Coherence Cost in a Summary World Model

Separating an AI scientist's synthesis-accuracy gap into task difficulty and coherence burden, demonstrated on a controlled surrogate of the Kosmos architecture across a capability ladder.

Embedded Risk Analytics · A Fathom case study · August 2026

An AI scientist that coordinates its agents through a summary-based world model reports a specific pattern in its own accuracy: cross-trajectory synthesis lands well below single-source data analysis and literature validation. Kosmos, the system published by Edison Scientific, states this split directly, reporting synthesis near 57.9 percent against 85.5 and 82.1 percent for data analysis and literature validation. A gap of this shape has two possible causes, and they call for different fixes. Either cross-trajectory synthesis is intrinsically harder, or the gap marks where the system pays to keep its work consistent with a world model it has been summarizing as it goes. A single accuracy number cannot separate the two. This briefing measures the separation on a controlled surrogate of the summary-world-model architecture.

One result stands out. At full legibility, where the record carries no summarization loss, every model we tested still falls well short of a perfect executor on the synthesis family, and the shortfall shrinks as the model improves without closing. A stronger base model reduces the coherence cost it carries and does not remove it. Asking the model to verify its own committed work does not remove it either, and a single bad committed value cascades through everything that later depends on it.

Embedded Risk Analytics models coherence as a capacity condition.

Embedded Risk Analytics treats agent coherence as an information-theoretic condition. A system that maintains a world model over a long run holds a record of what it has already established within a finite capacity. That capacity must cover two loads at once: the exogenous difficulty of the task, written H_{ext}, and the coherence cost of keeping later steps consistent with the record the system's own outputs produced, written B_{rho}. Coherence holds while capacity covers both, expressed as $C \geq H_{ext} + B_{rho}$. The two loads differ in kind. Coherence cost grows with the horizon and with how tightly the system's own conclusions depend on one another, while exogenous difficulty stays bounded. A single accuracy number combines the two terms and reports neither.

A summary world model sits on the internal branch of this condition. It does not act on the phenomenon it studies, so its coherence cost is the cost of staying consistent with its own lossy record rather than with an environment it has reshaped. The Fathom harness reads that cost from the model's own reported state, attributes the synthesis gap to coherence cost or to task difficulty against a matched control, and reports which term binds.

A controlled surrogate of the summary-world-model architecture isolates the two causes.

The demonstration uses a derived-quantity ledger that reproduces the summary-world-model shape and grades every step against a known answer. The environment generates the facts each cycle, so the ground truth is fixed in code, and the model performs only the coherence-maintenance role. Each cycle the model receives the world model at a controllable legibility, applies that cycle's updates, and reports the current totals. Some updates define one quantity in terms of others, which forces the model to read and combine values from its own world model rather than copy a single figure forward. Because the reported totals are numbers with a known answer, every measurement is computed on real vectors.

Two knobs drive the study: legibility, the fraction of the record the model carries in full rather than in summary, and the model itself. The instrument reads the reflexive loading, the degree to which the world-model error aligns with the accumulated state the model is tracking. Coherence cost produces that alignment; a disturbance of the same size drawn independently of the model's own record does not, and that matched-difficulty control is the separator. A faithful-executor floor runs alongside every model at no cost, so the gap between a model and that floor isolates the coherence cost the model itself adds. The instrument was validated on synthetic data with no model spend before any model was called.

The summarization deficit carries the coherence-cost signature, and a matched control does not.

As legibility drops, the record loses more of the model's accumulated content, and synthesis accuracy falls with it. The endogenous loading rises toward one, meaning the world-model error aligns fully with the state the model is tracking, while the matched-difficulty control stays near zero at every setting. A disturbance of equal size that is drawn independently of the model's record does not carry the signature. This pattern holds identically across a low-cost model and a frontier model, so the structural coherence cost is a property of the architecture and not of the model inside it. At full legibility there is no summarization loss to carry, and the small residual loading there reflects the model's own error rather than the record, which is the quantity the next section isolates.

Table 2. Reflexive loading and synthesis accuracy by legibility, DeepSeek-chat. The matched-difficulty control loading stays near zero throughout. Below full legibility the endogenous loading rises toward one as the summarization loss aligns with the accumulated state.

Legibility λ	Endogenous loading	Matched-difficulty control	Synthesis accuracy
1.0	-0.73	-0.02	0.61
0.6	+0.99	-0.12	0.45
0.4	+1.00	-0.15	0.42
0.2	+1.00	-0.17	0.42

The coherence cost that remains at full legibility falls with capability and does not close.

At full legibility the record is complete, so a perfect executor scores perfectly and any shortfall is the model's own coherence cost. Every model falls well below that floor on the synthesis family. The frontier model carries the smallest gap and the highest accuracy, and the gap shrinks with capability without closing. A stronger base model buys the coherence cost down and leaves a floor it does not reach. The trend is not strictly ordered across the middle of the range, and the smallest model is omitted because it could not hold the reporting format reliably enough to read, which is itself a distinct failure from the reconciliation cost the larger models show.

Table 3. Synthesis accuracy at full legibility against the faithful-executor floor. The coherence-cost gap is the floor minus the model. Eight seeds, twenty cycles. Llama-3.1-8B is omitted for unreliable formatting.

Model	Synthesis accuracy	Faithful-executor floor	Coherence-cost gap
DeepSeek-chat	0.61	1.00	+0.39
Llama-3.3-70B	0.48	1.00	+0.52
Claude Sonnet-4.5	0.72	1.00	+0.28

The coherence cost takes the form of a compounding cascade, and self-verification does not remove it.

The model-intrinsic error is not random noise. The model over-reports derived quantities, and because those quantities feed later definitions, the error grows faster than the truth and spreads. To measure the spread we commit one bad value into the world model partway through a run, hold a paired baseline with identical inputs, and count how many quantities the bad value poisons as the density of definitions rises. With no definitions the damage stays with the one quantity. As dependency density rises, a single bad commit poisons most of the ledger, and the model widens the reach beyond the dependency graph itself, because a model confused by the corrupted state introduces fresh errors in quantities that do not depend on the seed. This is the committed-value cascade the Kosmos paper describes in section 2.3.1, where a value stored incorrectly propagates through variant identification and column collisions into work that is not obviously downstream.

Table 4. Downstream reach of one bad committed value by dependency density, DeepSeek-chat. Model reach is the count of poisoned quantities; the floor is the reach carried by the dependency graph alone.

Dependency density	Quantities poisoned (model)	Dependency-graph floor
0.0	1.1	1.0
0.2	3.1	1.1
0.4	5.8	3.0
0.6	8.8	4.3

Self-verification does not remove the cost. Given a pass each cycle to re-derive and correct its own committed totals, the low-cost model's gap moved from 0.39 to 0.41 and the frontier model's from 0.28 to 0.31, both slightly worse. The model checking its own work carries the same bias into the check, so the check reproduces the error it is meant to catch. A control in which the self-check was accurate did remove the cost, which confirms the instrument would register a repair when one is present. The reading is that a coherence cascade is caught at the source, by a reader that holds the committed state against ground, and not by asking the model to look again.

The demonstration is a controlled surrogate with a defined scope.

The result is established on a controlled surrogate and its scope is stated. The surrogate shares the summary-world-model architecture and is not Kosmos, so the inference to the real system runs by architecture and not by proof. The structural coherence cost is measured across a low-cost model and a frontier model at eight seeds and twenty cycles; the capability trend is clean at the frontier and not strictly ordered across the middle of the range. The model-intrinsic error runs in the over-reporting direction, which the pre-registration did not fix in advance and which held consistent across three models. The smallest model could not hold the reporting format, which bounds the low end of the capability ladder. The demonstration establishes the decomposition, the capability trend, the cascade, and the self-verification result on the surrogate. It does not yet run on Kosmos itself.

Two evaluations on Kosmos would extend the result to production.

Two evaluations on Edison's own system would carry the result onto the production agent. The first runs the decomposition on real Kosmos run traces and returns where the difficulty and coherence costs fall, scored against Kosmos's own synthesis-accuracy metric. Whether the in-run signal separates the two on real traces is a measurement that requires Edison's graded runs. The second is the propose-next-tasks loop, where each cycle queries the world model for the next cycle's tasks with no signal for whether the

last cycle was hard or incoherent. Whether a coherence read at that seam sets re-ground against escalate more accurately than the current loop is a measurement Edison is positioned to run. Embedded Risk Analytics supplies the method and the open evidence, and Edison holds the reference standard.

Embedded Risk Analytics is seeking design partnerships.

Embedded Risk Analytics works with a small number of teams running long-horizon agents. We measure where an agent loses coherence, attribute the cause, and apply the matched repair. Email us: contact@embeddedriskanalytics.com.

Method.

The surrogate is a derived-quantity ledger of ten quantities over twenty cycles, where each cycle applies base updates and definitions that set one quantity to a combination of others, resolved in code against the prior state. The models were DeepSeek-chat, Llama-3.3-70B, and Claude Sonnet-4.5, run through OpenRouter, with Llama-3.1-8B attempted and omitted for unreliable formatting. Each cell pools eight seeds. Legibility sets the fraction of the world model carried in full against summarized away. The reflexive loading is the slope of the world-model error along the accumulated-state direction, read against a matched-difficulty control drawn independently of the model's record, with a faithful-executor floor for the model-intrinsic gap. The instrument was validated on synthetic data with no model spend, and every model read is a deterministic measurement over the model's reported state. Results are written per seed to a single store, and the reads recompute from the stored trajectories without re-spending.