

Structured Representation and Epistemic Integrity in AI Reasoning

Lynn Walker | WinMedia

Technical paper · Version 1.0 · 5 October 2026 · WMR-WHP-016

Research-stage report. Independent peer review and production qualification are not claimed.

Abstract

Explicit representation may change how an AI system accounts for its reasoning without improving the quality of its proposed solutions. This paper examines that distinction through IIE-009, an isolated prompt-scaffold ablation interpreted in the IIE-010 determination. Twelve frozen problem blocks across three domains were tested under ordinary reasoning (A), structured representation (R), and representation plus transformations (T). Four solver repetitions per problem and condition produced 144 artifacts, each receiving three blinded evaluator judgments. The primary endpoint was Solution Quality; Epistemic Integrity and Novelty were secondary endpoints. Representation improved measured Epistemic Integrity relative to ordinary reasoning by +0.701 on a 0–4 scale (95% bootstrap interval [+0.521, +0.875]; exact paired $p=0.00098$). The earlier Solution Quality advantage was not replicated, and neither an incremental transformation advantage nor a Novelty advantage was established. These findings support a limited claim about evaluator-rated rigor, failure-mode disclosure, and boundary candor. They do not establish improved invention outcomes or software efficacy. The paper develops the distinction between solution performance and epistemic accountability, specifies the evidential boundary, and proposes a focused follow-up design. [1–4]

Keywords: structured representation; epistemic integrity; AI reasoning; prompt scaffolding; ablation; evaluation.

1. Introduction

A reasoning system can offer a plausible solution while leaving its assumptions, uncertainties, and failure conditions difficult to inspect. Conversely, a system can describe those boundaries carefully without producing a better solution. These are distinct properties of a response. Evaluating a structured reasoning intervention therefore requires more than asking whether its output appears organized: it requires separate measurements of solution quality and the way the response represents the limits of its knowledge.

IIE-009 tested whether an explicit invention-state representation scaffold improved Solution Quality over ordinary reasoning, and whether adding design-state transformations improved it beyond representation alone. The three conditions formed an ablation: ordinary reasoning supplied the comparator, representation supplied the first intervention, and transformations supplied the incremental intervention. IIE-010 subsequently issued the interpretation of the executed trial. This paper develops the qualified aggregate report into a fuller technical account; it introduces no new solver runs, judgments, or statistical reanalysis. [1, 2]

The central finding concerns a secondary endpoint. Representation improved measured Epistemic Integrity, while the primary Solution Quality advantage was not replicated. Preserving both findings matters. Describing the trial solely as a success would substitute a different outcome for the original primary

question; describing it solely as a failure would discard a supported, narrower result. The contribution is a bounded empirical distinction between better rated epistemic conduct and better rated solutions, together with an account of what would be needed to test that distinction more strongly. [1, 3]

2. Study design and execution

2.1 Intervention structure

The experiment compared three independently requested prompt conditions. The descriptions below identify intervention classes rather than disclose the frozen prompts or held-out problems. The prespecified R–A contrast tested the representation scaffold relative to ordinary reasoning. The T–R contrast tested the addition of transformations to the representation condition. The latter is an incremental comparison: it does not compare transformations in isolation with an unstructured baseline. [2]

Table 1. Intervention classes and contrasts

Condition	Intervention class	Prespecified comparison
A	Ordinary reasoning	Comparator for R
R	Explicit invention-state representation scaffold	R–A: representation increment
T	Representation plus design-state transformations	T–R: transformation increment

The design comprised 12 frozen problem blocks across three domains, with all three conditions and four independent solver repetitions in each problem-condition cell. This yielded 144 completed solver requests. Each resulting artifact received three blinded evaluator slots, yielding 432 completed valid judgments. These counts describe execution volume; the paired inferential sample remained the 12 problems. The repetitions and evaluator slots supplied repeated measurements within those blocks. [2]

2.2 Models, isolation, and deviations

The direct-API harness used gpt-5.6-sol as the solver and gpt-5.6-luna as the blinded evaluator. The three evaluator slots per artifact were repeated judgments from the same evaluator model, rather than judgments from three distinct model families. The isolated, single-shot procedure excluded a continuing interactive workflow from the tested intervention. Findings therefore concern the response to a scaffold under the recorded procedure, rather than iterative reasoning with an evolving workspace. [1, 2]

The evaluator selection was a recorded deviation: gpt-5.6-luna replaced the package’s default gpt-5.6-terra by founder authorization before unblinding. A transient connection interruption during evaluation was handled by resuming missing slots without overwriting completed judgments. The completed analysis accounting contains no invalid or excluded judgment slots. Reporting these details preserves the actual execution identity and avoids presenting the package default as the model that generated the results. [2]

3. Endpoints and analysis

3.1 Separate outcome constructs

Solution Quality was the primary endpoint, on a 0–16 scale combining constraint satisfaction, mechanistic plausibility, utility, and testability. Epistemic Integrity and Novelty were secondary endpoints, each on a 0–4 scale. In the qualified interpretation, Epistemic Integrity concerns measured rigor, failure-mode disclosure, and boundary candor. It is a rating of the response under the study’s evaluation procedure, rather than an independent guarantee that every assertion, assumption, or proposed mechanism is correct. [1, 2]

The separation of endpoints is essential to interpretation. A higher integrity rating cannot be substituted for a higher quality score, and a plausible mechanism cannot be assumed to be novel. The aggregate report does not supply component-specific Solution Quality effects, so it does not establish a distinct improvement in utility, feasibility, or any other individual component. Nor does an integrity increase establish that a response would produce safer decisions or better human outcomes. Those require additional measurements. [2–4]

3.2 Aggregation, contrasts, and uncertainty

Scores were first averaged across the three evaluator slots for each artifact. Artifact scores were then averaged across the four solver repetitions within each problem-condition cell. For each endpoint, the analysis computed R–A and T–R within each problem and summarized those differences across the 12 problems. This hierarchy prevents the 432 judgments from being treated as 432 independent problem-level observations. The relevant contrast is between condition means for the same problem. [2]

Uncertainty was reported through problem-level bootstrap 95% intervals, with exact paired sign-flip p-values. The paper reproduces those statistics as recorded rather than recalculating them from rounded table entries. The aggregate methods identify the analysis unit and inferential procedures, but do not expose the complete execution parameters of the bootstrap or the underlying problem-level judgment records. This limits independent verification from the manuscript alone. [2, 3] The disclosed aggregate methods do not specify a multiplicity-adjustment procedure; the p-values are reproduced as reported.

The prespecified practical thresholds applied to mean Solution Quality improvement: +1.0/16 was a signal for further study and +2.0/16 was substantial. These thresholds answer a different question from a p-value: whether an estimated gain reaches the magnitude considered useful for the original quality claim. They were not thresholds for Epistemic Integrity or Novelty. The secondary integrity finding is reported alongside both quality contrasts and both novelty contrasts, preserving the endpoint hierarchy. [2, 3]

4. Results

All six endpoint-by-contrast results are shown below. Every row uses 12 paired problem blocks. Direction counts indicate positive, negative, and zero problem-level differences, respectively. Intervals and p-values are reproduced from the qualified aggregate derivative. [3]

Table 2. Complete endpoint-by-contrast results

Endpoint / contrast	Mean change	95% interval	Paired p	Direction + / - / 0
Solution Quality R-A (0-16)	-0.146	[-0.437, +0.146]	0.38721	4 / 8 / 0
Solution Quality T-R (0-16)	-0.069	[-0.306, +0.194]	0.64844	4 / 7 / 1
Epistemic Integrity R-A (0-4)	+0.701	[+0.521, +0.875]	0.00098	11 / 0 / 1
Epistemic Integrity T-R (0-4)	+0.104	[-0.007, +0.201]	0.12305	9 / 2 / 1
Novelty R-A (0-4)	-0.021	[-0.083, +0.035]	0.68750	2 / 3 / 7
Novelty T-R (0-4)	+0.000	[-0.076, +0.076]	1.00000	3 / 4 / 5

4.1 Solution Quality

Neither contrast established a Solution Quality improvement. R-A had a mean of -0.146/16, and T-R had a mean of -0.069/16. Both intervals crossed zero; neither estimate reached the prespecified +1.0 investigation signal or +2.0 substantial-effect threshold. The reported upper interval bounds were also below those thresholds. IIE-010 accordingly issued EARLIER_ADVANTAGE_NOT_REPLICATED for Solution Quality. The earlier advantage is background to the determination, rather than additional evidence pooled into this clean trial. [3]

This outcome should not be recast as demonstrated equivalence or noninferiority. The trial did not establish a quality gain, but a nonsignificant contrast does not by itself prove identical performance or prove that no degradation is possible. No equivalence or noninferiority result is reported here. The estimates and intervals concern this frozen experiment; they do not show that every possible representation or transformation must have zero effect. [3, 4]

4.2 Epistemic Integrity

Representation produced the strongest reported contrast: R-A increased Epistemic Integrity by +0.701/4, with a 95% interval of [+0.521, +0.875] and $p=0.00098$. Eleven problem blocks had positive differences, none had negative differences, and one had a zero difference. IIE-010 issued EPISTEMIC_GOVERNANCE_ADVANTAGE_SUPPORTED for this contrast. The supported claim is an increase in evaluator-rated integrity under the tested conditions, with the problem-level direction pattern providing useful context for the mean. [3]

Adding transformations yielded a smaller integrity estimate, +0.104/4, with interval [-0.007, +0.201] and $p=0.12305$. Although nine problem-level differences were positive, the reported analysis did not establish an additional transformation effect. This leaves the incremental benefit unresolved; it neither establishes a gain nor proves that transformations cannot contribute in another setting. The clear representation contrast should therefore be kept distinct from the weaker transformation contrast. [3, 4]

4.3 Novelty

Neither contrast established a Novelty advantage. R-A was -0.021/4, with interval [-0.083, +0.035]; T-R was +0.000/4, with interval [-0.076, +0.076]. The zero rounded mean for T-R is an aggregate result, not identical performance on every problem: its direction counts were three positive, four negative, and five zero. No claim of increased creativity or invention novelty follows from these ratings. [3, 4]

5. Interpretation: accountability and performance

The results distinguish two possible purposes of representation. One is to improve the proposed solution; the other is to make the response's reasoning boundaries more explicit and assessable. Only the latter received positive support on the reported measures. This distinction can matter in research and design work, where a useful response must be inspected for its constraints and uncertainties even when its central proposal is unchanged. That practical relevance is a motivation for further testing, rather than a measured deployment benefit.

A possible explanation is that an explicit representation scaffold encourages the solver to state limitations or examine failure conditions more visibly. The experiment does not identify that mechanism. It observes the scored response to a bundled prompt intervention, without separately varying each representational component or independently checking every disclosure. Higher ratings could reflect better underlying reasoning, more explicit presentation, closer conformity to evaluator expectations, or some combination. Discriminating these explanations requires an additional study.

The result also does not establish that the tested representation is uniquely necessary. The reported comparison was with ordinary reasoning, rather than with every alternative structured prompt. A matched generic scaffold could help determine whether the integrity difference depends on the particular representation or on the broader instruction to organize and disclose reasoning. This is a causal question left open by the present A/R/T design, rather than a conclusion against the representation.

The evidence concerns prompt scaffolds. IIE-009 did not exercise a bounded Design-Space Calculus executable, a bounded Novum executable, their novum-dsc/0.1 exchange contract, or their composition. Technical qualification of those separate assets cannot convert the prompt-level observation into an efficacy result for software. Conversely, this trial's quality result does not invalidate a separately tested implementation property. Each claim requires evidence from the intervention and outcome actually tested. [4]

6. Threats to validity and reporting limits

External validity is restricted by the 12 frozen problems, three domains, one solver model, one evaluator model, and single-shot procedure. Multiple repetitions improve characterization within a problem; they do not create new domains or independent model populations. Three evaluator slots similarly provide repeated ratings without evaluator-model diversity. Human assessment, different model families, interactive workflows, deployed systems, and real-world invention outcomes were not tested. [2, 4]

Construct validity is especially important for Epistemic Integrity. Evaluator-rated candor can be informative without guaranteeing truthful uncertainty estimates or correctly identified failure modes. A response may disclose a limitation that is irrelevant, omit a consequential limitation, or sound cautious while making an unsupported claim. The aggregate report does not supply an independent ground-truth audit of these possibilities. It also does not report judge-agreement coefficients or component-level rating distributions. These omissions constrain assessment of rating reliability and of the source of the integrity difference.

Blinded evaluation reduces the evaluator's direct access to condition identity as described in the study procedure. It does not establish that textual features could never reveal a condition, or that the evaluator had no preference for structured exposition. The aggregate record does not include a reported test of whether

conditions could be inferred from the artifacts. Using a single evaluator model leaves this potential interaction between treatment style and rating behavior unresolved. [2, 4]

The prompt interventions are bundled conditions. The available aggregate methods do not specify enough detail to isolate the contribution of each representation element, additional instruction content, or response length. They also do not expose every generation parameter or the complete scoring implementation. This is a limitation of the disclosed record, not evidence that those factors were uncontrolled. The manuscript consequently avoids component-level causal claims and does not present itself as a complete external replication protocol. [2]

Finally, a positive secondary endpoint must remain visible within the complete outcome pattern. The primary quality finding was negative with respect to the proposed advantage, and the additional transformation and novelty benefits were not established. The paper reports all six contrasts and does not combine their scales into a new success metric. Its interpretation follows the qualified claim limits, rather than expanding the integrity result into productivity, creativity, general intelligence, or executable efficacy. [1, 3, 4]

7. A focused follow-up

The most direct next experiment would test the specificity and validity of the integrity effect. A new task set could compare ordinary reasoning, the present representation class, and a matched generic structured prompt, while retaining a prespecified quality endpoint and explicit integrity criteria. A transformation arm would remain useful if its incremental contribution were still a research priority. All such conditions, endpoints, and practical thresholds should be fixed before outcome inspection. These are proposals, not completed extensions of IIE-009.

The follow-up should add evidence beyond repeated ratings from one evaluator. Independent evaluator families and human reviewers could assess whether the result persists under different judgment processes. Where tasks permit, a separate audit could check the accuracy of stated assumptions, uncertainties, and failure conditions against known constraints. This would help distinguish genuine epistemic accuracy from a presentation that receives higher integrity scores. Reporting reviewer agreement and component-level effects would make the interpretation more informative.

Interactive and tool-assisted reasoning should be examined separately rather than inferred from the single-shot study. Such a design would need to specify how state is retained, what information is available, and whether transformations alter decisions or only explanations. For the immediate question, however, the smallest useful follow-up is the matched-scaffold comparison with independent integrity checks. It targets the principal uncertainty left by the positive result without multiplying unrelated development goals.

8. Evidence provenance and availability

The empirical source is the qualified IIE-009/010 aggregate derivative at commit 9e1407d3ef89a3705d5a0e98f9f1e1d94513e0db. Its seven-file candidate manifest has SHA-256 c1dc1b0f75f56bdbb01615795dadcc2bfa46f5195896f4ec3a0ae3f3675f1dfd. This manuscript develops the existing report as pinned at commit 53b2c79c91a2329205988cc9937cb2359f260969 and preserves its result values and claim limits. The issued IIE-010 determination supplies the named dispositions. [1–5]

Executed records have durable private custody, with complete source authority at commit ac82917a9c6dfcdf862239390ed043fe5c05f483. Internal custody and public reproducibility are distinct: the

held-out tasks, frozen prompts, raw outputs, judgment records, and blind mapping are not included in this manuscript. The cited records are held in a private repository; the references identify them without providing public access. Accordingly, the paper permits assessment of the aggregate argument but does not enable independent external reproduction of the original trial. [2, 4, 6]

9. Conclusion

In this controlled prompt-scaffold ablation, structured representation improved measured Epistemic Integrity relative to ordinary reasoning. The trial did not replicate the earlier Solution Quality advantage, establish an incremental transformation benefit, or establish a Novelty benefit. The defensible contribution is therefore a bounded observation about epistemic accountability in evaluated responses. Separating that observation from invention performance preserves the value of the finding and makes the next test clearer: determine whether the integrity effect survives a matched structured comparator and independent checks of what the response actually gets right. [1, 3, 4]

References

- [1] Lynn Walker. Structured Representation and Epistemic Integrity in a Controlled AI Reasoning Ablation. IIE-009/010 qualified report draft. Private research record; not publicly accessible.
- [2] Intelligent Instruments Foundry. IIE-009/010 aggregate derivative: Methods. Private research record; not publicly accessible.
- [3] Intelligent Instruments Foundry. IIE-009/010 aggregate derivative: Aggregate results. Private research record; not publicly accessible.
- [4] Intelligent Instruments Foundry. IIE-009/010 aggregate derivative: Scope, nonclaims, and separate identities. Private research record; not publicly accessible.
- [5] Intelligent Instruments Foundry. IIE-010 clean trial results and causal determination. Private research record; not publicly accessible.
- [6] Intelligent Instruments Foundry. IIE-009 executed-evidence custody authority. Private research record; not publicly accessible.

Author note. Prepared with AI-assisted drafting from the existing IIE-009/010 report and its primary evidence records. This manuscript introduces no new experimental runs or statistical analyses.