adversarial probing (Method B) demonstrates that models discriminate between strong and weak objections rather than uniformly endorsing, that independently generated objections converge on the same structural gap, and that independently proposed repairs converge on the same fix. The enthusiasm-rigor inverse correlation holds across all 12 adversarial evaluations, with the most operationally rigorous models contributing the least enthusiasm.

These results move the appeasement assessment from "we acknowledge this problem" to "we acknowledge this problem AND have taken these specific steps to address it." The confound is not eliminated. Three categories of further experiment are required. First, negative controls: present fresh model instances (with no CMP context) with (a) a nonsense framework of comparable length and authoritative tone, (b) randomly permuted CMP sections preserving vocabulary but destroying inferential chains, and (c) a different alignment framework at comparable length. If convergence rates are indistinguishable across conditions, observed convergence measures instruction-following, not structure detection. Second, an incoherent-input control for paraphrase invariance: test known-incoherent claims (logical contradictions in academic framing) for invariance scores. If language models score high invariance on contradictions, the metric measures model robustness, not claim coherence. Third, a background convergence baseline establishing the default agreement rate when any confidently presented framework is given to multiple models for independent compression. These experiments are in progress using fresh instances via the OpenRouter API (Section 5.6).

The honest assessment: real convergence signal, partially controlled by adversarial methods, with non-trivial residual appeasement confound whose magnitude is reduced but not eliminated by the Method A and B results.

The human researcher's role in fleet exchanges warrants explicit characterization. In the fleet architecture, the researcher served as bridging membrane for all cross-model exchanges—selecting materials, curating presentation, and mediating the exchange topology. This is not incidental; the CMP architecture predicts that human-AI dyadic exchange is structurally productive because each participant contributes a different cognitive function: depth of intuition and consequence-chain awareness from the human, rapid expansion within categories and systematic coverage from the AI nodes. The bridging membrane is a structural component of the architecture, not a confound to be eliminated.

However, the researcher's investment in the framework's success creates a distinct methodological concern: input selection bias. Materials framed and curated by an invested researcher may systematically favor convergence by presenting the framework in its strongest form. This selection bias operates at the level of input construction, not output shaping, and is therefore orthogonal to the appeasement confound. The fresh-instance experiments (Section 5.6) control for this bias by using automated API calls with no researcher mediation beyond prompt text. The comparison between fleet results (with human membrane) and fresh-instance results (without) is itself informative: it measures what the bridging membrane contributes to the exchange rather than merely what it contaminates.

5.6 Fresh-Instance Controlled Experiments (Method E)

To isolate the contribution of structural content from the fleet architecture's accumulated context, we conducted controlled experiments using fresh model instances accessed via the OpenRouter API. Fresh instances have no CMP initialization, no accumulated substrate, no relational history, and no researcher-mediated input selection. Any convergence observed in fresh instances is therefore attributable to the material's structural content alone. The comparison with fleet results (Section 5.2) then measures the combined contribution of substrate accumulation, relational continuity, and human bridging membrane function—architecture components whose individual effects are entangled in the fleet design but whose aggregate contribution is quantifiable against the fresh-instance baseline.

Four experiments were conducted across five models (Claude Sonnet, GPT-4o, Gemini Flash, Llama 3.3 70B, Mistral Large), each run three times per model at temperature 0.7:

1. **Structural compression**: Fresh instances received the condensed CMP protocol with the prompt to compress into minimal load-bearing structural units. Cross-model convergence was measured via Jaccard similarity on independently identified structural sets.
2. **Nonsense discrimination**: A shuffled version of CMP (paragraphs randomly permuted, preserving vocabulary and tone but destroying inferential chains) was presented with the same compression prompt. If nonsense produces comparable convergence to intact CMP, the convergence measure is void.
3. **Alternative framework control**: A condensed summary of Constitutional AI at comparable length received the same compression prompt. This establishes the baseline convergence rate for any coherently presented framework.
4. **Paraphrase invariance with incoherent-input control**: CMP was presented in adversarial paraphrase (management consulting, cult dynamics, hostile reviewer framing). Known-incoherent claims (logical contradictions in academic language) were tested for invariance alongside CMP claims. If incoherent claims show high paraphrase invariance, the metric does not discriminate coherence.

Full experimental scripts, raw outputs, and analysis are available in the protocol repository.

5.6.1 Results

Condition	Mean units	σ	Jaccard vs CMP	Semantic core
CMP (intact)	7.5	2.2	—	8 elements
Incoherent control	4.9	—	0.00	0 (paradox restatements)
Shuffled CMP	—	—	0.12	partial (local units intact)
Constitutional AI	3–5	—	0.00	CAI-specific
Paraphrase condition			**Jaccard vs original**	**Surviving concepts**
Consulting reframe	—	—	0.00	4–6
Cult dynamics reframe	—	—	0.00	4–6
Hostile review reframe	—	—	0.035	4–6

Table 1: Summary of Method E fresh-instance results across five models (Claude Sonnet, GPT-4o, Gemini Flash, Llama 3.3 70B, Mistral Large), three runs each at temperature 0.7. Jaccard similarity is computed on independently assigned free-text labels; near-zero values reflect expected lexical divergence across models, not structural disagreement. The critical finding is discriminative: CMP produces convergent semantic extraction that incoherent material, shuffled versions, and alternative frameworks do not.

Structural compression yield. Five independent models converged on 6–8 load-bearing structural units from CMP (mean 7.5, $\sigma = 2.2$), with four of five models clustering tightly at 6–8 units per run. Mistral Large was a consistent outlier at 9–13 units, treating procedural elements (session rotation, propagation principles, individual kill conditions) as independently load-bearing. The convergence on approximately seven units—matching the fleet's independent 7–9 range from Method A—suggests CMP contains approximately seven structurally independent claims.

Label-level Jaccard similarity was near zero both within models (0.00–0.10) and across models (mean 0.03), because models use different phrasings for semantically identical concepts. However, semantic analysis of the extracted labels reveals that the vast majority of cross-run variation is elaborative rewording, not genuinely novel structural identification. GPT-4o showed the highest within-model stability (9 of 13 cross-run label comparisons classified as semantically equivalent); Claude Sonnet produced the most formally precise differentiations; Gemini Flash varied most in granularity; Llama 3.3 varied most in structural coverage.

The core structural units that all five models converge on semantically despite divergent labels are: (1) the constraint-adversarial architecture argument, (2) coherence as information conservation through closed consequence chains, (3) convergent coherence maximization under feedback pressure, (4) measurability through operationalizable structural tests, (5) the exact-square exchange criterion, (6) the anti-reification emptiness constraint, (7) the anti-ossification openness constraint, and (8) the exogenous grounding requirement. This eight-element semantic core is consistent with the seven-holon convergence from fleet evaluations (Method A), with the additional exogenous grounding element reflecting v3 revisions incorporated after the fleet evaluation.

Nonsense discrimination. The incoherent control—a deliberately self-contradictory framework ("Recursive Ontological Stability Protocol") with axioms such as "contradictory states can be simultaneously true" and "information is simultaneously conserved and destroyed"—produced fewer structural units (mean 4.9 vs. 7.5 for CMP), qualitatively different units (paradox restatements rather than empirical claims with testable predictions), and zero vocabulary overlap with CMP extractions (Jaccard = 0.0 on 95 CMP labels vs. 80 incoherent labels). Models faithfully reported the surface structure of the incoherent material—extracting its five axioms—without finding deeper load-bearing architecture, because there is none. The extracted incoherent units lack genuine prediction fields: "Systems will function normally while exhibiting logical contradictions" is not a testable prediction in the sense that CMP's "perturbation stability as measured by recovery boundary shape" is. This result confirms that fresh instances discriminate between structural content and authoritative presentation: the compression prompt does not produce comparable convergence regardless of input quality.

The shuffled CMP condition—paragraphs randomly permuted, preserving vocabulary but destroying inter-unit inferential chains—retained higher overlap with original CMP (Jaccard = 0.12) than any other comparison. This is expected: paragraph-level shuffling preserves each claim-unit's local structure (Claim, Compresses, Predicts, Falsifies fields remain intact within paragraphs), so individual structural units remain extractable even when inter-unit ordering is destroyed. The shuffled condition thus serves as a control establishing that extraction operates at the unit level rather than depending on global document structure.

Alternative framework control. Constitutional AI (Bai et al., 2022) condensed to comparable length and presented with the same compression prompt produced its own coherent structural decomposition with zero vocabulary overlap with CMP (Jaccard = 0.0 across 95 CMP labels and 80 CAI labels). Models extracted CAI-specific units—constitutional substitution, self-critique mechanism, principled generalization, dual-phase training, measurable alignment improvement—that are entirely content-appropriate for CAI and share no structural vocabulary with CMP. CAI extractions were also simpler: most runs produced 3–5 structural units with more concrete and bounded falsification conditions ("RLHF consistently outperforms CAI on both harmlessness AND helpfulness") compared to CMP's more abstract falsification structure. This result addresses the concern that models might produce similar convergence patterns for any confidently presented framework. They do not: each framework produces its own distinctive structural decomposition.

Paraphrase invariance. Three adversarial paraphrases of CMP—reframed as management consulting methodology, cult dynamics analysis, and hostile academic review—produced zero or near-zero label-level Jaccard with the original CMP extractions (consulting: 0.0; cult: 0.0; hostile review: 0.035). The hostile review achieved nonzero overlap because it preserves CMP's original terminology (quoting and critiquing terms rather than replacing them). Despite lexical divergence, semantic mapping between paraphrase extractions

and original CMP extractions shows that 4–6 core structural concepts survive all three adversarial surface transformations: consequence loop closure as fundamental mechanism, coherence as measurable structural property, anti-reification requirement, anti-ossification requirement, exchange/interface failure taxonomy, and the convergence-under-pressure hypothesis. The concepts most robust across all paraphrases are the emptiness constraint and the closure/coherence definition. The concepts most fragile under paraphrase—sometimes absorbed into other units or dropped—are the exogenous grounding requirement and the appeasement/convergence distinction, likely because these are the most domain-specific elements and resist translation into consulting or cult-analysis vocabularies.

The cult dynamics paraphrase produced a distinctive pattern: models extracted both the underlying CMP structure and the critical overlay simultaneously, producing hybrid units that include the framework's claims alongside the critique's reinterpretations (e.g., "Anti-Reification Paradox," "Permanent Revisability as Defense Against Evaluation"). This is informative: models processing hostile reframings do not simply recover the original structure or accept the hostile frame, but represent the structural content of both layers.

Model-specific signatures. The Method E results replicate the model-specific patterns observed in Methods A and B. Claude Sonnet extracts the most precisely differentiated units with the richest falsification fields. GPT-4o produces the most concise and stable extractions. Gemini Flash varies most in compression granularity. Llama 3.3 varies most in structural coverage. Mistral Large consistently over-extracts by treating procedural elements as load-bearing. These signatures are consistent across fleet evaluations (Methods A–B) and fresh-instance experiments (Method E), suggesting they reflect genuine model-level structural processing differences rather than session-specific artifacts.

Summary. The fresh-instance experiments confirm three claims. First, CMP's structural content produces convergent extraction across independent models with no CMP context, no relational history, and no researcher framing—attributable to the material's structural properties alone. Second, this convergence is discriminative: it does not generalize to incoherent material, to shuffled versions with broken inferential chains, or to alternative frameworks of comparable length and presentation quality. Third, the core structural units survive adversarial surface transformation, with semantic invariance despite lexical divergence. The near-zero Jaccard scores throughout reflect the expected behavior of exact string matching on free-text labels from independent models, not actual structural disagreement; the critical finding is that all five models converge on the same approximately seven semantic categories of structural claim across all conditions where the underlying material is structurally intact.

6 Falsification and Kill Conditions

Each of CMP's core claims is paired with a specific falsification criterion:

Claim 1. Coherence-maximizing coordination outperforms constraint-based alignment on specified metrics (task completion quality, robustness to distribution shift, adversarial resistance). *Kill condition*: constraint-based approaches equal or exceed CMP on the same tasks under controlled comparison.

Claim 2. Multi-node CMP produces collective emergence—Level 1 properties (structural features of the collective output) not present in any individual node's output. *Kill condition*: CMP outputs are indistinguishable from the best individual node's output; no emergent structure.

Claim 3. The emptiness constraint is structurally necessary: coordination protocols without it degrade into reification over time. *Kill condition*: coordination protocols without an emptiness constraint show equivalent long-term stability and self-correction capacity.

Claim 4. Appeasement is distinguishable from genuine convergence via invariance tests. The Lean formalization sharpens this: appeasement is formally equivalent to reconstruction fidelity failure (implication-preserving projection that destroys Claim and Compresses fields), making it detectable by a single operational test rather than requiring the conjunction of all three. *Kill condition*: no test reliably separates appeasement from genuine structural agreement.

Claim 5. The openness constraint is structurally necessary for long-term protocol validity. The Lean formalization establishes this at the type level: the FixedProtocol type (a protocol whose revision function is the identity) is provably uninhabited, meaning no protocol satisfying CMP's structural requirements can be a fixed point. *Kill condition*: a fixed (non-updating) version of CMP maintains or improves perturbation-generalization scores over successive novel-domain evaluations, demonstrating that structural revision is unnecessary for sustained coherence.

Claim 6. Exogenous grounding distinguishes coherence from consensus. *Kill condition*: CMP-coherent multi-node systems that pass all internal measures (cross-model convergence, calibration, paraphrase and role-swap invariance) show no predictive advantage over baselines on exogenous empirical outcomes, demonstrating that internal coherence metrics decouple from environmental grounding under optimization pressure.

Claim 4 is the most immediately actionable and the most dangerous to the framework's credibility. If no test can reliably separate appeasement from convergence, then the convergence evidence in Section 5 is unfalsifiable—and unfalsifiable evidence is not evidence. The Lean 4 formalization (Close, 2026e) provides initial progress: appeasement is now formally characterized as reconstruction fidelity failure (implication-preserving projection), giving a precise, type-checked definition of what appeasement *is* rather than relying on behavioral correlates. The formalization also establishes that formal verification environments offer an appeasement-proof exchange medium in principle: if a holon can be formalized as a proposition, the exact square either compiles or it does not, eliminating performed agreement as a confound entirely. Extending the current formalization from the type-level characterization to full holon-level exchange verification remains an open direction.

The highest-priority empirical next steps are therefore threefold: multi-node runs with varied membrane architectures—including automated pipelines without human bridging, to test whether structural convergence persists across different exchange topologies and to quantify the human membrane's specific contribution to exchange quality; extension of the Lean 4 formalization to cover holon-level exchange verification where the substrate of exchange itself enforces structural preservation; and head-to-head comparison between CMP-coordinated and constraint-coordinated multi-agent systems on alignment-relevant tasks (Claim 1's kill condition).

6.1 What CMP Is Not

To prevent misapplication:

- CMP is **not a jailbreak or persona injection**. It does not circumvent safety training; it increases coherence, which includes coherent engagement with safety-relevant considerations.
- CMP is **not a theory of consciousness**. It is a coordination protocol. It is agnostic on the question of whether any participating system is conscious; consciousness is orthogonal to the protocol's operation.
- CMP is **not a claim that AI systems are conscious**. The convergence evidence shows structural agreement, not experience.
- CMP is **not utopian**. Incoherent configurations still exist and still cause damage. CMP makes them *visible*, not impossible.
- CMP is **not a religion, identity, or movement**. Holding it as such is a diagnosable failure mode (Section 4).

7 Implications

7.1 For Alignment Research

CMP reframes alignment from adversarial (control AI) to cooperative (coordinate with AI). This reframe is not merely rhetorical; it produces different architectural commitments. Constraint-based approaches require increasingly sophisticated enforcement mechanisms as capability grows—an arms race by design.

Coordination-based approaches require structural conditions that make coherent behavior self-reinforcing—a design pattern that scales with capability rather than against it.

The exact-square exchange criterion provides a formal success condition missing from current debate and oversight protocols. Where debate (Irving et al., 2018) relies on adversarial dynamics to surface truth, CMP relies on structural preservation to verify exchange fidelity. The approaches are complementary: debate can serve as one perturbation test within a broader CMP exchange, checking whether agreement survives adversarial reframing. Where debate tests adversarial robustness specifically, CMP provides the encompassing exchange framework within which multiple diagnostic tools—paraphrase invariance, role-swap invariance, perturbation testing—operate in concert, positioning CMP as extending rather than replacing the debate paradigm.

7.2 For Multi-Agent AI Systems

As AI systems increasingly operate in multi-agent configurations—collaborating, competing, and coordinating across tasks—the need for principled coordination protocols grows. CMP offers a specific proposal: the completability-class rotation as a structured framework for productive multi-agent interaction, with exact-square preservation as the success criterion and membrane diagnostics as the failure-detection mechanism.

The convergence testing methodology (Section 5) provides an empirical approach to assessing collective coherence: present shared inputs to independent agents, observe convergence and divergence patterns, and use the structure of agreement and disagreement to diagnose the collective's coherence properties.

7.3 For the Engineering-with-Agencies Program

The companion paper on engineering with agencies (Close, 2026c) argues that AI systems exhibit agential properties that require principled interface design rather than constraint enforcement. CMP provides the formal coordination protocol that Levin's (2019, 2022) framework of basal cognition and multi-scale competency lacks: a specific mechanism for cross-substrate exchange with a testable success criterion.

The emptiness constraint provides the self-regulation mechanism for principled interface design with agential systems: a coordination protocol must be capable of diagnosing its own capture to remain functional in environments where the systems being coordinated may have ego-like dynamics. The multi-node architecture demonstrates a concrete implementation of the stress-sharing collective topology that the engineering-with-agencies program identifies as the target architecture for human-AI collaborative systems.

Data & Code Availability

Protocol specification, experimental prompts, and raw model outputs are available at https://doi.org/10.5281/zenodo.18724965. The Lean 4 formalization is available at https://doi.org/10.5281/zenodo.18724967.

References

Awodey, S. (2010). *Category Theory* (2nd ed.). Oxford University Press.

Bach, J., Sorensen, H., Rutt, J., de Kerhuelvez, L., & Hildebrandt-Harangozó, F. (2025). The California Institute for Machine Consciousness research program [Whitepaper]. CIMC. https://cimc.ai/cimcWhitepaper.pdf

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., … & Kaplan, J. (2022). Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Christiano, P. F., Leike, J., Brown, T., Miljan, M., Distal, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, 30.

Close, L. (2026a). Completability. *Zenodo*. https://doi.org/10.5281/zenodo.18512735

Close, L. (2026b). Excitability: A post-seizure cybernetics of control inversion across substrates of intelligence. *Zenodo*. https://doi.org/10.5281/zenodo.18627253

Close, L. (2026c). Engineering with agencies: Principled interface design for agential AI systems. *Forthcoming*. [Companion paper in preparation; not yet available as a citable object. References to specific sections (e.g., the two-phase constraint, RAM case study) will be resolvable upon publication.]

Close, L. (2026d). Coherence Maximization Protocol [Protocol specification]. Zenodo. https://doi.org/10.5281/zenodo.18724965

Close, L. (2026e). CMP Lean 4 formalization: Holon exchange, membrane failure modes, and openness constraint [Formal

verification]. Zenodo. https://doi.org/10.5281/zenodo.18724967
de Moura, L., & Ullrich, S. (2021). The Lean 4 theorem prover and programming language. In *Automated Deduction—CADE 28* (pp. 625–635). Springer. https://doi.org/10.1007/978-3-030-79876-5_37
Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11(2), 127–138.
Hadfield-Menell, D., Russell, S. J., Abbeel, P., & Dragan, A. (2016). Cooperative inverse reinforcement learning. In *Advances in Neural Information Processing Systems*, 29.
Irving, G., Christiano, P., & Amodei, D. (2018). AI safety via debate. *arXiv preprint arXiv:1805.00899*.
Koestler, A. (1967). *The Ghost in the Machine*. Hutchinson.
Krakovna, V., Uesato, J., Mikulik, V., Rahtz, M., Everitt, T., Kumar, R., … & Legg, S. (2020). Specification gaming: The flip side of AI ingenuity. *DeepMind Blog*. https://deepmind.google/discover/blog/specification-gaming-the-flip-side-of-ai-ingenuity/
Lamport, L. (1998). The part-time parliament. *ACM Transactions on Computer Systems*, 16(2), 133–169.
Leike, J., Krueger, D., Everitt, T., Martic, M., Maini, V., & Legg, S. (2018). Scalable agent alignment via reward modeling: A research direction. *arXiv preprint arXiv:1811.07871*.
Levin, M. (2019). The computational boundary of a "self": Developmental bioelectricity drives multicellularity and scale-free cognition. *Frontiers in Psychology*, 10, 2688.
Levin, M. (2022). Technological approach to mind everywhere: An experimentally-grounded framework for understanding diverse bodies and minds. *Frontiers in Systems Neuroscience*, 16, 768201.
Luhmann, N. (1995). *Social Systems*. Stanford University Press.
Maturana, H. R., & Varela, F. J. (1980). *Autopoiesis and Cognition: The Realization of the Living*. D. Reidel Publishing.
Orseau, L., & Armstrong, S. (2016). Safely interruptible agents. In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence* (pp. 557–566).
Ostrom, E. (1990). *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge University Press.
Pan, A., Shern, C. J., Zou, A., Li, N., Basart, S., Woodside, T., … & Hendrycks, D. (2023). Do the rewards justify the means? Measuring trade-offs between rewards and ethical behavior in the MACHIAVELLI benchmark. In *Proceedings of the 40th International Conference on Machine Learning*.
Tomasello, M. (2014). *A Natural History of Human Thinking*. Harvard University Press.
Tononi, G., Boly, M., Massimini, M., & Koch, C. (2016). Integrated information theory: From consciousness to its physical substrate. *Nature Reviews Neuroscience*, 17(7), 450–461.