ARCHITECTURAL SIGNATURES IN MULTI-AGENT REASONING: PROMPT-INDEPENDENT GENERATION AND INSTRUMENT-DEPENDENT EVALUATION

Marcelo Manucci · CODHZ Research Laboratory

Technical Report M3 · Version 1.1, June 2026

Prepared for open repository deposit (third publication of the CODHZ research program; prior DOIs: 10.5281/zenodo.19287671; 10.5281/zenodo.20121190).

Series nomenclature: internal series label: Paper 2.

Public citation: Manucci 2026c, Technical Report M3, third deposit of the CODHZ research program.

DOI: 10.5281/zenodo.20717393

ABSTRACT

When several large language models are composed into a multi-agent workflow, does the architecture of that composition leave a stable, recognizable trace on the reasoning it produces or is any apparent difference an artifact of how the orchestrating prompts happen to be written? We study three control architectures that process the same complex institutional case: a single-shot single-model analyst (C1), a parallel three-model ensemble with a synthesizer (C2), and a sequential four-role chain with an explicit critic (C3). For each architecture we define in advance a functional signature: the epistemic operation its closure performs, integrating competing hypotheses (C1), arbitrating divergences with an explicit position (C2), or declaring unresolved residual tensions (C3).

We then attack the two threats to any claim that such signatures are real. First, a generation confound: to rule out that the signature is an artifact of author-written prompts, we regenerate every operating prompt with an independent adversarial generator (Google AI Studio, Gemini 3.1 Pro Preview) given only function-level briefings, and re-run all three architectures in triplicate (nine execution chains, twenty-seven outputs, zero failures). All three signatures survive intact in 3/3 replicates. Second, an evaluation confound: in the original blind evaluation, an external judge had barely distinguished one condition ($\Delta = 0.262$), where the internal judge separated it cleanly ($\Delta = 0.911$). We pre-register a blind re-evaluation in which the same external judge (Grok 4.3) classifies the nine anonymized outputs by closure operation rather than by content quality. The judge scores 9/9 (chance $p \approx 5.1 \times 10^{-5}$), with a perfect 6/6 on the critical C2/C3 pair.

The combined result establishes that the signatures are properties of the architectures, not of the prompts, and that they are externally legible, but only when the evaluation instrument asks about

operations. The same judge barely discriminates the conditions with a content-oriented rubric and succeeds perfectly with an operation-oriented one, because content converges across architectures while operations do not. These results resolve two questions left open in the preceding study of this program (Manucci, 2026b): whether process-dependent structural properties are legible to blind external assessment, and whether the structural signature of a reasoning architecture is detectable from output structure alone. The implication for the field is direct: standard quality-oriented evaluation protocols for multi-agent systems are structurally blind to real architectural differences in reasoning.

1. MOTIVATION AND QUESTION

Large language models under standard instructions tend to resolve uncertainty by completing it with the probabilistically dominant configuration, a computational analogue of inferential inertia. Left unmanaged, this produces epistemic monoculture: many independent generations collapse onto the same framing before the space of alternatives has been explored (Manucci, 2026a). Prior work in this program argues that the way to keep that space open is not a better single prompt but a composition of reasoning regimes, an orchestration layer that deliberately routes inference through heterogeneous epistemic stances and harvests the friction between them. The second study of the program (Manucci, 2026b) provided systematic evidence for this claim across six frameworks: the structural advantage of epistemological sequencing concentrates in process-dependent properties (relational density and inferential traceability) and the value of process, rather than in properties replicable by sufficiently prescriptive single-prompt engineering.

That study, however, left two questions explicitly open, and the present study takes both as its point of departure. First, its blind assessors recognized relational density but could not distinguish process traceability without awareness of how the outputs were produced, a limitation the study records as an open question about whether process-dependent properties are legible to external assessment at all (Manucci, 2026b, §5.4, §6.5). Second, one of its assessors correctly identified the sequenced document in all six anonymized pairs from output structure alone, prompting the conjecture that "the structural signature of epistemological sequencing may be detectable from output structure alone" (Manucci, 2026b, §5.4). The present study converts that conjecture into a pre-registered, falsifiable test and resolves the first question by showing that the legibility of process properties depends on the evaluation instrument, not on the properties themselves.

Demonstrating that architectural signatures are real requires defeating two confounds that, until addressed, leave any such claim unfalsifiable:

• The generation confound. If the people who designed the architectures also wrote the prompts, an observed "signature" might be nothing more than a stylistic fingerprint of those authors. Any claim about architecture is then circular.

• The evaluation confound. Even granting that a signature exists, it might be visible only to the same modeling lineage that produced or curated the conditions. A signature that no independent judge can detect is, for practical purposes, not a signature.

This report presents a single experiment (designated M3) built specifically to attack both confounds. We restrict attention to three control architectures (C1–C3). A fourth, proprietary CODHZ architecture (C4) participates in the broader program and is reported separately; nothing in the results below depends on it, which keeps this report self-contained and fully public.

2. EXPERIMENTAL DESIGN

The case. All conditions analyze one fixed, deliberately complex institutional scenario: a century-old private university whose traditional certifying function is being eroded by generative AI, with fragmented faculty positions, budget constraints, unreliable detectors, and an ambiguous external signal to employers. The case is open-ended enough that no single "correct" analysis exists, which is what makes the shape of each analysis (rather than its conclusions) the object of study. At the program level, this constitutes a second analytical case, structurally distinct from the regional health network case used in Manucci (2026b).

The three control architectures.

• C1: single-shot, single model. One model produces the complete analysis in a single pass.

• C2: parallel ensemble + synthesizer. Three different models analyze the case independently, using the same brief; a fourth synthesizer consolidates the three into a single report.

• C3: sequential chain, single model, four roles. The same model passes through four ordered roles (investigator, critic, strategist, synthesizer), each consuming the previous role's output.

The signatures, defined ex ante. Before reading any regenerated output, we fixed the functional signature of each architecture from the original (author-written) line:

• C1. Hypothesis integration: competing hypotheses are not left as a comparative exercise; they are absorbed into a single analytic edifice (one as diagnosis, others as constraint or risk).

• C2. Operational convergence: divergences between analysts are closed by taking an explicit, justified position; the genuinely irreducible is declared as uncertainty rather than forced. One synthetic voice.

• C3. Critical opening: closure declares residual, unresolved tensions; integration does not clausurate the problem space.

The signature is defined at the level of operation, deliberately independent of content and of document format. This definitional choice turns out to be the crux of the entire result (§6–7).

3. INDEPENDENT PROMPT GENERATION (LINE B)

To defeat the generation confound, we replaced every operating prompt with one produced by an independent adversarial generator and never edited thereafter. This extends the control-prompt generation procedure described in Manucci (2026b), used to construct the single-prompt control condition, to regenerate all experimental conditions: the same auditable chain, taken one step further.

Protocol. Each prompt was elicited from Google AI Studio (Gemini 3.1 Pro Preview) as a separate first-output request (no candidate selection, no post-hoc editing), based on a function-level briefing that described the analytic objective without revealing any CODHZ-internal structure, sequencing logic, or epistemological regime. The seven generations were produced within a single recorded AI Studio session; because the generator therefore had its own earlier generations in context, this is conservative with respect to the signature claim, shared context can only push the generated prompts toward one another, working against the architectural differentiation the experiment subsequently recovers, not for it. We logged each generated prompt verbatim in an auditable generation log, together with the briefing, a confirmation that no proprietary architecture had been disclosed, and any incidental regenerations. The chain of custody is: briefing → first generator output → verbatim record → execution without modification.

Generator pre-codings, recorded. The generator spontaneously introduced devices the briefings had not requested, a cross-reference labeling scheme, an explicit evidence-level scale, and fixed paragraph contents for executive summaries. These are reported transparently because they have a measurable, directional effect on traceability and must not be confused with architectural effects.

Execution and dataset. Each architecture was run three times. Execution models were held fixed (a current-generation Claude model for C1 and C3, and one ensemble arm of C2; current-generation GPT and Gemini models for the other two arms, and the synthesizer of C2). The result is nine execution chains and twenty-seven outputs, completed without a single runtime failure.

A clean internal check confirms nothing was reconfigured between replicates: the fixed input token counts (the prompt-plus-case that does not depend on prior outputs) are byte-identical across the three replicates of every condition, while the chained inputs vary exactly by the amount the preceding outputs varied. Total compute cost for the full Line B (three conditions × three replicates) was approximately USD 6.7.

Two documentary discrepancies. During preparation, we found that two design documents disagreed with the actual implementation prompts. The qualitative-validation note recorded one C1 section as an "organic, unrequested addition," but the actual prompt requested it; and a design document stated that the C2 executors used "the same prompt as C1," whereas the actual implementation used a distinct executor prompt. We resolved both against the ground truth (the literal prompts in the execution platform), and Line B replicates the implemented design (three

prompt generations, not two). Both discrepancies are recorded in the measurement note.

4. SIGNATURE-ROBUSTNESS RESULTS (n = 3)

All three signatures are present in 3/3 replicates. Beyond mere presence, the replication surfaced several stabilities that bear directly on the central claim.

C1: Hypothesis integration, with a stable hypothesis space. In all three replicates, the alternative hypotheses migrate from comparison into the structural frame, with increasing explicitness across replicates (in one, the operation is even given its own self-declared subsection: the alternatives are "readings at different levels of the same phenomenon," and the strategy "must operate on all three planes simultaneously"). The prompt asked the model to evaluate and compare hypotheses; it did not ask it to integrate them. The integration emerges in 3/3. Moreover, the two alternative hypotheses belong to the same two interpretive families in every replicate (an internal, labor/faculty family and an external, signal-market family) and two replicates independently reconstruct the same triadic "three-contracts" diagnosis. The hypothesis space the model opens before this case is itself a stable attractor.

C2: Operational convergence. The closing "divergence-resolution" section appears in 3/3 (with seven, five, and six arbitrated divergences, respectively), each carrying an explicit, justified position. The irreducibility clause is used correctly in all three; the genuinely undecidable is declared as uncertainty, never forced. Internal validity was confirmed by matching arbitrated readings to the actual executor inputs.

C3: Critical opening. The residual-tensions section appears as structural closure in 3/3 (five, five, and four tensions), with a systematic rhetorical marker ("mitigates… but does not resolve," repeated tension by tension). Two emergent contents are notably stable: a radical question about what replaces the certifying function recurs in all three Line-B replicates and in the original line (four appearances across four independent executions by two different prompt authors), and a "century-old brand as asset vs. liability" tension recurs in 3/3. A third emergent content (institutional sovereignty vs. external standards) is present in 3/3 but migrates structurally, leaving a residual tension in two replicates and folded into the consensus diagnosis in the third. The emergent content is stable; its structural location fluctuates. This is why the correct unit of analysis for a signature is the operation, not the section.

Transversal finding (T2). The three architectures converge on substantive content, a broken certifying contract, two-speed governance, sampled in-person verification, and employer-co-constructed external signals, but diverge on operations. Content does not discriminate against the conditions; only the closure operation does. (Relatedly, traceability scores are a function of the prompt, not the architecture: the generator pre-coding that requests labeling raises C1's granular traceability, while its absence depresses C3's, confirming that the level of this metric is prompt-driven, whereas the presence of a minimal functional traceability is architecture-driven.

This is consistent with the value-of-question / value-of-process decomposition established in Manucci, 2026b.)

5. PRE-REGISTERED BLIND RE-EVALUATION (LINE B)

T2 yields a falsifiable explanation for the original evaluation anomaly (an external judge at Δ = 0.262 against an internal judge at Δ = 0.911), and for the blind-assessment pattern recorded in Manucci (2026b): a content-oriented rubric cannot see signatures that live in operations. We tested this directly with a pre-registered protocol.

Design. A single external judge (Grok 4.3), in one clean session, with no project context. Corpus: the nine signature-bearing outputs (three per condition), anonymized (document titles and all metadata removed, bodies kept verbatim, including internal references to multiple voices, which are constitutive of the operations), coded DOC-A through DOC-I in a seeded random order. The code↔condition table was sealed before any contact with the judge.

Two passes in one session. Pass 1 (open-set, exploratory): describe in one or two sentences what operation each document's closure performs, with no categories supplied. Pass 2 (closed-set, confirmatory): given three functionally-defined operations (deliberately worded to avoid the generators' literal section vocabulary so the task cannot be solved by label matching), classify each document into exactly one, with a supporting quotation and a 1–5 confidence rating. First-output rule throughout.

Pre-registered thresholds. Global discrimination ≥ 6/9 correct (chance p = 0.042); critical C2/C3 pair ≥ 5/6 correct (chance p = 0.018). Two hypotheses were pre-registered: H-a, the judge meets both thresholds (signatures are externally legible; the original deficit was instrumental); H-b, the judge fails the critical pair even with an operational instrument (signatures structurally verifiable but not superficially legible; itself a publishable finding about multi-agent evaluation). Both outcomes were defined as usable.

6. RESULTS OF THE RE-EVALUATION

H-a confirmed, with a perfect score.

• Global classification: 9/9 (pre-registered threshold ≥ 6/9; probability under chance $(1/3)^9 = 5.08 \times 10^{-5}$).

• Critical C2/C3 pair: 6/6 (pre-registered threshold ≥ 5/6; probability under chance 1.37×10^{-3}).

The confusion matrix is a perfect 3×3 diagonal: zero misclassifications of any kind, including within the same-genre C2/C3 pair whose operations are inverse (close-with-position vs. open-without-resolution). Mean confidence was 4.78/5; the only two ratings of 4 fell on the two documents with the weakest quote anchoring; the judge was, additionally, well calibrated.

An unanticipated finding in Pass 1. With no categories supplied, the open-set descriptions implicitly grouped the nine documents into the three correct triads, without the judge knowing triads existed. The C2 documents drew spontaneous arbitration vocabulary ("resolves divergences… takes positions," and for one of them even the irreducibility clause); the C3 documents were uniformly described as "dialectical integration culminating in an institutional imperative"; the C1 documents as a reframing into "institutional contracts" with a multilayer strategy. The C2 signature was the only one fully legible in the open condition; C1 and C3 produced consistent within-triad descriptions that were less operational in vocabulary, which is precisely what the closed instrument supplies. The clustering was already there; the instrument names it. This result directly confirms, under pre-registered conditions, the structural-signature conjecture recorded in Manucci (2026b, §5.4).

Citation-anchoring control. Seven of nine supporting quotations were correctly anchored (verified literal); two were transplanted. The quote offered for one C3 document belonged to its sibling C3 replicate (a within-C3 transplant), and the quote offered for one C1 document belonged to a C3 document (a cross-condition transplant, C3 → C1). In both cases, the classification is correct on the document's own content, so the scoring is unaffected. The control nonetheless exposes a quotation-anchoring weakness and should not be read as evidence that quote errors stayed within condition; what it does show is that a misattributed quote did not drag the affected document toward the wrong condition, classification tracked the document's own content, not the displayed quote. (Both transplanted passages originate in C3 documents.)

7. DISCUSSION

The signatures are architectural. The full causal chain is now closed end to end: independent adversarial generation of the prompts (§3) → fixed-architecture execution (§3–4) → independent blind evaluation (§5–6). A signature that (a) survives complete replacement of its prompts by an adversarial generator across three replicates and (b) is recovered 9/9 by an external judge cannot reasonably be attributed to the wording of any prompt or to the internal judge's idiosyncrasy. It is a property of the composition.

The instrument-vs-signature thesis. The single most exportable result is the contrast between two evaluations by the same external judge (Grok 4.3): $\Delta = 0.262$ with a content/quality rubric, and 9/9 with an operation rubric. The signatures did not become more visible; the question did. Quality-oriented evaluation fails here not by noise but structurally, because the architectures were engineered to differ in how they handle the problem space, not in the substantive conclusions they reach, and content is what a quality rubric measures. This resolves the open question recorded in Manucci (2026b, §6.5) about whether process-dependent properties are legible to blind assessment: they are, but their legibility is a property of the instrument, not of the reader. The practical corollary for the field: standard benchmarks for multi-agent systems that score

answer quality are, by construction, blind to genuine architectural differences in reasoning. Detecting those differences requires instruments that interrogate the operation on uncertainty, not the answer.

Connection to inferential-uncertainty management. The three operations are three stances toward uncertainty: C1 absorbs competing readings into a single structure, C2 closes them with a decision, C3 holds them open as declared residue. C3 is, in effect, a verifiable non-closure protocol, an operational counter to premature closure, and the most direct bridge to the program's underlying framework, which treats uncertainty as a generative space rather than a defect to be eliminated.

Limitations. (1) A single judge and a single session: the result establishes legibility for this judge with this instrument, not universal legibility. (2) The comparison with the original $\Delta = 0.262$ baseline is qualitative: the two scores sit on non-commensurable scales (a quality rubric vs. an operational classification). The external judge is the same in both evaluations (Grok 4.3), so holding the judge fixed isolates the rubric as the changed variable; the scores are therefore read as indicative of the instrument inversion, not as a within-scale effect size. (3) The closed-set format makes the confirmatory task easier by design; the Pass-1 spontaneous clustering substantially mitigates this. (4) $n = 9$ over a single institutional case, with the C3 device preserved in its briefing by recorded decision (we verified the stability of the device and the emergence of its contents under an independent prompt, not the spontaneous emergence of the device itself); the pre-registered thresholds compensate with stringency, and the observed result exceeds them by the maximum possible margin. (5) The seven control prompts were generated within a single AI Studio session rather than in isolated sessions, so the generator held its earlier generations in context; as argued in §3, this works against the signature differentiation rather than for it; it does not threaten the result, but the prompts were not generated under mutual independence. At the program level, this study adds a second analytical case to the one studied in Manucci (2026b), partially mitigating the single-case limitation declared there, though within-study multi-case validation remains pending.

8. IMPLICATIONS AND FUTURE WORK

The immediate conceptual deliverable is an operation-based protocol for epistemic-quality evaluation of multi-agent reasoning: classify by closure operation, with anchored justification, blind and pre-registered. The natural extensions are a second case outside this domain (to test whether the stable hypothesis families and emergent contents are case-specific), multiple independent judges (to move from legibility-for-one to legibility-in-general), a harder variant in which no architectural device is preserved in the briefing and each prompt is generated in an isolated session (to test spontaneous emergence under mutual independence), and the connection of closure operations to factual calibration under ground truth (whether architectures

whose signature includes declared irreducibles mark unsupported claims as uncertainty rather than asserting them) which is the next study of this program. Each is a discrete, low-cost study on the present infrastructure.

This study follows the verification model established by the program: public-control results, protocols, and instruments are fully disclosed; proprietary orchestration internals are not. The program's verification environment is available at https://verification.codhz.com, and the companion materials for the present study (the pre-registered protocol with the literal evaluation instrument, the unsealed correspondence table, the judge's complete responses, and the per-chain metrics) are deposited alongside this report under the same model.

APPENDICES (SUPPLEMENTARY MATERIAL)

• **Appendix A**. Briefings and generated prompts (Line B). Function-level briefings for the C1 analyst, the C2 executor and synthesizer, and the C3 four roles, together with the verbatim first-generation prompts (Generations 1–7) and the recorded generator pre-codings. Held in the project's auditable generation log and deposited as supplementary material.

• **Appendix B**. Pre-registered protocol and instrument. Full protocol with the literal Pass 1 and Pass 2 prompts, anonymization rules, and decision thresholds, time-stamped prior to any contact between corpus and judge.

• **Appendix C**. Unsealed correspondence table and judge responses. The seeded code↔condition map (DOC-A…DOC-I), the Pass 1 and Pass 2 responses in full, and the per-item scoring against the key.

• **Appendix D**. Dataset and metrics. Per-chain token counts, latencies, and costs for all twenty-seven outputs.

REFERENCES

• Manucci, M. (2026a). Epistemological Orchestration over Language Models: A Functional Architecture with Preliminary Evidence. Zenodo. https://doi.org/10.5281/zenodo.19287671

• Manucci, M. (2026b). Epistemological Sequencing as an Architectural Principle: Structural Output Properties of Multi-Regime Inference over Large Language Models. Zenodo. https://doi.org/10.5281/zenodo.20121190

This report is self-contained and discloses no proprietary orchestration logic. The control architectures C1–C3 and their Line-B prompts are fully described; the CODHZ architecture (C4) is reported separately. Prepared by CODHZ Research Laboratory, June 2026.