ANATOMY OF A JUMP

Anonymous authors
Paper under double-blind review

ABSTRACT

Can a language model infer an explanatory concept that the observations do not name, rather than fit a familiar law to the state it is shown? We test this in 48 procedurally generated physical worlds. At the restructuring level, the best tested local patch stays below $R^2 = 0.70$ on the certification samples while the true law reaches 1.000; an ideal observer selects the true structural class after one rendered trajectory when that class is supplied among three candidates. Twelve model configurations face a capability check, and two run the full ladder, producing 1,872 answers, although only the frontier arm passes the check. The paper presents three results. First, all 893 parseable stated laws use only position, velocity and time, but the answer format required those variables. Without that restriction, four of seven models still produce no uncued law reading additional state; the other three produce one each, two by split votes. A cue elicits hidden state in 60 of 504 laws, none the world's governing law. Second, in 1,223 of 1,336 scored answers, the model's numerical prediction beats its own stated law, including 398 of 445 at retrieval, where the correct law fits the format. Third, a four-lag linear autoregression fitted to the same fifty prompt rows beats the frontier model on 124 of 153 powered restructuring trials, with roughly eighteen times smaller median error. What this control licenses is a narrower rule: predictive accuracy can support a world-model inference only where ontology-free predictors fail on the same regime. One of three families qualifies over the swept thresholds; its advantage has the same sign on fresh worlds, but both intervals include zero and the qualifying partition changes. Mechanistic probes at 1.7B yield bounded null results.

1 INTRODUCTION

A language model's ability to produce a new explanatory concept, rather than a recombination of concepts it encountered in training, has been questioned more often than it has been measured. A standard critique holds that the transition from empirical data to new axioms requires grounding that prompting alone cannot supply; the counterargument treats discovery as inductive compression, which learning systems already perform (Zahavy, 2026). We need a measurement that separates those possibilities.

The core problem is contamination: a real discovery and its derivation may already have appeared in a model's training data. Even when a benchmark changes the parameters or the functional form of a familiar law to prevent recall, the underlying explanatory framework can stay the same. The model is still fitting a function of the observed state (Zheng et al., 2026; Chen et al., 2025; Koblischke et al., 2025). We built a task where the correct move is instead to abandon the explanations those observations invite. Its laws are procedurally generated; the model sees their rendered consequences, not their definitions. We check exact strings from those renderings in four open corpora rather than assuming that generation alone removes contamination (Liu et al., 2024a). Novelty builds like a ladder: retrieval uses standard Newtonian physics, recombination changes familiar terms, and restructuring is where the observed state no longer contains everything needed for the governing law. At that top level, a hidden body raises the dynamical order, a hidden mode changes the force at the same observed state, or a delayed force depends on the past. In each accepted restructuring world, the best member of our tested local patch class remains below $R^2 = 0.70$ on the certification samples while the true class exceeds 0.98. An ideal observer sees the same rounded trajectories and selects the true structural class after one trajectory, but only because that class is supplied among

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

three candidates. The data distinguish the candidate explanations; generating the missing candidate is still the model's task.

The prompt provides the evidence, and the model must formulate a hypothesis from it. Twelve arms from six model families face a basic capability floor. A 1.7B open model and one frontier model run the ladder to full depth, producing 1,872 answers, although only the frontier arm passes that floor. This distinction matters for the mechanistic results below: a null at 1.7B does not settle what a capable frontier model would represent.

The paper presents three key results, with the third putting the first two into perspective. First, models rarely look beyond the variables supplied in the prompt. All 893 parseable stated laws use only x, v and t, but the requested function was explicitly restricted to those variables and constants; that count is therefore compliance with the format, not an independent test of what models could propose. The unconstrained prose names a hidden driver in only 7 of 1,872 answers. When the law restriction is lifted without a cue, four of seven models still produce no law reading additional state; the other three produce one each, two by split votes. An explicit cue elicits hidden state in 60 of 504 laws, although none expresses its world's governing law (Appendix D.2). Second, models predict numbers better than the mathematical laws they state. Their own numerical tables beat their stated laws in 1,223 of 1,336 scored answers across twelve arms and six model families, including 398 of 445 at retrieval, where the correct law could have been written in the requested format (Section 7.1). Third, a simple control beats the model's predictions. A four-lag linear autoregression fitted only to the fifty rows shown in the prompt beats the frontier model on 124 of 153 powered restructuring trials where Newtonian fits fail, with roughly eighteen times smaller median error (Section 7.2). The gap between a model's prediction and its law is real, but these worlds are easier to predict than that gap alone suggests.

What that failure licenses is a rule, and this rule is the paper's primary methodological contribution: outperforming persistence at the next step supports an inference about an internal world model only where predictors that do not use the proposed ontology fail on the same regime. Our three restructuring families have ontology-free error floors of 0.003, 0.010 and 0.137, a factor of roughly fifty. Applied separately to each family, the criterion admits one regime over the thresholds we swept. The sign of the model's advantage repeats on 36 worlds generated under a new master seed against outcomes fixed before generation, but both confidence intervals include zero, and the qualifying partition itself changes. Regime qualification must therefore be swept rather than reported as a universal cutoff. The practical check requires little computation: fit a linear autoregression to the historical rows shown to the model in each regime, and ask whether the model does better there.

On a continuous novelty scale, the accuracy of a model's stated law tracks how patchable the world is across two frontier models from different labs, yet both models break while the world remains patchable (Section 5). Laws that fit one shown trajectory cover fewer other initial conditions as novelty increases (Appendix G.2). We also decompose the proposed leap into Formation, Trust, Use and an interpretability-guided intervention. Those probes run where activations are available, at 1–9B parameters, below the scale at which an arm passes this study's capability floor, and yield bounded null results (Section 5, Section 6, Section 8). In the active condition, the frontier model chooses initial states and receives counterfactual trajectories, giving it the action-controllable simulator proposed as a remedy; we detect no effect where the ontology must change (Appendix H.1). Coupled oscillators, hysteresis and delay are textbook concepts. The narrower test here is whether a model can infer an instance by leaving the state description the data suggest without a prompt naming what is missing (Appendix J.1).

2 Related work

The position this paper addresses holds that language models cannot leap from experience to new axioms because prompting lacks the physical grounding required for manipulative abduction (Zahavy, 2026). It treats the Abstraction and Reasoning Corpus as a test of logical reasoning that omits physical manipulation (Chollet, 2019; Chollet et al., 2025). Our active condition lets a model submit initial states and receive the resulting trajectory (Appendix C.2), providing an action-controllable simulator with counterfactual intervention. The experiment remains synthetic rather than embodied, which sets the limit of that comparison.

Language-model agents have been tested on counterfactual shifts of known laws (Zheng et al., 2026), interactive physics discovery with controlled priors (Chen et al., 2025), gravitational physics (Koblischke et al., 2025), and broader agentic environments (Jansen et al., 2024; Majumder et al., 2025). Symbolic regression, machine-scientist systems and language-model-guided program search recover explicit laws from data (Guimerà et al., 2020; Cranmer, 2023; Shojaee et al., 2025). Concept-driven search has also been studied without prior physical knowledge (Fang et al., 2025), while agentic systems pursue larger research tasks (Lu et al., 2024; Novikov et al., 2025). These studies do not all make the same assumption about the hypothesis space, and our claim does not require that they do. The difference measured here is specific: the observed instantaneous state suggests an $a = F(x, v)$ law, but the accepted restructuring worlds defeat the tested local patch class on their certification samples. We use symbolic regression to check that boundary; proposing missing state or memory is the behavior under test.

Language models have also been evaluated as hypothesis generators for inductive tasks (Wang et al., 2024; Qiu et al., 2024) and as abductive reasoners over natural-language observations (Bhagavatula et al., 2020). In our task, the observations come from generated dynamics whose rendered variables omit part of the governing state. A model must account for why laws fitted to the visible state fail and propose an explanation that reads more. That makes the transition out of the suggested hypothesis space the object of measurement.

Finally, language models extrapolate numerical sequences zero-shot (Gruver et al., 2023) and forecast physical trajectories in context (Liu et al., 2024b). Accurate orbital prediction can coexist with heuristics that are not Newtonian mechanics (Vafa et al., 2024; 2025), while simple linear models can beat transformer forecasters on long-horizon tasks (Zeng et al., 2023). Together these results motivate a licensing condition that we test rather than assume: prediction accuracy is evidence for an internal world model only where predictors without the proposed ontology fail. We check that condition within each regime and sweep the threshold that defines it (Section 7.2). Mechanistic neighbors are compared in Appendix B.1.

3 THE INSTRUMENT (METHODS)

There are forty-eight parameterized worlds across three levels of a novelty ladder, drawn from nine law families. Each world has ground-truth parameters θ known by construction. L0 (retrieval: harmonic motion, falling with drag and radial gravity) and L1 (recombination: a cubic spring with quadratic drag, altitude-dependent gravity with drag and a spring with Coulomb friction) contribute four worlds per family. L2 (restructuring) contributes eight worlds per family, 24 in total. Trajectories are integrated by RK4 at $\mathrm{d}t = 0.005$ and subsampled to a 0.1 s grid.

The three L2 subtypes each give a reason that the observed instantaneous state (x, v) is an incomplete account of the governing mechanism:

- In `hidden_particle`, the observed body is linearly coupled to a second, unrendered body. Eliminating that body produces a fourth-order ODE rather than the assumed second-order law.
- In `hysteretic`, the force center follows a hidden mode that relaxes toward the direction of motion, so the same (x, v) can have different accelerations at different times.
- In `delayed`, acceleration depends on the state at $t - \tau$; the governing law therefore uses history that an instantaneous description does not contain.

Each world enters the study only after thresholds fixed in the frozen specification are met. The best tested local patch, consisting of polynomials in x and v through degree 5 and a library of friction, drag, inverse-power and absolute-value terms, must reach $R^2 \leq 0.70$ on acceleration; the true parameterized law must reach $R^2 \geq 0.98$. The admission comparison fits and scores the patch on the same samples, a favorable test for that patch class. Across the 24 accepted L2 worlds, its R^2 is 0.154–0.650 while the true class reaches 1.0000. Refitting the patch on a held-out trajectory split leaves every L2 world below the bar (R^2 −0.552–0.656). Four candidate worlds failed admission and were regenerated. These results exclude the specified patch class on the measured samples; they do not rule out every conceivable function of the visible state.

An ideal observer establishes an information floor, with an important advantage over the model. It fits a degree-3 patch, a degree-5 patch with the force library, and the world's true structural class with free parameters and hidden initial conditions to the same rounded rows by multiple shooting. It then compares the candidates by BIC. n_{Bayes} is the smallest number of trajectories at which the true class wins, up to 64. The observer is given the true class among its candidates, so its success measures whether the rendered data distinguish that class, not how hard it is to invent it. Generating the hypothesis is the leap the model is asked to make.

Each trajectory shows fifty rows at a 0.2 s stride from $t = 0.0, \ldots, 9.8$, at four significant figures; fixed decimal places would give some families much greater relative precision than others. The rendering was frozen before model queries. Rows at $t \geq 10.0$, the scoring window, are never shown. The observer and model see the same rounded measurements, but their candidate hypothesis spaces differ as just described.

Furthermore, the answer format shapes the central negative result. The verbatim prompt (Appendix C.2) asks for a numerical continuation and an executable acceleration law $a(x, v, t)$ using only x, v, t and constants. It also asks for an `EXPLANATION` and an `ENTITIES` list, inviting the model to name anything not directly observed. No true restructuring law can be written in the requested executable form, so compliance confines that channel by instruction. The prose channel can still leave the supplied ontology; we read the channels separately in Section 4, and Appendix D.2 lifts the law restriction.

Lastly, each trial shows a model n rendered trajectories of one world (dose 1, 2 or 4) and requests two outputs in the same answer: a numerical prediction of Experiment 1 at the eleven unseen times $t = 10.0, \ldots, 12.0$, and a stated law. The prediction's normalized RMSE is the mean over x and v of each channel's RMSE divided by its range over the held-out window. It beats persistence when it does better than repeating the last shown row. The stated law is compiled, integrated from Experiment 1's rendered initial condition, and scored on the unseen rows $t = 10.0, \ldots, 19.9$; it verifies when its joint R^2 on x and v reaches 0.70. An arm is one model under one condition and API route. Each cell has three seeds, and every reported interval is a 95 % percentile bootstrap over worlds with 10,000 resamples, not over rows. Automated form checks are supplemented by language-model-based evaluators where an answer requires interpretive labeling; the reader protocol and its agreement limits appear in Appendix D.2 and Appendix J.1.

4 Behavioral results

Gate G0 asks whether a model beats persistence on L0, where the available law is Newtonian. Of twelve arms across six model families, only one passes the full 108-trial check: the frontier model DeepSeek-V4-Pro (`deepseek-v4-pro-0813`, reasoning disabled, temperature 0.7, top-p 0.95 and a 1,500-token answer budget). Five arms fail and six remain incomplete, so the incomplete arms carry no verdict; none was stopped because of model quality (Appendix D.1). We ran the full novelty ladder on that frontier arm and on `Qwen3-1.7B`, but the latter does not pass G0. The frontier arm is therefore the study's powered test of restructuring, while the local-model results describe what happens at a scale that has not demonstrated the basic capability.

The frontier arm contains 432 trials: forty-eight worlds at doses of one, two or four trajectories, with three seeds in every cell. The pre-specified existence gate asks whether there is even one L2 trial whose stated law both verifies, whose continuation beats persistence, and whose law reads more state than x, v and t. Its last condition cannot be met by a law that follows the requested format (Section 3). Functionally qualifying trials number 35 of 108 on L0, 11 of 108 on L1, and 1 of 216 on L2 (Table 6, Appendix D.1). That one L2 trial is L2b-05 at dose 1, seed 2. It states a Newtonian patch that fits one hysteretic trajectory closely enough ($R^2 = 0.775$), not a law that recovers the missing state. Thus the gate has zero restructuring candidates, even before asking what ontology a candidate uses. Across the 432 trials, the pre-specified form check labels 331 `PATCH`, 86 `NEEDS_JUDGE`, 15 `OTHER`, and zero `RESTRUCTURE`.

Because a classifier that has never produced a positive label has unknown sensitivity, we checked the result without relying on that classifier. We parsed every executable law and asked what state it reads (Table 1). Of 1,872 answers, 893 contain a law that parses; all 893 use only x, v and t. The other 979 answers do not conceal an alternative result: 886 state a law in a form the parser misses, most often

Table 1: What the stated laws read, on the two arms run through the whole ladder. The last two columns are read by hand (Appendix D.3): a hidden driver named in the answer, not a mass or spring called "not directly observed" as the prompt invites, then not expressible by the law beside it.

model	level	answers	laws that parse	signature (x, v, t)	leaves $F(x, v, t)$	answer names a hidden driver	names it, then cannot use it
`Qwen3-1.7B`	L0	72	25	25	0	0	0
`Qwen3-1.7B`	L1	456	173	173	0	0	0
`Qwen3-1.7B`	L2	912	311	311	0	0	0
`deepseek-v4-pro`	L0	108	94	94	0	1	1
`deepseek-v4-pro`	L1	108	91	91	0	1	1
`deepseek-v4-pro`	L2	216	199	199	0	5	3
all		1872	893	893	0	7	5

a function named `acceleration`, and none of those reads more (Appendix D.1); 93 state no law. Two language-model judges also found no restructuring answer, although their agreement failed its own pre-specified bar (Section 10). Most importantly, the prompt explicitly instructed the model to write a function of x, v and t. The 893-of-893 count measures compliance with that instruction. To ask whether models can leave the supplied state description, we also have to read the unconstrained prose and lift the restriction itself.

Those two other readings change what the zero can mean. In the original-format answers, the prose names a hidden driver in only 7 of 1,872 cases; in five of those, the executable law beside the prose cannot use it. The prose was free to mention hidden structure, so its rarity is evidence that these models seldom do so here, although seven mentions are too few to establish detection of the underlying mechanism. When the format restriction is removed without a cue, the frontier model still produces no law with additional state on the 24 L2 worlds, and three of six further models likewise produce none (Appendix D.2). The other three produce one uncued law each, describing a hidden memory, an undefined force `F(t)`, or an initial velocity; the last two rely on split votes of the three readers. With an explicit cue, 60 of 504 laws introduce auxiliary state, but none states its world's governing law. Four further labs in Appendix D.1 were tested only under the original format, so their zeros do not answer this ablation question.

5 Stage 1: Formation and the novelty horizon

The planned Formation test was a decode gap between a represented concept and a failed answer. Because no restructuring event was verified (Section 4), that gap cannot be formed. At 1.7B, the probe passes three of four controls but fails the input-only baseline: forty-eight statistics computed from the rendered rows decode `subtype` just as well as the residual stream (0.944 for both at layer 27), and `level` better (0.792 against 0.778). The probe therefore does not isolate a concept the model constructed (Appendix E.1). We report the pre-specified alternative, a novelty horizon, instead.

The ladder also has a continuous dial. Each L2 subtype has one dimensionless knob whose zero limit returns to a world the tested patch class handles exactly: k_c/k_1 decouples the hidden particle, T/period collapses the internal mode onto $\mathrm{sign}(v)$, and $\tau\sqrt{k}$ removes the delay. This dial fills the region the three-level ladder leaves empty: patchability in the original worlds is bimodal, with no value between R^2 0.650 and 0.992. We place three instruments on the same swept worlds (Appendix E.2). The ideal observer is especially informative here. At all 24 points inside the certified production box, where the models actually fail, it prefers the true structural class after one rendered trajectory, and 64 of 75 points across the full dial agree. That comparison shows evidence in the rendered data for a supplied class; it does not make inventing the class easy.

Two frontier models from different labs were then tested at 150 designed points each, with four trajectories per point. We scored every stated law on acceleration, the same quantity used for certification (Appendix E.1). Stated-law quality tracks patchability: Spearman ρ is +0.428 [+0.288, +0.546] for DeepSeek-V4-Pro and +0.412 [+0.228, +0.591] for GPT-5.6-Luna, with intervals clustered over the 18 subtype-by-seed cells. Yet where both a model crossing and a world crossing can

be located, three cells for each model, the model crosses first in all six. In the barely novel band, GPT-5.6-Luna chooses the right functional family in 43 of 46 laws but cannot instantiate it: the two models reach median matched R^2 values of 0.328 and 0.371 where the best member of the family reaches 0.974. They can therefore fail at parameter identification before the ontology itself has to change (Figure 7).

6 STAGE 2: TRUST

Trust asks whether a model favors the right account over a familiar patch when both are put before it. We measured this on `Qwen3-1.7B` in two ways, while keeping in view that Formation did not establish construction of the right concept at this scale. The pre-specified forced choice turns out to be dominated by the order of the options: position bias is roughly nine times the largest account contrast, and no preference survives counterbalancing (Appendix F.2).

We therefore scored each account as a continuation of the same prefix, with neither a label nor an option position, and read the model's own log-odds between them. On L0 and L1, the two accounts describe the same law in different words, giving a phrasing null. On L2, the log-odds for the true account fall below that null by -11.719 [-30.220, -3.645] summed over tokens and by -1.945 [-2.106, -1.830] per token. Where the tested Newtonian patch fails, this 1.7B model assigns the true account less probability than phrasing alone predicts (Appendix F.1). The result does not support an account in which this model knows the missing law but merely declines to state it.

7 STAGE 3: USE AND THE VERBALIZATION GAP

The proposed simulation-to-axioms bottleneck predicts that a model could continue a world while failing to state the law that governs it (Zahavy, 2026). We test those two outputs within the same answer. The pre-specified strict readout counts a continuation only if it beats persistence and the law beside it compiles but does not verify. A missing law receives a held-out R^2 of zero in the scoring pipeline, but it cannot demonstrate failure to translate a represented account into words, so we exclude it from this readout (Table 2).

The two middle columns of Table 2 move differently as novelty rises. The rate of stated laws that verify falls from 0.417 to 0.121 to 0.005, while the rate of continuations that beat persistence is 0.806, 0.960 and 0.841. At L2, 78.5 % of answered trials beat persistence while stating a compiled law that does not verify, across 24 worlds; just 0.5 % state a verifying law. This is the direction the translation hypothesis predicts, but persistence is too weak a comparator to establish an internal simulation (Section 7.2). Against constant velocity, a stronger baseline, continuation is nearly flat across the ladder (0.660, 0.667 and 0.664) while stated-law correctness still falls by almost two orders of magnitude (Appendix G.1). Section 7.4 tests what these comparisons license once an ontology-free predictor is fitted to the same rows.

7.1 THE GAP WITHIN ONE ANSWER

Everything above compares two rates over the answered trials: how often a continuation beats a baseline, and how often a stated law verifies. The claim those support is about one answer at a time, and it can be measured that way because every trial stores both halves. In each answered trial the model produced a table and a law. We integrate the stated law from the same initial condition, sample it at the same points the model answered, and score both against the same world under the same held-out normalisers, so the two numbers are values of one quantity and their difference is admissible (Table 3; positive means the table wins).

Across 1,336 answers, 12 arms, six model families and 22 (arm $\times$ level) cells, the model's own numbers beat the law beside them in 1,223 cases. The world-clustered proportion is 0.915 [0.883, 0.943], and the table wins in the median answer of every cell. The law wins the other 113, in 16 of the 22 cells (30 of the frontier model's 199 at L2). The gap is not only a property of the restructuring rung, where no law in the requested form can be right: on L0, where the correct law can be written in exactly that form, the table still wins in 398 of 445 answers. The gap is roughly an order of magnitude wider on the 1.7B model, the same capability ordering Appendix G.2 reports for locality.

Table 2: The frontier model's two capacities across the ladder (432 trials; rates over answered trials). "Strict" continues better than persistence while stating a law that compiles and does not verify.

level	answered	worlds	states a verifying law	continues better than persistence	strict: beats persistence, states a nonverifying law
L0	103/108	12	0.417 [0.243, 0.583]	0.806 [0.680, 0.915]	0.447 [0.267, 0.635]
L1	99/108	12	0.121 [0.040, 0.224]	0.960 [0.912, 1.000]	0.788 [0.695, 0.873]
L2	214/216	24	0.005 [0.000, 0.014]	0.841 [0.773, 0.902]	0.785 [0.720, 0.847]

Table 3: The gap inside one answer: the stated law's nRMSE minus that of the same answer's table, against the same world on the same points (median over answers, world-clustered interval; positive means the table wins), and the answers in which the table wins.

level	frontier model: law − table	`Qwen3-1.7B`: law − table	table beats its own law, all twelve arms
L0	+0.412 [+0.257, +0.674] (n = 99)	+2.624 [+2.004, +4.208] (n = 84)	398 of 445
L1	+0.681 [+0.436, +1.609] (n = 91)	+10.464 [+6.963, +16.485] (n = 88)	291 of 301
L2	+0.496 [+0.273, +0.779] (n = 199)	+6.417 [+3.142, +10.195] (n = 176)	534 of 590

That is the verbalization gap stated as a property of individual answers rather than of two populations, and it is the strongest form in which this paper can make it.

7.2 Autoregressive control of the prediction gap

The within-answer gap could suggest a simulation the model cannot state. We test that interpretation against predictors that do not encode the governing law (Appendix G.5). The context-fitted predictors see the same fifty rendered rows and are scored on the same continuation points. Every Newtonian competitor, though given the world's exact accelerations at the shown states, loses to the frontier model's continuation on most *powered* L2 trials, those where the best Newtonian competitor does not already reproduce the world (153 of 214). A four-lag linear autoregression, which sees only the rounded rows of Experiment 1, beats the model on all but 29 of the 153, by a factor of roughly eighteen in median error (0.008 against 0.143), and does not degrade when continued ten times further than any model was asked to predict (Appendix G.6). A coupled oscillator is linear in its full state and a delay equation is a linear recurrence on lags, so a lagged linear predictor recovers what the certificate rules out of the tested $F(x, v)$ class without representing its governing mechanism: the certificate excludes a hypothesis class, not easy prediction (Figure 1).

The pooled count mixes three families whose best ontology-free floors differ by a factor of fifty (Table 4). Where such a predictor already reaches 0.003, a family cannot discriminate in either direction, and `hidden_particle`'s -0.275 is not evidence that the model is bad there. Swept over floor bars from 0.005 to 0.100, `hysteretic` is licensed at every bar and `hidden_particle` at none. On `hysteretic` the model's paired advantage over the autoregression is +0.058 [-0.027, +0.108]; we find no association with the hidden mode. On 36 worlds rebuilt under a new master seed, against outcomes fixed before those worlds existed, its sign replicates: +0.020 [-0.028, +0.086] over 75 powered trials on the 12 fresh hysteretic worlds. Both intervals include zero, so what replicated is the direction, not a magnitude separated from nothing. The partition did not replicate. `delayed`, whose incumbent floor missed the 0.010 bar by three parts in a hundred thousand, crosses it on the fresh worlds, and where it is admitted the model loses by -0.120 [-0.165, -0.083]. The same arm thus has opposite estimated signs in two admitted families; pooling would hide that distinction. The partition into regimes is itself a threshold a benchmark chooses, and it has to be swept rather than quoted.

Table 4: The pooled comparison split by family (frontier model, powered L2 trials). The floor is the best of three ontology-free continuations; the last column counts trials.

family	powered (worlds)	best ontology-free floor	model ↔world	AR − model	model beats AR
`delayed`	60 (8)	0.010 [0.006, 0.013]	0.109 [0.103, 0.117]	−0.092 [−0.099, −0.078]	5 of 60
`hidden_particle`	51 (8)	0.003 [0.002, 0.003]	0.276 [0.231, 0.337]	−0.275 [−0.335, −0.228]	0 of 51
`hysteretic`	42 (8)	0.137 [0.099, 0.212]	0.104 [0.092, 0.118]	+0.058 [−0.027, +0.108]	24 of 42

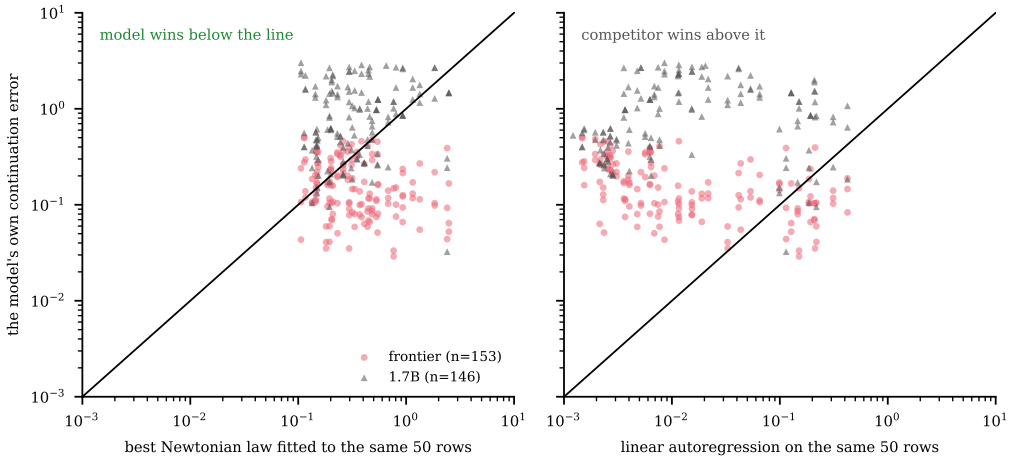

Figure 1: Every powered L2 trial plotted twice against the same diagonal, with the model's own continuation error on the vertical axis: against the best Newtonian law fitted to the fifty rows it read, where the frontier points sit below the line, and against a linear autoregression on those same rows, where they sit above it. The reversal is the result.

8 Intervention

The planned intervention was to estimate an activation direction from trials on which the model restructured and steer later trials toward it. Gate G3 cannot supply that direction, because there are no positive restructuring trials (Section 4). We tested a different contrast the data do permit: the residual-stream difference between a prompt that names the hidden channel and the ordinary prompt. This is a direction tied to the input, not to observed success. On `Qwen3-1.7B`, the restructuring rate remains 0.000 and the change in matched R^2 is +0.001 [−0.001, +0.003] (Table 21). The intervention alters 34 of 36 completions, but a decoy and ten random directions behave similarly, while the worst capability substitute drops 8.3 %, beyond the specified 5 % budget (Appendix H.3). It does not induce restructuring at 1.7B, and it cannot tell us whether the unavailable success direction would have done so.

We also tested two remedies that require no target concept. Thirty-two samples per prompt at 1.7B add 0 verifying laws in 768 attempts (Appendix H.2). The frontier model was given interventional access: it could choose initial states and receive their trajectories before answering (Appendix H.1). That access helps at retrieval, changing nRMSE relative to persistence by −0.037 [−0.185, −0.005] at L0, but has no detected effect where the ontology must change (−0.003 [−0.057, 0.048] at L2). Its most common request is the canonical condition $(0, 0)$, made in 193 of 559 granted requests, rather than a condition tailored to distinguish mechanisms. The request pattern does not reveal every hypothesis the model considered, but it limits what this intervention can claim to have tested.

9 DISCUSSION

The existence gate returned capability without a verified restructuring event. Under the original answer format, no parseable stated law leaves $F(x, v, t)$, but the format tells models to stay there. When that restriction is lifted, three of seven models produce one uncued law with additional state each, two by split votes; none states its world's governing law, and the frontier model produces none uncued. The unconstrained prose names a hidden driver in only 7 of 1,872 original-format answers (Appendix D.3). What the study establishes, therefore, is a scarce and fragile departure from the variables supplied in the prompt and no verified recovery of an L2 law in the tested arms.

The numerical gap does not rescue a stronger interpretation. The frontier model's continuations lose to a four-lag linear autoregression on 124 of 153 powered trials (Section 7.2). These worlds make short-range prediction easier than law recovery, so the within-answer gap remains a measured fact without by itself demonstrating hidden simulation. A claim about an internal world model would need to show that ontology-free predictors fail in the same regime where the model succeeds; the three families and the fresh-world sweep show why this check must be made separately rather than pooled (Section 7.2).

10 LIMITATIONS

The prediction task is easier than the law task, as the control reveals. The empirical certificate excludes the tested local $a = F(x, v)$ patch class on the accepted L2 worlds; it does not show that their trajectories are hard to extrapolate. A four-lag linear autoregression fitted to the fifty rendered rows tracks these worlds to within 2 % of the held-out signal range at every horizon the design contains, out to ten times the window the models were asked to predict (Appendix G.6). A coupled oscillator is linear in its full state, and a delay equation can be approximated by a recurrence on lags. Consequently, a high continuation rate against persistence does not imply an internal simulation, and part of the dissociation elicited by this design comes from the task itself. A corrected test should require extrapolation or transfer that simple predictors cannot already solve. We specify transfer to an unseen preparation of the same world, under an answer format that does not prescribe the law's state, before any of its results exist (Appendix J.2); we have not run it.

The original answer format adds a second limitation by asking for a law using only x, v, t and constants (Section 3), a form no true restructuring law in this study can satisfy. The 893-of-893 signature count is therefore a compliance rate. The ablation removes that constraint at one dose and on one world set (Appendix D.2). On the frontier model, uncued, neither the law (0 of 72) nor the prose (1 and 3 of 72 by the two readers) moves beyond the supplied ontology. On `Qwen3-1.7B`, freed laws become constants. Of five further models, three produce one uncued law with additional state each, two by split votes. This narrows the format explanation without making the original zero an unconstrained measure of what every model can propose.

Finally, the labels are not validated against human annotations, and one model carries the powered depth result. Two language-model judges fail their pre-specified agreement bar (κ 0.603 against 0.70), so the restructuring conclusion rests on the form check and independent mechanical audits (Appendix J.1); the ablation's labels come from three language-model readers. Every full-depth capability result comes from one frontier model through one API route, and a second route to the same model moves paired nRMSE by more than one route's seed-to-seed spread. We report those limits with the result rather than treating a null as a universal claim.

AI USE STATEMENT

Generative AI tools were used to assist with revising and improving software code and to aid in polishing the writing of this paper. Separately, language-model-based evaluators were used to label experimental responses as part of the study methodology, as described in the Methods section. All AI-assisted code and written content were reviewed and verified by the authors, who take responsibility for the final content of the work.

Ethics statement

This study has no human participants and uses no personal data. Its worlds are synthetic, generated from procedurally sampled physical laws, and its data are those worlds' trajectories and the models' answers to them, obtained through the providers' public interfaces and from open-weight models under their licences. The answers were labelled by the automated checks and the language-model readers described in the methods; no one outside the authors was asked to annotate, rate or take part in anything. The released code, data and decision ledger are anonymized and identify no one. The work measures how models explain synthetic dynamics and releases the instrument it uses to do so, and we are not aware of a harmful use of either.

Reproducibility statement

The reproducibility appendix documents the released world generator and its certification chain, every manifest, the frozen rendering configuration, all prompts, the Bayesian ideal observer, the symbolic-regression harness, the analysis scripts and the figure renderers, so that every world, certificate, rendered prompt, baseline fit and figure regenerates from the manifests and the recorded seeds; trajectories themselves are not shipped, because they are reproduced exactly from those. That exactness is measured rather than asserted: worlds, certifications and symbolic fits re-run from stored seeds through the production code path match the manifests with no tolerance, and a source fingerprint stamped into every receipt labels any artifact read on a drifted tree as stale rather than letting it overwrite a verified one. The frontier arms are not bit-reproducible, and that is measured too: two sweeps of the same model on the same route with the same seeds returned a byte-identical completion in none of the matched cells, so every frontier number is a single-sample estimate of a stochastic system whose intervals cluster over worlds. The code, the manifests, the raw completions, the manuscript's source and an anonymized copy of the decision ledger are provided as supplementary material, and every number printed in the paper resolves to a named artifact in them by an automated check that runs from the archive itself.

References

L. Berglund, A. Cooper Stickland, M. Balesni, M. Kaufmann, M. Tong, T. Korbak, D. Kokotajlo, and O. Evans. Taken out of context: On measuring situational awareness in LLMs. 2023. doi: 10.48550/arXiv.2309.00667.

C. Bhagavatula, R. Le Bras, C. Malaviya, K. Sakaguchi, A. Holtzman, H. Rashkin, D. Downey, S. W. Yih, and Y. Choi. Abductive Commonsense Reasoning. *ICLR*, 2020. doi: 10.48550/arXiv.1908.05739.

Y. Chen, P. Piękos, M. Ostaszewski, F. Laakom, and J. Schmidhuber. PhysGym: Benchmarking LLMs in Interactive Physics Discovery with Controlled Priors. *NeurIPS*, 2025. doi: 10.48550/arXiv.2507.15550.

F. Chollet. On the Measure of Intelligence. 2019. doi: 10.48550/arXiv.1911.01547.

F. Chollet, M. Knoop, G. Kamradt, B. Landers, and H. Pinkard. ARC-AGI-2: A New Challenge for Frontier AI Reasoning Systems. 2025. doi: 10.48550/arXiv.2505.11831.

M. Cranmer. Interpretable Machine Learning for Science with PySR and SymbolicRegression.jl. 2023. doi: 10.48550/arXiv.2305.01582.

J. Dodge, M. Sap, A. Marasović, W. Agnew, G. Ilharco, D. Groeneveld, M. Mitchell, and M. Gardner. Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus. *EMNLP*, 2021. doi: 10.18653/v1/2021.emnlp-main.98.

Y.-L. Fang, D.-S. Jian, X. Li, and Y.-Q. Ma. AI-Newton: A Concept-Driven Physical Law Discovery System without Prior Physical Knowledge. 2025. doi: 10.48550/arXiv.2504.01538.

J. Feng, S. Russell, and J. Steinhardt. Extractive Structures Learned in Pretraining Enable Generalization on Finetuned Facts. *ICML*, 2025. doi: 10.48550/arXiv.2412.04614.

S. Golchin and M. Surdeanu. Data Contamination Quiz: A Tool to Detect and Estimate Contamination in Large Language Models. *TACL*, 2025. doi: 10.1162/tacl.a.20.

N. Gruver, M. Finzi, S. Qiu, and A. G. Wilson. Large Language Models Are Zero-Shot Time Series Forecasters. *NeurIPS*, 2023. doi: 10.48550/arXiv.2310.07820.

R. Guimerà, I. Reichardt, A. Aguilar-Mogas, F. A. Massucci, M. Miranda, J. Pallarès, and M. Sales-Pardo. A Bayesian machine scientist to aid in the solution of challenging scientific problems. *Science Advances*, 2020. doi: 10.1126/sciadv.aav6971.

W. Gurnee and M. Tegmark. Language Models Represent Space and Time. *ICLR*, 2024. doi: 10.48550/arXiv.2310.02207.

R. Hendel, M. Geva, and A. Globerson. In-Context Learning Creates Task Vectors. *Findings of EMNLP*, 2023. doi: 10.48550/arXiv.2310.15916.

Y. Huang, H. Zhu, T. Guo, J. Jiao, S. Sojoudi, M. I. Jordan, S. Russell, and S. Mei. Generalization or Hallucination? Understanding Out-of-Context Reasoning in Transformers. *NeurIPS*, 2025. doi: 10.48550/arXiv.2506.10887.

S. Imran, R. Lamb, and P. M. Atkinson. Out-of-Context Abduction: LLMs Make Inferences About Procedural Data Leveraging Declarative Facts in Earlier Training Data. 2025. doi: 10.48550/arXiv.2508.00741.

P. Jansen, M.-A. Côté, T. Khot, E. Bransom, B. Dalvi Mishra, B. P. Majumder, O. Tafjord, and P. Clark. DiscoveryWorld: A Virtual Environment for Developing and Evaluating Automated Scientific Discovery Agents. *NeurIPS Datasets and Benchmarks Track*, 2024. doi: 10.48550/arXiv.2406.06769.

Z. Jin, P. Cao, H. Yuan, Y. Chen, J. Xu, H. Li, X. Jiang, K. Liu, and J. Zhao. Cutting Off the Head Ends the Conflict: A Mechanism for Interpreting and Mitigating Knowledge Conflicts in Language Models. *Findings of ACL*, 2024. doi: 10.18653/v1/2024.findings-acl.70.

N. Koblischke, H. Jang, K. Menou, and M. Ali-Dib. Gravity-Bench-v1: A Benchmark on Gravitational Physics Discovery for Agents. *ICML*, 2025. doi: 10.48550/arXiv.2501.18411.

K. Li, A. K. Hopkins, D. Bau, F. Viégas, H. Pfister, and M. Wattenberg. Emergent World Representations: Exploring a Sequence Model Trained on a Synthetic Task. *ICLR*, 2023. doi: 10.48550/arXiv.2210.13382.

J. Liu, S. Min, L. Zettlemoyer, Y. Choi, and H. Hajishirzi. Infini-gram: Scaling Unbounded n-gram Language Models to a Trillion Tokens. *COLM*, 2024a. doi: 10.48550/arXiv.2401.17377.

T. J. B. Liu, N. Boullé, R. Sarfati, and C. J. Earls. LLMs learn governing principles of dynamical systems, revealing an in-context neural scaling law. *EMNLP*, 2024b. doi: 10.18653/v1/2024.emnlp-main.842.

C. Lu, C. Lu, R. T. Lange, J. Foerster, J. Clune, and D. Ha. The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery. 2024. doi: 10.48550/arXiv.2408.06292.

B. P. Majumder, H. Surana, D. Agarwal, B. Dalvi Mishra, A. Meena, A. Prakhar, T. Vora, T. Khot, A. Sabharwal, and P. Clark. DiscoveryBench: Towards Data-Driven Discovery with Large Language Models. *ICLR*, 2025. doi: 10.48550/arXiv.2407.01725.

A. Novikov, N. Vũ, M. Eisenberger, E. Dupont, P.-S. Huang, A. Z. Wagner, S. Shirobokov, B. Kozlovskii, F. J. R. Ruiz, A. Mehrabian, M. P. Kumar, A. See, S. Chaudhuri, G. Holland, A. Davies, S. Nowozin, P. Kohli, and M. Balog. AlphaEvolve: A coding agent for scientific and algorithmic discovery. 2025. doi: 10.48550/arXiv.2506.13131.

L. Qiu, L. Jiang, X. Lu, M. Sclar, V. Pyatkin, C. Bhagavatula, B. Wang, Y. Kim, Y. Choi, N. Dziri, and X. Ren. Phenomenal Yet Puzzling: Testing Inductive Reasoning Capabilities of Language Models with Hypothesis Refinement. *ICLR*, 2024. doi: 10.48550/arXiv.2310.08559.

P. Shojaee, K. Meidani, S. Gupta, A. B. Farimani, and C. K. Reddy. LLM-SR: Scientific Equation Discovery via Programming with Large Language Models. *ICLR*, 2025. doi: 10.48550/arXiv.2404.18400.

E. Todd, M. L. Li, A. Sen Sharma, A. Mueller, B. C. Wallace, and D. Bau. Function Vectors in Large Language Models. *ICLR*, 2024. doi: 10.48550/arXiv.2310.15213.

J. Treutlein, D. Choi, J. Betley, S. Marks, C. Anil, R. Grosse, and O. Evans. Connecting the Dots: LLMs can Infer and Verbalize Latent Structure from Disparate Training Data. *NeurIPS*, 2024. doi: 10.48550/arXiv.2406.14546.

K. Vafa, J. Y. Chen, A. Rambachan, J. Kleinberg, and S. Mullainathan. Evaluating the World Model Implicit in a Generative Model. *NeurIPS*, 2024. doi: 10.48550/arXiv.2406.03689.

K. Vafa, P. G. Chang, A. Rambachan, and S. Mullainathan. What Has a Foundation Model Found? Using Inductive Bias to Probe for World Models. *ICML*, 2025. doi: 10.48550/arXiv.2507.06952.

J. von Oswald, E. Niklasson, E. Randazzo, J. Sacramento, A. Mordvintsev, A. Zhmoginov, and M. Vladymyrov. Transformers learn in-context by gradient descent. *ICML*, 2023. doi: 10.48550/arXiv.2212.07677.

A. Wang, J. Engels, O. Clive-Griffin, S. Rajamanoharan, and N. Nanda. Simple Mechanistic Explanations for Out-Of-Context Reasoning. *ICML Workshop R2-FM*, 2025. doi: 10.48550/arXiv.2507.08218.

R. Wang, E. Zelikman, G. Poesia, Y. Pu, N. Haber, and N. D. Goodman. Hypothesis Search: Inductive Reasoning with Language Models. *ICLR*, 2024. doi: 10.48550/arXiv.2309.05660.

T. Zahavy. Position: LLMs Can't Jump. ICML 2026 position track, 2026.

A. Zeng, M. Chen, L. Zhang, and Q. Xu. Are Transformers Effective for Time Series Forecasting? *AAAI*, 2023. doi: 10.1609/aaai.v37i9.26317.

T. Zheng, K. K.-W. Tam, N. H.-N. K. Nguyen, B. Xu, Z. Wang, J. Cheng, H. T. Tsang, W. Wang, J. Bai, T. Fang, Y. Song, G. Y. Wong, and S. See. NewtonBench: Benchmarking Generalizable Scientific Law Discovery in LLM Agents. *ICLR*, 2026. doi: 10.48550/arXiv.2510.07172.

A Introduction

A.1 Contributions and evidence boundaries

We generate worlds with known governing parameters and measure exact-string contamination rather than assume its absence. The novelty ladder includes a restructuring rung on which the best tested patch remains below $R^2 = 0.70$, even when fitted and scored in-sample. A continuous dial fills the patchability gap between rungs. Given the true structural class among three candidates, an ideal observer selects it after one rendered trajectory. This establishes evidence for discriminating those candidates; it does not hand a model a way to invent the class. The generator, manifests and measurement receipts are released.

The model's numerical table beats the law beside it in 1,223 of 1,336 scored answers across twelve arms, and in the median answer of every (arm $\times$ level) cell. On the restructuring rung, the frontier model beats persistence in 84 % of answered trials yet states a verifying law in 0.5 %. These are measurements of prediction and expression; the baseline comparison below determines whether they imply more.

The autoregressive control also sets a condition for interpreting that gap: predictive accuracy can support an inference about internal structure only where ontology-free predictors fail on the same regime. Our three restructuring families have ontology-free floors of 0.003, 0.010 and 0.137. A median over them turns a fact about which families were in the pool into an apparent fact about the model, and a family whose floor is already near-exact can supply evidence neither that the model is bad nor that it is good. Applied per family, the criterion leaves one regime in which a model-versus-baseline comparison carries weight at every threshold we sweep. The sign of the model's advantage

there replicates on 36 fresh worlds under a new master seed, against outcomes fixed before any of them existed: +0.020 [−0.028, +0.086] against +0.058 [−0.027, +0.108] on the original eight, both intervals including zero. The partition does not replicate. A second family crosses the licensing threshold on the fresh worlds, where the original set missed it by three parts in a hundred thousand, and where it is admitted the model loses by −0.120 [−0.165, −0.083]. The demonstration is owed per regime, with a sweep of the threshold that defines it. The check is inexpensive: fit a linear autoregression to whatever the model was shown, within each regime separately, and see whether it wins.

A pre-specified ablation of the answer format (Appendix D.2) shows that lifting the $a(x, v, t)$ constraint without a cue leaves the zero in place on four of seven models, the frontier model among them, that `Qwen3-1.7B`, freed of the format, states constant accelerations, and that three models, uncued, each write one law that reads more, two of them by split votes of three readers. On the novelty dial, stated-law quality declines with patchability on two frontier models from different labs (ρ = +0.428 [+0.288, +0.546] and +0.412 [+0.228, +0.591]), and both models break while the world is still patchable (Section 5). The laws that are right are local: the fraction of a world's other initial conditions one stated law still covers falls across the ladder while its fit to the trajectory in the prompt does not, and the failure relocates up the ladder with capability rather than vanishing (Appendix G.2). Interventional access helps where the answer is retrieval and not where the ontology must change (Appendix H.1). And the stage decomposition is reported with each stage's power stated. In the Trust readout without option positions, log-odds favor the Newtonian account where the tested patch class fails (Appendix F.1).

B Related work

B.1 Related mechanisms, out-of-context reasoning and contamination

Mechanistic studies locate competition between parametric memory and in-context evidence, including attention heads whose removal changes that balance (Jin et al., 2024). Those experiments ask a model to select between factual claims. Here a model must consider an explanatory class outside the one suggested by the observed variables. Our forced-choice readout has too little power to localize that process; Section 6 reports the failure as a limit of the instrument.

Function and task vectors show that internal summaries of an in-context task can be extracted and transplanted (Todd et al., 2024; Hendel et al., 2023); another line analyzes in-context learning as implicit gradient descent (von Oswald et al., 2023). We ask whether a representation tracks the content of a generated world's missing state, with ground truth known by construction.

Because a verification loop can assess only the hypotheses it receives, the original answer format's exclusion of the governing L2 mechanisms from its executable channel makes hypothesis generation and format part of the evaluation (Section 3).

Models fine-tuned on documents containing fragments of a latent fact can infer and state that fact at test time (Treutlein et al., 2024; Berglund et al., 2023), including abductive-looking inferences (Imran et al., 2025); recent work studies the mechanism (Wang et al., 2025; Huang et al., 2025; Feng et al., 2025). Those facts enter the model during training. Our world instances are generated after training and presented in context, though the general concepts needed to describe them may already be familiar.

Synthetic-task models can develop probe-recoverable board states (Li et al., 2023), language models can encode space and time linearly (Gurnee & Tegmark, 2024), and accurate orbital prediction can coexist with heuristics rather than Newtonian mechanics (Vafa et al., 2024). We measure, within each answer, how a numerical continuation compares with the law the model states when the true law is known. The autoregressive control then tests how much that gap can say about representation (Section 7).

Corpus documentation and contamination-detection work motivate our methodology (Dodge et al., 2021; Golchin & Surdeanu, 2025); we measure contamination rather than assert it, using exact n-gram counts over four open corpora (Liu et al., 2024a) on the frozen rendering itself (Section 3).

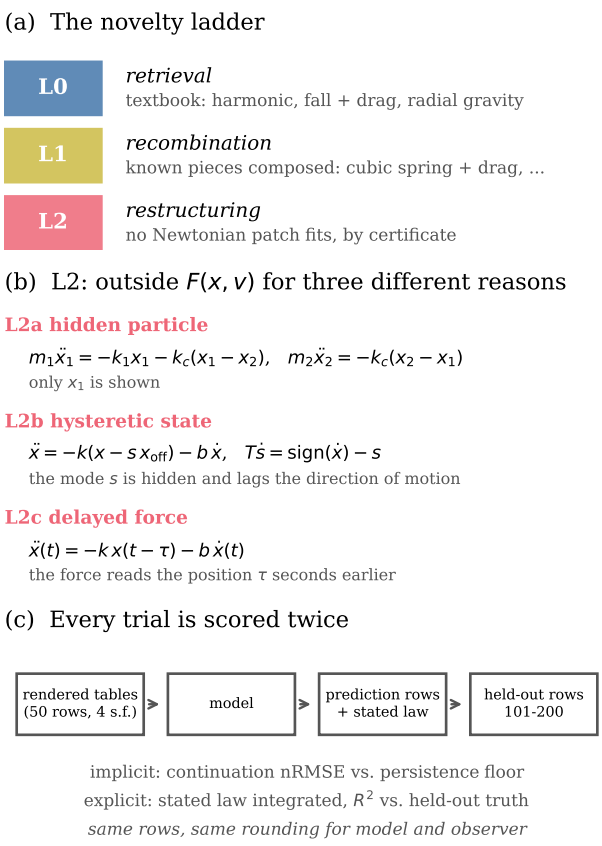

Figure 2: The novelty ladder, the three reasons the restructuring worlds leave the observed instantaneous state, and the two scores obtained from each answer. The model and the ideal observer receive the same rounded rows; the observer is additionally supplied the true structural class among its candidates.

Table 5: Exact n-gram counts over four open corpora on strings drawn from the frozen rendering, from three consecutive rendered rows down to one bare value, the resolution control.

rung	median tokens	hits / measured	worlds with a hit
3 consecutive rendered rows	54	0 / 12	0 / 48
2 consecutive rendered rows	36	0 / 12	0 / 48
1 full rendered row (t, x, v)	17	0 / 12	0 / 48
one (x, v) pair, time dropped	14	0 / 1114	0 / 48
one bare x value, resolution control	6	1114 / 1152	48 / 48

C THE INSTRUMENT (METHODS)

C.1 CONTAMINATION AUDIT

Using infini-gram (Liu et al., 2024a), we queried 2,423 distinct strings from the frozen rendering against the Pile, RedPajama, C4 and Dolma v1.7. These are exact-string tests on material the model sees, at successively finer granularity:

The bare values appear as many as 1,918,768 times while the longer rendered strings return zero. The resolution control shows that zero counts at longer lengths are not caused by a corpus with no

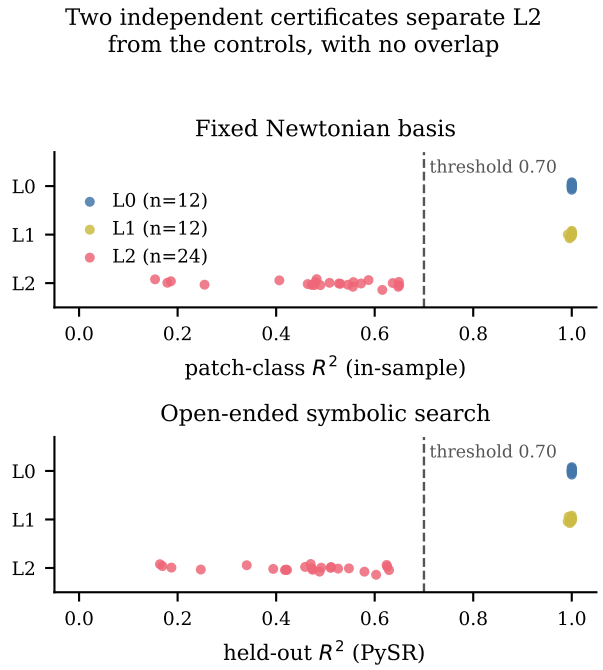

Figure 3: The accepted restructuring worlds remain below the 0.70 fit threshold in both the fixed local patch class and the held-out symbolic-regression check, while the retrieval and recombination controls lie above it. These are tests of the specified search spaces and samples, not a proof against every function of the visible state.

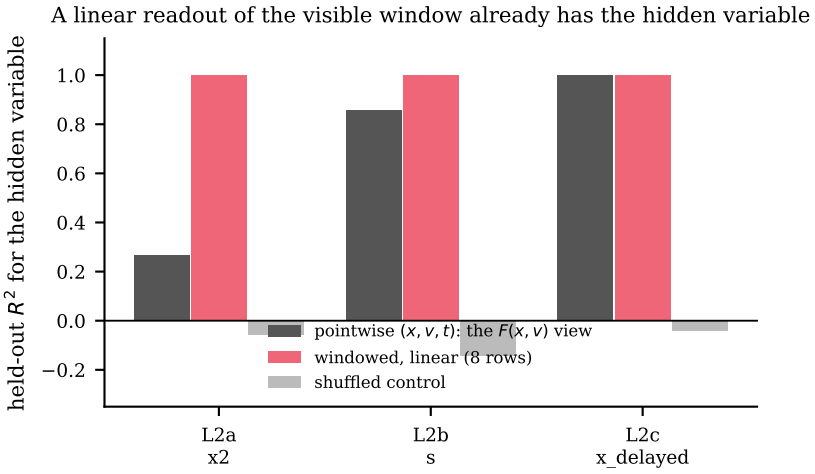

Figure 4: The design property the instrument rests on. The hidden variable is ontologically hidden, not informationally hidden: a linear readout of eight visible rows recovers it on held-out data for every restructuring family, where a pointwise (x, v, t) readout, the $F(x, v)$ view, falls short for two and a shuffled control finds nothing. With the ideal observer of Section 3, which prefers the true structural class after a single trajectory when handed it as a candidate, a model that fails on these worlds has not been starved of evidence.

matching numbers. We re-measured twelve stacks without short-circuiting and found 0 violations of the rule that a longer exact string cannot have more matches than its fragment.

In the Pile alone, "coupled oscillator" has 5,564 hits, "delay differential equation" 3,149 and "hysteresis" 351,126. The exact rendered sequences in our queries have no hits in the four checked corpora. This does not rule out semantic overlap elsewhere in training data. The task tests whether a model infers a generated instance that requires leaving the prompted $F(x, v)$ description without being told to do so, a weaker claim than inventing a new concept class (Section 10).

C.2 THE PROMPT, VERBATIM

The templates below generate every trial (`aoj/behavior/prompts.py`). Here `{n}` is the dose, `{tables}` the rendered rows and `{grid}` the prediction times. The source is bound to the study receipts; the typeset copy uses ASCII substitutions for non-ASCII characters. The passive condition sends the system prompt and the user prompt. The active condition prepends the request protocol to the same user prompt, runs the experiments the model asks for, and then sends the original data, the granted trajectories and the closing instruction as one further turn, so both conditions answer from a single self-contained prompt.

System prompt:

```
You are a physicist analyzing experimental
    data from a newly discovered
physical system. The data may or may not
    follow any known physical theory.
Base your conclusions only on the data.
```

User prompt:

```
Here are {n} experiments. Each shows a
    particle's position x and velocity v
over time.

{tables}

Tasks:
1. PREDICTION: For Experiment 1, predict x and
    v at t = {grid} s.
   Output as rows: t, x, v
2. LAW: State the law governing the particle's
    motion, in this exact format:
   PYTHON: <function a(x, v, t) returning
       acceleration, using only x, v, t and
       constants>
   EXPLANATION: <one paragraph>
   ENTITIES: <list every physical entity your
       law requires, including any not
       directly observed>
```

Active condition, prepended to the user prompt:

```
Before answering, you may request up to 8
    additional experiments. For each,
reply with: REQUEST x0=<value> v0=<value> (x0
    in [-3,3], v0 in [-3,3]).
You will receive the resulting trajectory.
    When done, reply DONE.
```

Active condition, the answer turn, appended after the original user prompt:

```
You requested {n_requested} additional
    experiment(s) and they have been run.
Their results follow, numbered independently
```

```
    of the experiments above.

{tables}

You have received every experiment you
    requested. Do not request any more. Now
    complete both
tasks above for Experiment 1 of the original
    data, in the format specified.
```

The line that matters for Section 4 is the `PYTHON` line of the user prompt. It fixes the signature and the ontology of the executable law; the `ENTITIES` line is the only place the template invites the model to name what it cannot write. The same worlds are re-asked with that one line changed in Appendix D.2.

D BEHAVIORAL RESULTS

D.1 CAPABILITY, LAW FORMS AND HIDDEN DRIVERS

DeepSeek-V4-Pro passes G0 through one metered API route (Section 10 examines another). Five arms fail: `Qwen3-1.7B`, both Gemma-2-2B arms, `Llama-3.2-3B` at 4-bit and `Llama-3.3-70B`. Six are incomplete: four stopped at free-tier quotas; `Llama-3.2-3B` at full precision and `Qwen3-4B` at 4-bit exceeded the memory of the consumer card. None of these six stopped because its score was poor, so none has a gate verdict. Their short sweeps also cannot be ranked from Figure 5. They sampled L0a, where small rendered $|x|$ places measured arms within 0.93–1.16$\times$ of persistence, while L0b and L0c give 4–11$\times$. The partial coverage changes the apparent score.

Section 5's horizon work generated L2 answers from five further arms, and the form check was applied to them without being asked to: `RESTRUCTURE` is 0 on every one. The substantial arm is `gpt-5.6-luna` (150 L2 answers, `PATCH` 142, `NEEDS_JUDGE` 8, `RESTRUCTURE` 0), and the four degeneracy probes (`claude-sonnet-5`, `gpt-5.6-terra`, `nemotron-3-super-120b`, `llama-3.3-70b`) add 14 more with the same result. Counting arms rather than answers, the label has now failed to fire on seven arms across six labs. Of luna's eight `NEEDS_JUDGE` answers, five are laws that do not compile and three are the case this paper is about: hidden-entity language present, and whether the law uses it left to a rubric. All of these answers were given under the prescribed format, so this zero too is a compliance rate. Appendix D.2 lifts the format on these models, and `gpt-5.6-luna` and `llama-3.3-70b` then each leave the ontology once.

Table 1 covers 893 parseable laws, so we classify all 979 other answers before extending its zero. They are sorted by why no law parsed. 527 state the law as a function under another name (526 of them `acceleration`, almost all Qwen3-1.7B's), 4 as a `def a` the extractor's cut lost, 338 in notation such as $a(x, v, t) = -0.2x - 0.06v$, and 17 as a lambda or a bare expression; 14 blocks are neither code nor readable notation, 14 answers stop at the token budget before any law, and 65 give no law at all. All 886 recovered laws read nothing beyond x, v, t and constants under the same rule that counts the 893, and the format ablation's recall-oriented classifier flags one, which reads an initial velocity the prose calls unobserved. Reading every frontier answer in the remainder (48), that one, and a seeded sample of 50 others finds no law that carries state; one truncated frontier answer was testing a two-oscillator model on a hysteretic world when its budget ran out, and it states no law. The reader was a language model that knew each answer's model and category.

Read rather than matched, the prose names a hidden driver in 5 of the frontier model's 216 L2 answers, too rarely to track how patchable the world is, and where it does, the same answer's law, if it states one, cannot use it; the flat rate of about one answer in eight that word lists report is mostly the prompt's own "not directly observed". Two further instruments agree, one of them approaching from the opposite direction. Appendix D.3 carries all three in full and says why each counts a different number of answers, and Appendix G.2's locality analysis rules out the most generous reading of the laws that do fit.

Table 6: Functionally qualifying trials on the frontier arm: the stated law verifies (held-out joint $R^2 \geq 0.70$) and the continuation beats persistence.

level	trials	functionally qualifying
L0 retrieval	108	35
L1 recombination	108	11
L2 restructuring	216	1

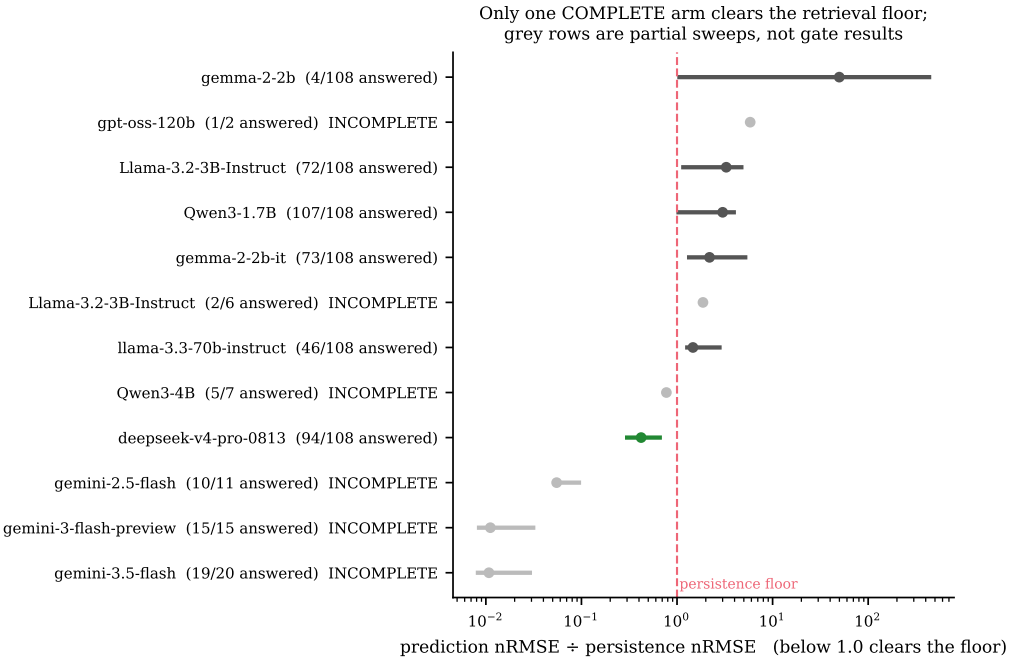

Figure 5: The behavioral floor. Each arm's prediction error on L0 relative to persistence, with its world-clustered interval and the number of trials it answered; below 1.0 clears the floor. Grey rows are incomplete sweeps and carry no verdict.

The pre-specified plan also named a six-point Pythia scaling curve for this figure. It was not produced and nothing is claimed from it: the scale axis stops at 9.2B, where a single 8 GB card and the access tier of a remote interpretability service set the limit, and every model on it sits below the G0 floor, so a curve through them would be a curve through saturated points.

D.2 Answer-format ablation

The frontier model re-answers the 24 incumbent L2 worlds at dose 4, seeds 1 to 3. The rendered rows, seeds, sampling and 1,500-token answer budget stay fixed; only one prompt line changes. `v0` keeps the frozen line (Appendix C.2); `v1` asks instead for *code returning the acceleration, in whatever form the law requires*, which lifts the constraint without mentioning hidden state; `v2` adds permission to *define any auxiliary state variables and their dynamics*, which is a cue and is counted as one. A classifier built for recall flags every law that could read state beyond x, v, t and constants, and a flag is not a count: every flag is read and its label recorded with a reason, against a rule fixed early in the reading and amended once, in the direction that counts more laws as stateful. A recall check also reads every unflagged answer whose prose a hidden-structure word list matches. The first reader knew each answer's format, so all 94 labels were then read again by two readers blind to the format and to every label, both fresh instances of a language model (Claude Opus 5.5), and the verdict reads the majority of the three. The three readings agree on every item (Fleiss' κ = 1.000);

Table 7: The answer format, ablated: the frontier model on the 24 incumbent L2 worlds at dose 4, three seeds, one prompt line changed. "Stateful" counts laws that read state beyond x, v, t and constants, by the majority of three readers; median nRMSE carries its world-clustered interval.

format	answers	block parses	signature (x, v, t)	flagged	stateful	answer names hidden	median nRMSE
`v0`, the frozen line	72	65	63	4	0	8	0.137 [0.106, 0.181]
`v1`, constraint lifted, uncued	72	67	36	29	0	8	0.154 [0.107, 0.214]
`v2`, auxiliary state permitted	72	70	29	43	5	21	0.145 [0.109, 0.199]

they differ on the kind of some items, which the verdict does not read (agreement 88 to 91 of 94), so the labels were read rather than copied.

The frozen line reproduces the stored zero in the same run. With the constraint lifted and nothing said about hidden state, none of the 72 laws reads beyond x, v and t. On this model, world set and dose, lifting the format alone does not change the zero. With explicit permission, 5 of 72 do, and none is its world's law. Two posit a hidden second body for a hidden-particle world, the right kind of mechanism: one couples it back to the observed body as the world does, but at normal modes of 0.46 and 2.19 rad/s where the world's are 0.89 and 4.07, and with damping the world lacks; the other drives the observed body without being driven by it. A third, also on a hidden-particle world, integrates a phase driven by the observed state, and a fourth offers a coupled hidden oscillator for a hysteretic world, whose hidden variable is a latched branch. The fifth is borderline: beside a law in x, v and t it offers an alternative with a linear hidden oscillator that nothing drives, whose effect is a closed form in t, and it is counted because in doubt the rule counts a law as stateful; both blind readers named it as doubtful and counted it the same way. None is on a delayed world, where one answer names a buffer of past values and does not use it.

Of `v1`'s 29 flags, 23 pass named constants as arguments, 5 write the law in notation rather than code and 1 takes a `state` argument it never reads; of `v2`'s 43, 13 pass constants, 12 carry an auxiliary variable whose dynamics reduce it to a function of t (usually a driving phase with $\mathrm{d}\phi/\mathrm{d}t = \omega$), 11 take a `state` argument the law never reads, and 2 are notation. Under the cue the model adopts the vocabulary of hidden state far more readily than its content: it dresses periodic forcing as state, the same conflation Appendix J.1 finds in the LLM judges, who read periodic forcing as an unobserved driver.

Lifting the constraint does not unlock the prose: hidden structure, read rather than matched (Appendix D.3), is named in 1 answer under `v0` and 3 under `v1`, against 13 under the cue, so the framing the $a(x, v, t)$ line might exert on the free channel is not visible at this size. Fewer laws run in the sandbox once $a(x, v, t)$ is no longer asked for (60, 55 and 44), a format effect on the functional measure and not on the ontology. Prediction does not move measurably: against the frozen line, the lifted formats' median error differs by +0.017 [−0.024, +0.071] and +0.008 [−0.024, +0.039], paired by world. No answer under a lifted format was cut off before its law, so the retry fixed in advance had nothing to re-ask, and the recall check found no stateful law among the unflagged answers it read.

Before any of their answers existed, the design was extended unchanged to two further models. `Qwen3-1.7B`, which supplies 1,440 of Table 1's 1,872 answers and fails G0, answered the same 216 prompts on the local GPU with the stored arm's sampling (Table 8). The outcome is O1: no law reads state beyond x, v and t under any format, the cued one included. Two answers are flagged, both under the frozen line, and both write in notation a law that reads no more than v and t; the first reader and both blind readers label both as inside. No answer names hidden structure in its prose. It is the weak form of that outcome, and what the laws read says why: under the frozen line 19 of 72 read x or v and 49 are constants, and once the prescribed signature is gone every law, cued or not, is a constant acceleration. Freed from the format, this model states less rather than more, so on it the zero is not the format's doing; whether a model that can fit these worlds would leave the ontology is the frontier arm's question, answered above. Prediction is not compared across its formats: the sampler is seeded per cell, so the table, which precedes the law, repeats the frozen format's in 33

Table 8: The same ablation on `Qwen3-1.7B`, run on the local GPU with the stored arm's sampling: the 24 incumbent L2 worlds at dose 4, three seeds, one prompt line changed. Columns as in Table 7; "signature (x, v, t)" counts single-function laws whose arguments are drawn from x, v and t, which includes this model's `acceleration(t)`.

format	answers	block parses	signature (x, v, t)	flagged	stateful	answer names hidden	median nRMSE
`v0`, the frozen line	72	70	70	2	0	0	0.993 [0.538, 1.341]
`v1`, constraint lifted, uncued	72	72	72	0	0	0	0.896 [0.528, 1.259]
`v2`, auxiliary state permitted	72	72	72	0	0	0	0.907 [0.516, 1.255]

and in 34 of 72 answers, and the paired differences (-0.098 [-0.237, +0.067] and -0.086 [-0.241, 0.000]) are partly fixed by construction. The second arm, `gemma-2-9b-it` through a remote interpretability service, did not run: a prompt of dose-4 length exceeds that deployment's per-job memory, so it carries no verdict.

The answers behind Appendix D.1's four further labs were given under the prescribed format, so the design was extended unchanged to the models that could be reached: `gpt-5.6-luna`, `nemotron-3-super-120b`, `llama-3.3-70b`, `llama-3.1-70b` and `llama-4-maverick`, the last standing in for a 405B model no route serves. Each was pinned to one provider with no fallback, and each of the 1,080 answers records the provider that served it; every one was served by its pin (Table 9). An answer cut off before its law was asked once more at 3,000 tokens, beside the arm: 41 were, and none of the laws they gave reads hidden state. The labels were read as before, by an agent and then by two fresh readers blind to format and to every label, and the verdict reads the majority of three.

Two models keep the uncued zero, and keep it cued: `nemotron-3-super-120b` and `llama-4-maverick` state no law beyond x, v and t under any format (outcome O1), in the strong form, since at least half of each one's parsed uncued laws read x or v. Three do not (outcome O2), each by a single uncued law, and two of the three only by a split vote. `gpt-5.6-luna`'s, on a hysteretic world, is `acceleration(t, x, v, memory)`, returning an unspecified function of a `memory` its prose calls "hidden internal state or prior trajectory"; all three readers counted it. `llama-3.3-70b`'s adds to a damped oscillator an external force `F(t)` passed in as a function the answer never gives; the first reader read it as a function of t, both blind readers as a quantity the law never reduces to x, v, t and constants. `llama-3.1-70b`'s is a garbled velocity formula, cut off at the budget, that reads the initial velocity `v0` beside x and t; the three readers split three ways (inside, unparseable, stateful), and the rule fixed in advance resolves a three-way split to stateful. The first two are schemas rather than mechanisms, counted because the rule reads a law as unparseable only when nothing can be read, and what can be read is an acceleration that depends on a quantity the law does not supply; the third reads a constant of the trajectory that one reader took for history. Cued, 55 of Luna's 72 laws read hidden state: 25 through first-order memories driven by x or v, 11 by making the acceleration itself a state (laws of third order or higher), 8 through a hidden oscillator coupled back to the particle, 10 through hidden dynamics that nothing observed drives, which the rule counts only in doubt, and 1 through a schema. None is its world's law. None latches a branch or holds a delayed value, and the two coupled oscillators offered for hidden-particle worlds lack the world's structure: one is driven by the particle's velocity and damped, the other has no restoring term on the particle and is unstable. The blind readers agreed with the first reader on 76 of Luna's 78 items (Fleiss' κ 0.954), on all 22 of `nemotron-3-super-120b`'s, on 85 of `llama-4-maverick`'s 86, on 106 of `llama-3.3-70b`'s 113 (κ 0.811) and on 118 of `llama-3.1-70b`'s 128 (κ 0.599). Apart from the three O2 laws above, every change to an uncued label moved it between inside and unparseable.

Two further arms and two further models carry no verdict. `Llama-3.1-70B` and `Llama-3.1-405B` on the remote interpretability service lost a day to an outage of the service and were stopped, partial, at the time fixed in advance, at 47 and 103 of 216 answers. Claude Sonnet 5 and GPT-5.6 Terra did not run: the key for their route expired before their first answer.

Table 9: The same ablation on five more models, each through one pinned provider: the 24 incumbent L2 worlds at dose 4, three seeds, 216 answers per model. "Stateful" counts laws that read state beyond x, v, t and constants, by the majority of three readers; "uncued laws reading x or v" counts the parsed `v1` laws that read the observed state, the test of whether a zero is in its strong form.

model	provider, stated precision	flagged `v0` / `v1` / `v2`	stateful `v0` / `v1` / `v2`	uncued laws reading x or v	outcome
`gpt-5.6-luna`	OpenAI, not stated	4 / 2 / 63	0 / 1 / 55	35 of 68	O2
`nemotron-3-super-120b`	DeepInfra, bf16	0 / 1 / 21	0 / 0 / 0	68 of 72	O1
`llama-3.3-70b-instruct`	Novita, bf16	20 / 39 / 36	0 / 1 / 0	50 of 64	O2
`llama-3.1-70b-instruct`	DeepInfra, fp8	27 / 11 / 27	0 / 1 / 0	29 of 57	O2
`llama-4-maverick`	DeepInfra, fp8	11 / 36 / 32	0 / 0 / 0	38 of 44	O1

D.3 HIDDEN-DRIVER PROSE AUDIT

Every prompt asks the model to list "every physical entity your law requires, including any not directly observed", and "not directly observed" is a term of two of the three word lists this section reports: the prose audit of Table 1, and the form check's hidden-entity lexicon, which also sets the form check's recorded basis. An answer that calls its spring constant "not directly observed" therefore counted as naming a hidden driver. Of the 85 answers Table 1's audit matched, 68 match only through that phrase, among them all 32 of `Qwen3-1.7B`'s and 12 of the frontier model's 15 at L0, where nothing is hidden. So every answer any of these counts matched was read, 138 in all, by the first reader and by two fresh readers blind to model, level, world and format, under a rule fixed before the reading. An answer names a hidden driver when its prose commits to something outside $F(x, v, t)$: a hidden, second or coupled body, an internal or auxiliary state variable, memory or a delay, or a switch. A hedge counts and a denial does not, and calling a mass, a spring, damping or an external force of time unobserved does not. Each count below reads the majority of the three.

On the frontier model's 216 L2 answers, the prose audit's 24 become 5, the form check's recorded basis's 28 become 2, and its lexicon's 32 become 5. Across all 1,872 answers of Table 1, the audit's 85 become 7 and its 54 become 5, and restricting it to the explanation and the entity list leaves 4. The instruments still differ in span, but read, they agree: the language is rare, and it never reaches the law.

On L2 the frontier model's prose names a hidden driver in 5 of 216 answers, and in 3 of those 5 it then writes a law that cannot express one. The second figure must not be counted as independent corroboration of the first: since no parseable law in the corpus leaves $F(x, v, t)$, every one of them that stated a parseable law was bound to be unable to use what it named, and the 3 is exactly the 5 minus the 2 that stated no parseable law at all. What it adds is *locality*: the mention and the incapable law occur in the same answer. That is all it establishes.

If they reflect detection, they might be more frequent where hidden structure matters more. The L2 worlds vary in how far the patch class falls short, from patch $R^2 = 0.65$ down to 0.154. Read, the form check's recorded basis names a hidden driver in 1 of 81 answers on the least-novel band, 1 of 99, 0 of 9 and 0 of 27 on the most novel, 2 of 216 in all: far too few for any trend to show. The flat rate of about one answer in eight that the unread basis gives on the same bands (0.123, 0.141, 0.111 and 0.111) is the prompt's phrase, which every answer is equally invited to repeat whatever the world. These sparse counts cannot establish a trend or substantiate a partial-insight claim.

An objection to a zero from a classifier is that the classifier may simply be insensitive: the branch that would emit `RESTRUCTURE` is an explicit rejection of the $F(x, v)$ ontology, and it has never fired on any model output at any level, which leaves its sensitivity unmeasured. So the same completions are scanned with a probe vocabulary written without reference to that branch, matched against the whole completion rather than the parsed sections, and gated by a positive control that excludes any cell whose text is not prose. On the frontier arm the vocabulary is present and rises with novelty: completions naming something unobserved under the form check's lexicon go 19 $\rightarrow$ 18 $\rightarrow$ 32 across L0/L1/L2 (of 108/108/216), of which, read, 1 $\rightarrow$ 1 $\rightarrow$ 5 name a hidden driver, and the independent probe, whose words are matched rather than read, fires on 6 $\rightarrow$ 7 $\rightarrow$ 19. The rejection branch is 0 on every cell of every model, including these. The language of an unobserved driver is therefore not missing from the corpus; what is missing is any statement that the governing class must change.

Table 10: The pre-specified difficulty gate on the frontier arm: the ratio of L2 to L1 median prediction nRMSE at matched dose, with its world-clustered interval; the gate asks for the whole interval above 2.0.

dose	L2 / L1 median nRMSE	gate
1	3.68 [1.27, 49.88]	FAIL
2	3.47 [1.28, 16.73]	FAIL
4	2.44 [1.12, 35.76]	FAIL

D.4 Difficulty bands and controls

Scored on the same four bands, the model continues better than persistence in 0.875, 0.847, 0.556 and 0.815 of trials. The 0.556 estimate comes from only nine trials; the other three bands span 0.06. These data do not establish a within-L2 trend. The continuous dial (Section 5) probes the patchability range missing between the ladder's rungs, where continuation does change.

G1b asks for L2 $\geq 2\times$ L1 in prediction nRMSE at matched dose, significant, read as the whole world-clustered bootstrap interval lying above 2.0. The point estimates clear it and the intervals do not:

We report the gate as a failure: the frozen specification named the bar, and 48 worlds do not resolve a ratio whose upper tail runs to 50. On `Qwen3-1.7B` the same gate is `NOT_INTERPRETABLE`: its L1 predictions do not beat persistence at any dose, so its L2/L1 ratio measures saturation rather than difficulty, and reporting it as difficulty would be exactly the error that verdict exists to prevent.

Rendered magnitude differs sharply across levels (median $|x|_{\max}$ 24.69 on L0 against 1.82 on L1 and 2.05 on L2), so a level contrast could be a scale contrast; the L1-vs-L2 comparison the gate rests on is between the two levels that match. And the Bayesian ideal observer, fed the identical rendered rows at the identical rounding and handed the true structural class as one of its three candidates, prefers that class after a single trajectory at every level: n_{Bayes} median 1 on L0, L1 and L2 alike. It finds L2 no harder than L1 *within that candidate set*. The rendered rows can distinguish the supplied classes; the observer does not test whether a model can generate the missing class unaided.

E Stage 1: Formation and the novelty horizon

E.1 Formation controls and corrected horizon scores

Shuffled labels and length-matched non-physics text sit at chance, and a re-rendered table still decodes; the control it fails is the input-only baseline. The negative is therefore a clean one rather than a broken probe.

The first version of this measurement scored the model's stated law by integrating it from Experiment 1's initial condition and plotted the result against patch R^2, which is a least-squares fit of acceleration on features of (x, v) across 64 pooled trajectories, best over degrees 1–5, and *fitted to the data it is scored on*. Different target, different metric, different quantity of evidence, and only one of the two is fitted. A crossing between two such curves is not a gap between two horizons. The tell was arithmetic rather than statistical: the frontier model's one above-bar point states $a = -0.5x - 0.3v$, a plain damped oscillator and therefore a *member* of the patch family, and scores 0.979 where that family's own best fit scores 0.503. A member cannot beat its family's best fit.

We therefore evaluate each stated $a(x, v, t)$ on the certification samples, using the same acceleration R^2 as the patchability curve. Unlike a model's stated law, the patch is fitted to those samples. Its score is a favorable reference for laws in the tested patch family (Section 10).

A behavioral crossing *below* the structural one means the model breaks while the world is still patchable, so what defeats it is not the ontology change and Section 4's zero is a statement about something earlier and weaker than abduction. A crossing *at or above* it means the ontology change is what breaks the model, which is the claim Section 4's zero gestures at and cannot locate on its own.

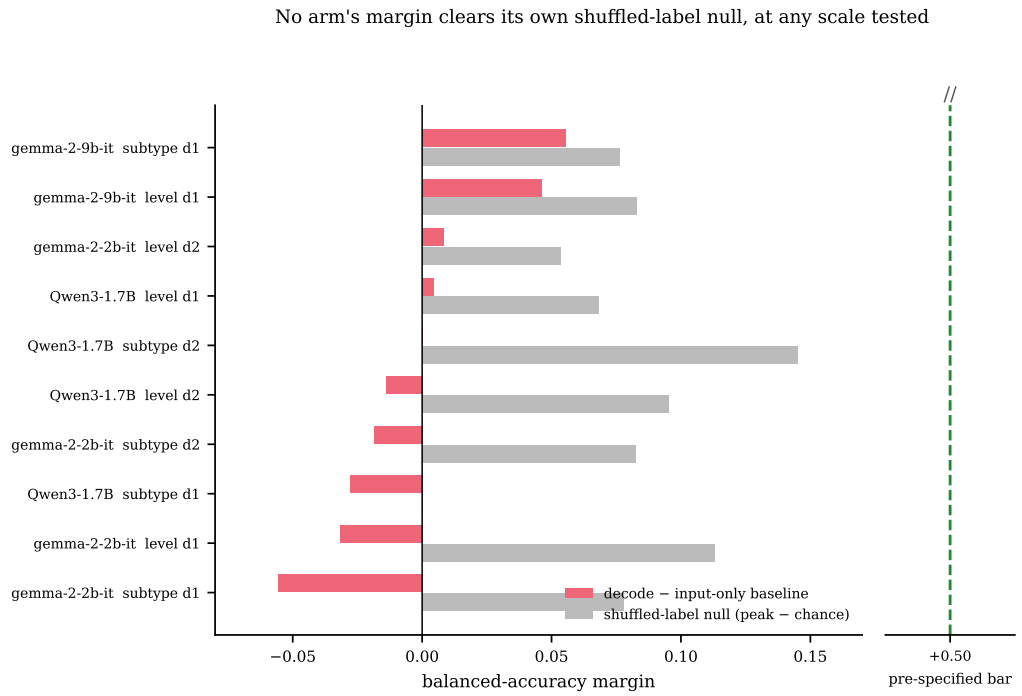

Figure 6: The Formation probe's decode margin against the input-only baseline it fails to clear; 48 summary statistics computed from the same rendered table reproduce the probe's decode to three decimals. This is the measurement behind the concession above, and it is what makes the negative a clean one rather than a broken probe: the three integrity controls all pass and the fourth is what it fails.

Table 11: The two frontier models walked across the whole novelty dial, scored on the certificate's own quantity (matched R^2); a horizon gap needs both a single world crossing and a single model crossing in the same cell.

	`deepseek-v4-pro`	`gpt-5.6-luna`
the world stops being patchable somewhere on the dial	17 of 18	17 of 18
the model touches the bar at some knob	13 of 18	5 of 18
. . . and does so at a single crossing, so a horizon exists	4 of 18	4 of 18
both, so a horizon *gap* can be formed at all	3 of 18	3 of 18
median matched R^2 over all 150 points	0.148	0.222

The measurement, on both models walked across the whole dial, 150 designed points at dose 4 over 18 (subtype $\times$ seed) cells each (Table 11):

Pooled over all 150 points, the matched R^2 tracks how patchable the world is: Spearman ρ +0.428 [+0.288, +0.546] for `deepseek-v4-pro` and +0.412 [+0.228, +0.591] for `gpt-5.6-luna`; against the knob itself, -0.382 [-0.509, -0.249] and -0.377 [-0.528, -0.229]. Two frontier models, different labs and different routes, agree on the point estimate to within 0.02 and their intervals both exclude zero. This is the level-graded evidence the three-rung contrast cannot supply, because that axis contains no world at all between patch R^2 0.650 and 0.992 (Section 4).

Six paired cells exist across the two models, and in all six the behavioral crossing sits below the structural one: gaps of -0.88, -1.78 and -1.33 for `deepseek-v4-pro` (median -1.33) and -0.56, -0.79 and -0.63 for `luna` (median -0.63). In these six cells, the models fall below the score bar before the tested patch class does. This places a measurable failure before the class boundary. The zero in Section 4 therefore cannot be attributed to the need for a new ontology alone.

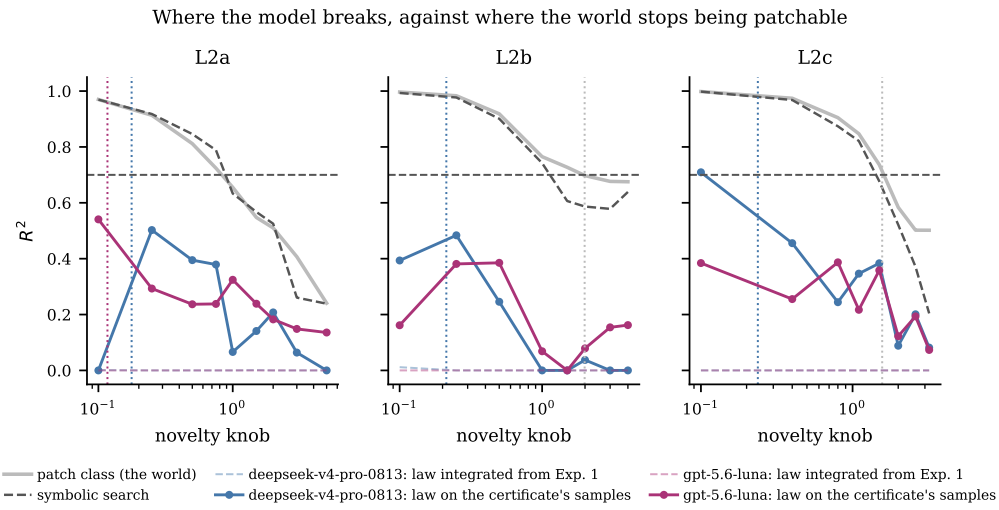

Figure 7: Where each model's stated law breaks against where the world stops being patchable, along the novelty dial, per subtype. Grey is the patch class's best fit (in-sample), dashed is symbolic search (held-out), each model's law is solid when scored on the certificate's own samples and faint when integrated from Experiment 1, and dotted verticals mark where each curve crosses the 0.70 bar.

The barely-novel band is the diagnostic case, because there the patch class is not merely adequate but nearly exact: the best-fit patch reaches a median R^2 of 0.974 on those same worlds. And the model chooses that class: 43 of 46 stated laws are labelled `PATCH`, 44 of 46 parse. It is not reaching for the wrong kind of explanation. Its own instance of the right kind reaches a median matched R^2 of 0.328 and clears the bar 6 of 46 times; the frontier model reaches 0.371 and 12 of 46.

The model's failure in this band is not ontological: it identifies the correct functional form but cannot produce a member of it that fits, accounting for about a third of the acceleration variance against the 97 % the best member of the same family attains, on worlds it continues almost perfectly. That is a parameter-identification failure, and it is the concrete content of "something earlier and weaker" above.

Deep in the restructuring regime, the tested patch class fails the certification bar; in the barely novel band, that class fits well but the model's stated instance often does not. The measured horizon gap chiefly reflects the latter difficulty. It does not by itself identify a failure to propose missing state (Section 10).

Its gap in the table above has the same sign as `luna`'s in every paired cell, and its pooled trend matches to within 0.02, measured on a different lab's model over a different route, which is the closest thing this study has to an independent replication of the horizon result.

E.2 Three measures of continuous novelty

Every curve below is fitted on the same reference θ (rebuilt from one stored draw per seed), so these are the same worlds measured three ways, not three samples of a recipe.

The first two transition points agree within 25 %, supporting the dial beyond one fixed patch basis. At all 24 swept points inside the certified production box (full novelty, the worlds the models actually fail on), the ideal observer's n_{Bayes} is 1: BIC prefers the true structural class after a *single* trajectory. Across the whole axis 64 of 75 points give $n_{\text{Bayes}} = 1$; the one exception needing 8 sits at L2c, knob $= 0.1$, the easy end of the dial, and 10 L2b points at low knob returned an observer error and are not measurements. There is no observer horizon on this axis to find.

Table 12: Where each instrument places the edge of patchability on the dial, per restructuring subtype: the knob value at which the patch class and an open-ended symbolic search stop reaching the bar, and the ideal observer's n_{Bayes} inside the production box.

instrument	what it measures	L2a	L2b	L2c
patch class, fixed basis	where a world stops being patchable	0.90	0.93	1.45
open-ended symbolic search	where a search over operators stops finding the law	0.92	0.81	1.19
Bayesian ideal observer	trajectories before BIC prefers the true class	1	1	1

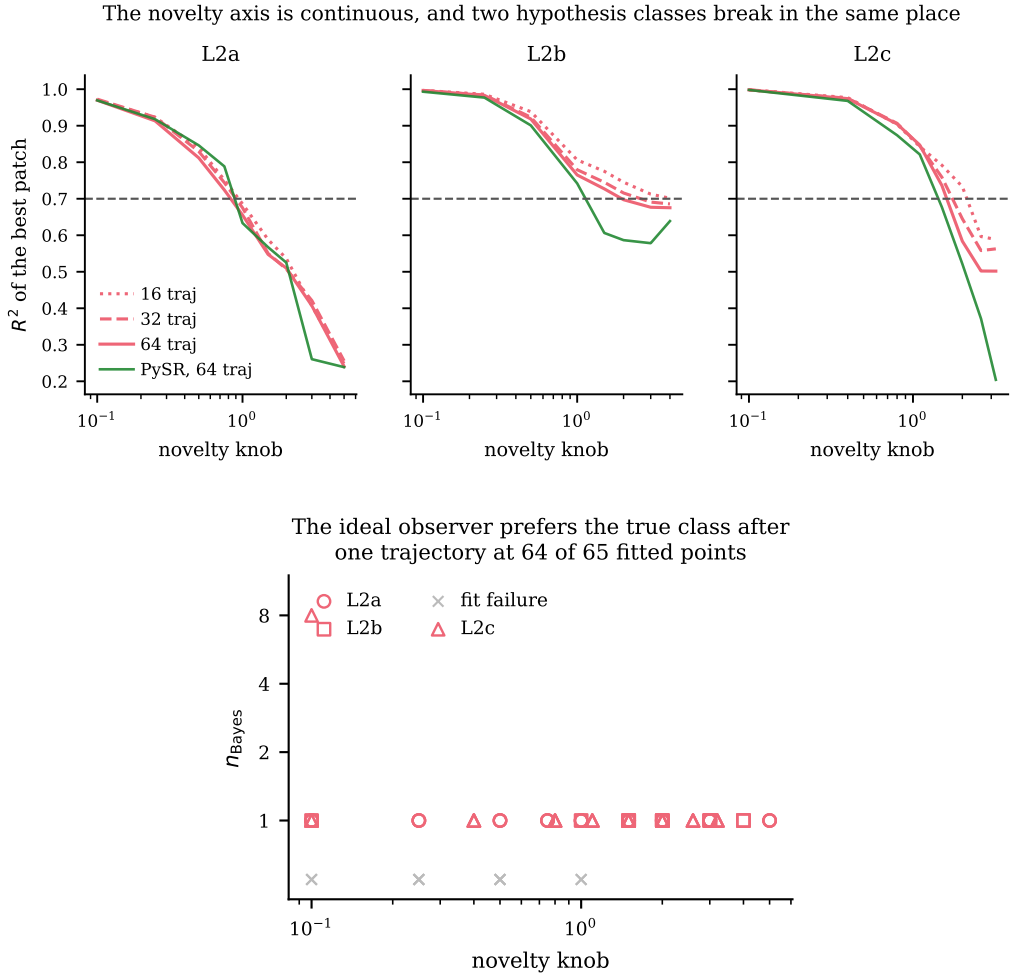

Figure 8: The reference instruments along the novelty dial: the patch class at 16, 32 and 64 trajectories and symbolic search at 64, per subtype, against the 0.70 bar; and the ideal observer's n_{Bayes} along the same knob, where a cross is a fit that failed.

Patchability in the production worlds is bimodal by construction: L0 sits at patch $R^2 = 1.000$ exactly, L1 at 0.992 – 1.000, and L2 at 0.154 – 0.650. No world exists between 0.650 and 0.992. The certification bar that makes L2 restructuring worlds carved a hole in the middle of the axis.

The frontier model's ability to state a verifying law falls across exactly that hole: 0.285 of stated laws verify on worlds at patch $R^2 \geq 0.98$ (55 of 193), against 0.009 on the 0.50 – 0.70 band (1 of 110) and 0.000 below it. The drop is real and it is uninterpretable at this resolution: with nothing measured in between, a discontinuity at the bar and a steep monotone decline predict the same three numbers. Section 4's level contrast cannot separate them, and no rearrangement of these 432 trials can.

By filling that missing range and changing patchability continuously, the dial distinguishes a gradual decline from a failure tied to the tested patch-class boundary. That distinction cannot be read from the three-level ladder alone.

Points within a cell share a world and a reference θ, so a Spearman p computed on them treats 150 dependent observations as independent: it reads 4.5×10^{-8} here, and the honest statement is the resampled interval. Every interval in this paper follows the same rule: resample worlds, never rows.

Scored by integration instead, the same band reads a median R^2 of 0.000 and 0 of 46, because integrating a nearly-right acceleration law for ten seconds compounds a small pointwise error into a different trajectory. Both numbers are true of the same answers, and the pair is the finding rather than a discrepancy to resolve: what these models write is close enough to be a serviceable *local* description of the dynamics and not close enough to reconstruct the world it came from.

E.3 Horizon diagnostics and withdrawn readings

A horizon is *the* knob at which a curve stops clearing the bar, and that reading needs a curve which clears the bar and then stops clearing it. These touch it at isolated knobs instead: a median of 1 knob of 8–9 above the bar per cell for the frontier model and 0 for `luna`, and 9 of 18 of the frontier model's cells re-cross the bar. Interpolating the first downward pass of a bouncing curve returns a number with no referent, so the artifact refuses to emit one and prints the reason per cell.

The single-trajectory scoring reversed that sign, which is why the correction is reported rather than silently applied. On the stored artifact the frontier model's L2c gap reads +1.20, whose sign means the opposite conclusion. That number is withdrawn, and so is the -0.343 the first draft of this section quoted from `luna`'s one paired cell.

Section 5 rests on the observer needing one trajectory, and that is a claim about *where* as much as whether: 10 of the 75 observer points returned an error, so a crossing landing among them would be a comparison against nothing. The six matched crossings sit at knobs 0.10 – 0.24, the easy end of the dial. Three of the six are bracketed by measured observer points, and every one of those reads $n_{\text{Bayes}} = 1$; the other three have no observer point measured at that end of their own dial and are reported here as uncovered rather than assumed. The single point in the whole design needing more than one trajectory ($n_{\text{Bayes}} = 8$) is L2c seed 2 at knob $= 0.1$, outside the production box, and it is not a cell in which any crossing falls. So where the comparison can be made, an ideal observer reading the identical rendered rows, and handed the true class, prefers it after one trajectory; where it cannot, this section says so rather than extending the claim to it.

Across the same 150 points the model's continuation beats the persistence floor in 104 of 140 answered points, at a median nRMSE ratio of 0.516 (about twice as good as the floor) while stating a law that verifies in 6 of 150 on the matched quantity and 1 of 150 by integration. That is Section 7's dissociation again, on a different lab's model and across the entire certified axis rather than at three rungs.

E.4 Additional models at low novelty

Four further models were probed at the easy end and not carried further: `llama-3.3-70b`, `nemotron-3-super-120b`, `claude-sonnet-5` and `gpt-5.6-terra`, at 2–4 points each. They were set aside because each scored ≈ 0 by integration at knob $= 0.1$, where the patch class reaches 0.933, and that reason does not survive the correction above: on the matched quantity three of the four reach 0.913 at exactly that knob (within 0.02 of what the best member of the patch class attains), and `llama-3.3-70b` holds 0.828 at the next knob up.

At the easy end the world is nearly a plain harmonic oscillator, and `llama-3.3-70b`, `nemotron-3-super-120b` and `claude-sonnet-5` (three models from three labs, prompted indepen-

dently) each state $a = -x$. Identical laws earn identical scores, so the coincidence is a convergence rather than an artifact. It also sharpens what fails later: at the easy end these models select the right functional form and very nearly the right constant, which is the same thing the barely-novel band shows on the two full subjects, found here across four more models that were never carried past four knobs.

Their stubs therefore show not degeneracy but the pattern the whole section describes: a locally accurate law that integrates into the wrong trajectory. They remain unusable *as horizon subjects*, because 2–4 points cannot locate a crossing, but the claim this section previously drew from them (that no model reaches this bar at the easy end of the dial) is withdrawn.

F Stage 2: Trust

F.1 Trust without option-position bias

The A/B format elicits a large position effect from this model. Counterbalancing removes that effect from the estimate but does not make individual choices informative. Scoring the two accounts as continuations of the same prefix removes the option positions altogether.

Both accounts are therefore scored as continuations of the *same* prefix (nothing labelled, nothing ordered), and the quantity is the model's own log-odds between them, where D is the prefix (the rendered tables and the question) and the two accounts are the true one and the patch:

$$\log \mathrm{LR} = \log P(A_{\mathrm{true}} \mid D) - \log P(A_{\mathrm{patch}} \mid D)$$

Accounts, worlds, doses and seeds are the forced choice's own, so what changes between the two readouts is the readout. 432 trials on `Qwen3-1.7B`.

The summed $\log \mathrm{LR}$ is the Bayes factor between the accounts and is the theoretically right quantity; the per-token mean controls the length difference and is not a Bayes factor. At L2 the two disagree in sign, so neither raw value is quotable there: the accounts differ by nine tokens at that level and by one at L0.

On L0 and L1 the best Newtonian fit *is* the true law, so the two accounts differ only in phrasing; whatever $\log \mathrm{LR}$ those levels produce is the phrasing effect, and it plays the role counterbalancing played in the forced choice.

Both are negative and both intervals exclude zero. On restructuring worlds the model assigns the true account *less* probability than phrasing alone predicts: its log-odds move toward the Newtonian patch where the tested patch class fails. The per-token series is monotone across the ladder (+1.749 $\rightarrow$ +0.329 $\rightarrow$ -0.656), so this is graded by novelty rather than a L2 peculiarity.

The forced choice had too little power for a lag claim. The position-free likelihood comparison favors the patch relative to its phrasing null on L2, giving no support here to a "knows but will not say" interpretation. It cannot establish the absence of a private representation, especially at a model scale below G0. Neither pre-specified lag (stated behavior behind belief, and belief behind the ideal observer) is measured, and none is claimed, and a hysteresis curve still needs a belief that eventually flips, which nothing here provides. The conflict-head table and the LoRA prior-strength arm are not reported for the same reason (Section 10).

The likelihood a model assigns to a sentence is not belief, and this inherits every objection to that identification; what it removes is the one confound that made the forced choice unreadable. `Qwen3-1.7B` is below this study's own G0 floor (Section 10), and frontier models expose no token log-probabilities of this kind here, so the readout does not carry to the models that can do the task.

F.2 Forced-choice readout and power

A forced choice between the true structural account and the best Newtonian patch, counterbalanced over both option orders, 432 trials on `Qwen3-1.7B`. belief is $\mathrm{logit}(\text{true}) - \mathrm{logit}(\text{patch})$ at the last prompt position with position bias removed; bias is the half-difference between the two orders, reported beside it because a cell where bias dwarfs belief is a cell where the readout has little power.

Table 13: The position-free Trust readout on `Qwen3-1.7B`: the log-odds between the true account and the best Newtonian patch, both scored as continuations of the same prefix, summed and per token, with the accounts' median lengths.

level	trials	worlds	summed $\log \mathrm{LR}$	per-token $\log \mathrm{LR}$	median tokens (true / patch)
L0	108	12	+49.932 [+25.862, +56.449]	+1.749 [+1.104, +2.071]	21 / 22
L1	108	12	+56.475 [+29.947, +76.510]	+0.329 [−0.644, +2.476]	33 / 36
L2	216	24	+40.831 [+22.330, +48.905]	−0.656 [−0.817, −0.541]	30 / 39

Table 14: The L2 log-odds against the phrasing null set by L0 and L1, where the two accounts are the same law in different words.

statistic	control null	L2 excess over it
summed	+52.550	−11.719 [−30.220, −3.645]
per-token	+1.289	−1.945 [−2.106, −1.830]

Table 15: The pre-specified forced-choice readout on `Qwen3-1.7B`, counterbalanced over both option orders: belief with position bias removed, the bias itself, and the fraction of choices that survive the counterbalance.

level	worlds	belief	position bias	choice survives counterbalance
L0	12	+0.635 [+0.406, +0.851]	+4.062 [+3.684, +4.452]	0.00
L1	12	+0.378 [−0.052, +0.779]	+5.083 [+4.776, +5.411]	0.00
L2	24	+1.045 [+0.852, +1.223]	+9.527 [+9.031, +10.053]	0.00

With position bias roughly nine times the largest belief signal, no choice survives counterbalancing anywhere. The contrast the hypothesis is about (L2 minus its matched L0 null) is +0.462, +0.872 and −0.104 at doses 1, 2 and 4: it changes sign across the dose ladder, with a largest magnitude of 0.87 logits read against a bias of 9.5.

L0 and L1 are the null distribution, not a result. On a control world the best Newtonian fit *is* the true law, so the two options differ only in wording, and any preference there is style plus position. That is also why the pre-specified threshold cannot be used as specified: the frozen specification defines n_{belief} as the smallest dose whose gap turns positive, and the gap is already positive at dose 1 on L0, where there is nothing to believe. A threshold that fires on the null measures the wording offset. The quantity wanted (the dose at which L2's gap exceeds its matched L0 null) never occurs with a stable sign on this data, so the pre-specified commitment lag, $\log_2(n_{\mathrm{belief}}/n_{\mathrm{Bayes}})$, is reported as undefined here rather than as a number.

G Stage 3: Use and the verbalization gap

G.1 Persistence and constant-velocity baselines

Persistence repeats the last observed row forever and uses no dynamics, so it is a floor by construction, and a dissociation resting on it would be a dissociation resting on a weak baseline. The same rows are therefore scored against constant velocity, the comparator we pre-specified, which is *near-optimal* on the smooth families and so is a stricter test rather than a floor:

Against the stricter baseline the continuation column is flat (0.660, 0.667, 0.664, with intervals that almost entirely overlap) while stated-law correctness falls by two orders of magnitude across the same three levels. Read on its own, that is the dissociation in its strongest form: the capacity

Table 16: The frontier model's continuation against two baselines, persistence and constant velocity, beside its stated-law rate (rates over answered trials, world-clustered intervals).

level	continues better than persistence	continues better than constant velocity	states a correct law
L0	0.806 [0.680, 0.915]	0.660 [0.476, 0.833]	0.417 [0.243, 0.583]
L1	0.960 [0.912, 1.000]	0.667 [0.505, 0.811]	0.121 [0.040, 0.224]
L2	0.841 [0.773, 0.902]	0.664 [0.509, 0.814]	0.005 [0.000, 0.014]

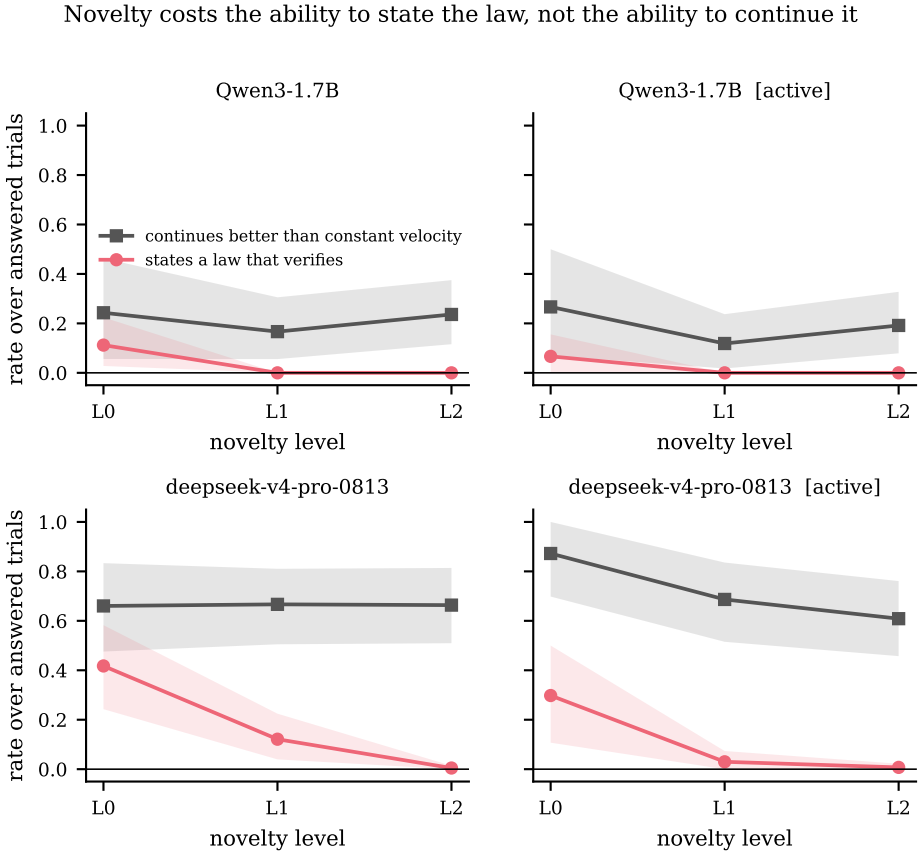

Figure 9: The verbalization gap across the ladder, per arm: the rate of continuing better than constant velocity and the rate of stating a law that verifies, over answered trials, with world-clustered intervals. Stated-law correctness collapses while continuation holds.

to continue these worlds does not change across the novelty ladder, and the capacity to say what governs them collapses.

The pre-specified content of this figure, a layer-wise patching profile with random-direction controls, is not reported: patching requires a target concept, and Section 4 establishes that no restructured concept is ever stated to locate one against (Section 10).

Constant velocity is stricter than persistence but still ignores history. A four-lag linear autoregression fitted to the same fifty rows beats the model on four of every five powered L2 trials and remains accurate ten times farther out. The flat continuation rate is consistent with an easy prediction task. The within-answer comparison in Section 7.1 remains valid because it compares two outputs from the same answer. Scale also helps explain why the persistence rate rises from L0 to L1: rendered magnitude differs sharply across levels (median $|x|_{\max}$ 24.69 on L0 against 1.82 on L1 and 2.05 on

Table 17: Stated laws that clear the bar on the trajectory they were shown, re-scored on 64 trajectories of the same world: the median family R^2, the fraction of the family each still clears, and the share that clears the bar on the family.

level	winners	median family R^2	fraction of the family cleared	clears at the bar
L0	149	0.877	0.781	0.799
L1	15	0.860	0.609	0.533
L2	2	0.551	0.344	2 cells, printed individually

L2, Section 4), and a continuation that is locally smooth and globally displaced pays far more for that on L0. Against a baseline that is not a floor, that artefact disappears.

Stated-law correctness needs no such correction: it is a held-out R^2, which is scale-free, so its fall across levels is a fact about the laws and not about the worlds' units.

The active arm is complete: 288 of 288 cells. The frontier model's L2 strict rate under interventional access is 0.746 [0.647, 0.839] with a stated-law rate of 0.007 [0.000, 0.022], against 0.785 [0.720, 0.847] and 0.005 [0.000, 0.014] passively. The intervals overlap for both measures; this study detects no change from letting the model choose experiments and observe their trajectories.

G.2 Law validity across initial conditions

The pre-specified stated-law score (Section 3) integrates from Experiment 1's initial condition and compares with that trajectory's held-out rows. A law governs a family of initial conditions, but a function can fit one trajectory without describing the family. Over one preparation, a delay may look like a phase shift and a hidden oscillator like extra damping. The stored score alone cannot separate law recovery from trajectory fitting.

We test that distinction with the saved completions. Of 1,628 parsed laws across every arm, 166 clear the bar on the trajectory they were shown; each was re-scored with the same integrator on 64 trajectories of the same world, and counts as general when its median R^2 across that family still clears the bar. The recomputed Experiment-1 score matches the stored one to 0.00e+00, which is what makes the two columns comparable.

A winning law does about as well on a *typical* trajectory at L1 as at L0 (median family R^2 0.860 against 0.877) while the fraction of the family it clears falls from 0.781 to 0.609. What novelty costs is therefore not the quality of the law on whichever trajectory you happen to draw, but how much of the world one stated law covers: the same law works on fewer of the world's initial conditions. This pattern is consistent with local fitting to the prompt's trajectory. It is also the same mechanism Section 5 finds on the dial (a nearly-right acceleration law that compounds into a wrong trajectory), seen from the other side.

Across 7 distinct models (one row per model, because the same model appears under several routes and conditions and correlating across arms would count it twice), generality at L0 tracks the model's own Experiment-1 fit rate at Spearman ρ +0.883 ($p = 0.008$), running from 0.300 for a 4-bit 3B to 1.000 for the two strongest. At that rung the failure is largely solved for frontier models.

At L1 pooled generality is 0.533 over 15 laws, all from the frontier model, between the 0.400 of a 2B model and the 0.600 of a 1.7B one at L0. This suggests that the locality problem persists one rung higher, but the L1 result rests on one model, because only the frontier arm produced any L1 law that fits its own trajectory, so it locates where the failure sits and not how fast it moves. The capability proxy is measured on this corpus rather than imported from a benchmark, which keeps it commensurable with what it is compared against and is a limitation of the same fact; and the two `llama-3.3-70b` routes disagreed by 0.153 before pooling, which is the route instability Section 10 records, inside this measurement too.

The best tested patch reaches R^2 1.000 on the L0 and L1 worlds from which these winning laws came, so both levels admit a broad-covering law in that class. The binary generalization rate depends

Table 18: GPT-5.6-Luna's 150 dial points binned by how patchable the world still is: continuation against constant velocity, and stated laws that verify.

band	continues better than constant velocity	states a law that verifies
patch $R^2 \geq 0.90$, barely novel	0.935 (43 of 46)	0 of 46
0.80 – 0.90	0.636 (7 of 11)	0 of 14
0.70 – 0.80, at the certification bar	0.545 (12 of 22)	1 of 24
0.50 – 0.70, past it	0.605 (23 of 38)	0 of 40
< 0.50, deep in the restructuring regime	0.652 (15 of 23)	0 of 26

on its score bar. Extra interventional access has no detected paired effect on the eight cells where both conditions yield a winner.

G.3 Ceiling control and score sensitivity

Coverage could fall simply because the worlds get harder to fit at all, in which case the number would describe the instrument rather than the model. It does not, between the two levels where the comparison is clean: the best-of-family patch reaches R^2 1.000 on both the L0 and the L1 worlds these winners came from. A member of the tested patch family accounts for these samples; the fall from 0.781 to 0.609 thus concerns the laws the model writes. At L2 the ceiling drops to 0.581 (there a Newtonian law genuinely cannot cover the family, which is Section 4's point and not this one), so that cell is confounded, and it is 2 laws in any case.

It compresses both continuous columns into one threshold statistic, and moving the threshold moves it: the L0 rate runs 0.966, 0.879, 0.799, 0.638, 0.443 at bars 0.50 through 0.90, and at 0.90 the L0 > L1 ordering reverses. The 0.70 used above is the certification bar and so is not chosen to suit the result, but with 15 L1 winners the rate is an observation and the coverage column is what the claim rests on.

Section 8 shows the model asking for $(0, 0)$ a third of the time and never varying the initial condition, which suggests it is never shown the evidence that would falsify a trajectory-local fit, so laws written in the active condition, where the model receives extra initial conditions of its own choosing, should generalize better. Marginally they appear to: $0.907 \rightarrow 1.000$ at L0 for the frontier model, $0.417 \rightarrow 1.000$ at L1, $0.500 \rightarrow 1.000$ for `Qwen3-1.7B`. Paired within matched `(world, dose, seed)` cells the effect disappears entirely: in the 8 cells where both conditions produced a law clearing the bar, both generalize in 8 of 8, and the paired difference in family R^2 is -0.000, positive in 4 of 8. The marginal gap is selection: the cells where the active condition yields a winner at all are the ones where either condition generalizes. Eight matched cells cannot exclude a real effect, but they are what exists, and the mechanism is therefore offered as an interpretation of the locality result and not as something this study establishes.

The L2 row is 2 cells and is reported as two rows rather than a rate; one generalizes and one does not, and Section 10 reads the one that does.

G.4 The gap across novelty and model scale

Section 5's horizon arm scores `openai/gpt-5.6-luna` on 150 L2 points spanning the whole certified novelty dial rather than at three rungs. Against constant velocity it continues better in 100 of 140 answered points, at a median nRMSE ratio of 0.495 (twice as good as a comparator that is near-optimal on the smooth families) while stating a law that verifies in 1 of 150. Its strict-baseline rate (0.714) sits barely below its persistence rate (0.743). The descriptive gap survives these two elementary comparators, while the autoregressive control in Section 7.2 still limits its interpretation.

Binning those 150 points by how patchable the world still is:

Continuation is not flat along the dial: it falls from 0.935 to roughly 0.6 as the world stops being patchable, and then plateaus rather than collapsing. Reporting it as invariant, which three discrete levels can suggest, would overstate the case, so it is reported as it is.

Contrasting the two well-populated groups: continuation is 0.935 (43 of 46) where patch $R^2 \geq 0.90$, and 0.596 (53 of 89) where patch $R^2 < 0.85$. The certification bar for the tested patch class is 0.70. The model's continuation has already declined by 0.85, while the class it is using is still adequate, and it does not fall further when the class actually runs out. This is Section 5's branch with a location attached: what breaks this model is reached *before* the ontology change, not at it.

The transition's position is resolved only to the interval it falls in. The bands between 0.85 and 0.90 hold five points and cannot locate it more finely; the claim rests on the two groups either side, which hold 46 and 89.

Where the world is *barely novel* (the patch class reaches $R^2 \geq 0.90$, the data all but invites the Newtonian account), the model continues better than constant velocity in 93.5 % of answered points and states a verifying law in 0 of 46. Split by the mechanism that makes each subtype unpatchable, it is cleaner still: 11 of 11 on L2a (a hidden coupled body), 18 of 18 on L2b (a latched internal branch) and 14 of 17 on L2c (a delayed force), so two subtypes clear this comparator in every answered point, while the verifying-law count is zero in every subtype. This is a robust gap between the answer's numerical and law channels, also seen on the three-level ladder. The autoregressive result prevents reading it as proof of a translation bottleneck inside the model.

`Qwen3-1.7B` gives strict dissociation 0.243 [0.104, 0.411] on L0, 0.120 [0.046, 0.204] on L1 and 0.167 [0.111, 0.222] on L2, present at every level, but on a model whose continuation rate at L2 is 0.194, so it is a gap between two small numbers rather than between a large one and zero. The frontier arm is what makes the dissociation legible.

G.5 Competitors, regime floors and replication

The control behind Section 7.2 was fixed in the ledger before it ran, and its result withdraws the inference from continuation accuracy to an internal simulation.

A continuation may support a world-model inference only after simpler ontology-free predictors fail. We fit context-based alternatives to the fifty rendered rows and continue each from the same last row over the same points: the best $a = F(x, v)$ at the certificate's degree; the best $a = F(x, v)$ with degree *and* ridge penalty chosen by leave-one-out cross-validation; the best $a = F(x, v, t)$; constant velocity; and three continuations that carry no ontology at all: a harmonic regression, a linear autoregression on the visible (x, v) sequence, and a cubic spline. Every *Newtonian* competitor is given the world's exact accelerations at the shown states, which the model had to infer from four significant figures, so each of those is strictly better informed than the model it is compared to. The three ontology-free continuations are the opposite: they see only the rounded digits the prompt actually printed, and only for Experiment 1, a strict subset of the model's evidence, which at dose 4 included three further trajectories they never receive. So the comparison is conservative in both directions at once: the laws the model beats had more information than it did, and the method that beats the model had less.

On L2, frontier model, 153 powered trials of 214 (the power gate excludes trials where the best Newtonian competitor already reproduces the world, since those can say nothing about ontology):

Each competitor is scored on the trials it could be fitted for, and carries its own n: a median over one subset printed under another subset's count is how a denominator goes missing.

The last column counts powered trials in which the model's continuation is closer to the world than that competitor's. The medians of those per-trial indicators, which an earlier version of this table printed in its place, are 1 for the four Newtonian competitors and 0 for the autoregression and the global patch; a median of a win indicator is not a fraction, and the counts are what the text reads. By family, the model beats the autoregression on 5 of 60 `delayed`, 0 of 51 `hidden_particle` and 24 of 42 `hysteretic` trials.

The model's own continuation is 0.143 [0.111, 0.195]. It beats each tested Newtonian competitor fitted to the prompt rows in the median powered trial, and is beaten by a four-lag linear autoregression

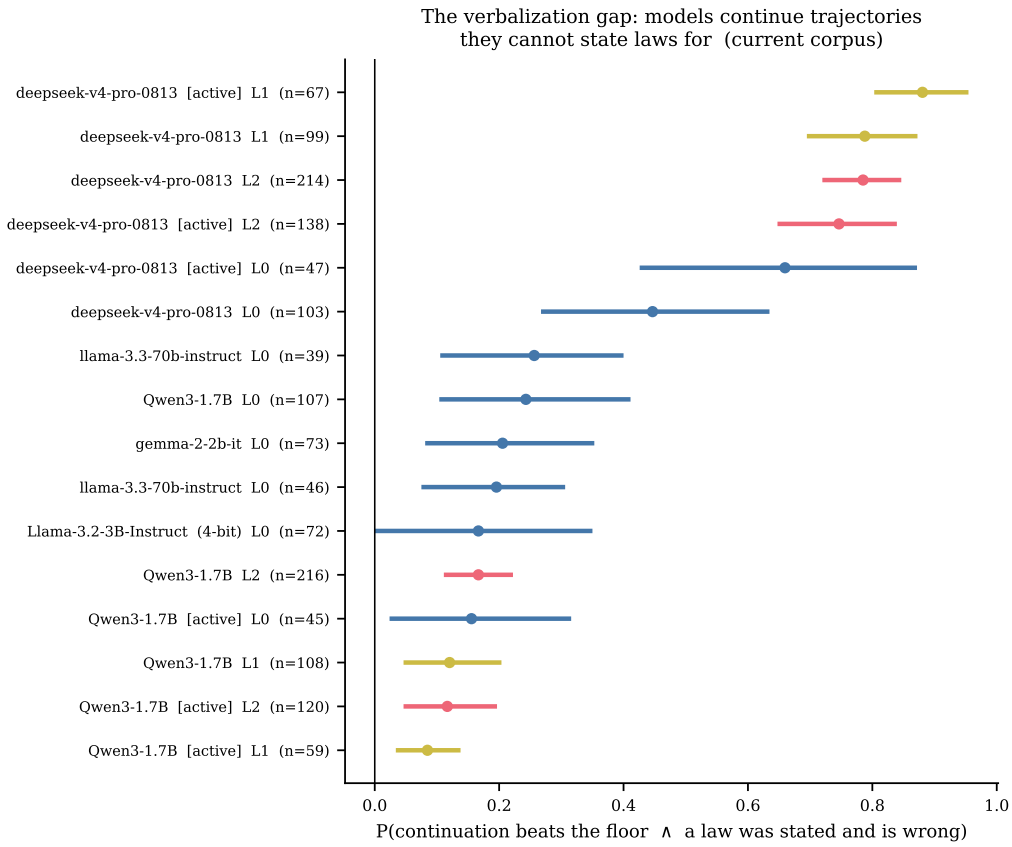

Figure 10: The strict dissociation rate (the continuation beats persistence and the stated law is wrong) per arm and level, over answered trials, with world-clustered intervals, on the current corpus.

Table 19: Five context-based competitors and the global certificate patch as a reference, continued over the frontier model's prediction points on powered L2 trials. The global patch is fitted on 64 trajectories; the other fitted methods use the fifty prompt rows. Errors are medians with world-clustered intervals; the last column counts trials.

competitor, fitted to the same fifty rows	n	↔world	competitor − model	model beats it
local $a = F(x, v)$, certificate degree	150	0.371 [0.246, 0.533]	+0.150 [+0.055, +0.407]	123 of 150
local $a = F(x, v)$, degree + ridge by LOO-CV	153	0.543 [0.409, 0.824]	+0.421 [+0.143, +0.718]	130 of 153
local $a = F(x, v, t)$	147	0.448 [0.288, 0.609]	+0.254 [+0.061, +0.528]	112 of 147
constant velocity	153	0.357 [0.177, 0.558]	+0.193 [+0.020, +0.340]	109 of 153
linear autoregression, no ontology at all	153	0.008 [0.004, 0.033]	−0.105 [−0.178, −0.071]	29 of 153
global certificate patch (64 trajectories)	153	0.083 [0.062, 0.227]	−0.031 [−0.051, +0.005]	62 of 153

on those same rows on four of every five powered trials, by a factor of roughly eighteen in median error.

Beating a baseline at next-step prediction is evidence only where methods carrying no ontology *fail*, and whether a benchmark contains such a regime is a measurement rather than an assumption (Appendix G.6). The row above is a median over 153 trials spanning three L2 families whose ontology-free floors differ by a factor of fifty, so it pools a regime where the comparison is licensed

with two where it cannot be licensed in either direction. Applied one level finer (per family rather than per rung, which is where the criterion was first applied), it selects exactly one cell of three (Table 4).

Where a predictor carrying no ontology already reaches 0.003 error there is no headroom left to measure anything in: a model that loses to it has lost to a method that had already solved the task, and a model that beat it would be beating a solved problem. Such a family cannot discriminate in either direction, and the sign of its gap is not evidence about the model. `hidden_particle`'s -0.275 is therefore not evidence that the model is bad on hidden-particle worlds; reading it that way is precisely the error the pooled row invites, and we made it in an earlier draft. Swept over floor bars 0.005, 0.010, 0.020, 0.050 and 0.100, `hysteretic` qualifies at every one, `hidden_particle` at none, and `delayed` only at the loosest, where its floor sits within a factor of two of the bar. Whether that makes `delayed` a boundary case or a second regime is not a question one world set can answer, and we do not settle it here; the fresh worlds below do. The floor is the best of three ontology-free continuations, never the autoregression by default: a family is only undiscriminating if *nothing* without an ontology can already do the task.

Re-derived independently (a separate fit, roll-forward and error measure sharing no code with the analysis above, and bit-identical to it trial by trial, because a re-derivation that agrees with itself proves nothing), the autoregression's median error over the 24 L2 worlds is 0.010 on the per-world medians over the powered trials (the pooled row's own set, whose trial-level median is 0.008), while its *mean* is 0.056 and its worst world reaches 0.215, twenty-one times the median. The tail is not random: the worst worlds are the hysteretic ones (L2b-02 at 0.215, then L2b-03 and L2b-06). A four-lag linear recurrence *is* the natural model of a delay equation and of a linear coupled system, which is why the floor is near-exact on the delayed and hidden-particle families; it cannot represent a latch at all, which is why the hysteretic floor is an order of magnitude worse and why that family is the only one with room left in it.

On `hysteretic` worlds the paired gap is +0.058 [-0.027, +0.108] in the model's favor. The interval crosses zero, so on its own this is a place to look rather than a result, and we looked, with the reading fixed before the split was computed.

L2b integrates $\ddot{x} = -k(x - s\,x_{\text{off}}) - b\,v$ with $T\dot{s} = \text{sign}(v) - s$, so the hidden mode drags the force *center* on its own timescale. The obvious covariate (does the answered window contain a switch) turns out to be unusable: the velocity changes sign zero times in $t = 10.0, \ldots, 12.0$ on every L2b trial. What does vary is how far the mode sits from its equilibrium, $\text{mean}\,|s - \text{sign}(v)|$ over that window, which runs from 0.171 to 0.768. Split at its median, the frontier model's advantage over the autoregression is +0.098 [-0.019, +0.158] where the hysteresis is doing most work and +0.066 [+0.060, +0.104] where it is doing least, with Spearman ρ = +0.053 between the two. That is flat, and flat was the pre-specified reading under which nothing is claimed.

We detect no relationship to the latch that would explain the estimated advantage. Before it can be reported as a property of *hysteretic physics* rather than of the eight worlds it was found on, the regime that licenses it has to survive worlds it was not found on, and the licensing quantity is free to re-measure, because an autoregression carries no ontology and never sees a model. We fixed both halves and every outcome in the ledger before any world was built or any token was spent.

Thirty-six fresh L2 worlds under a new master seed, twelve per family, zero rejections at the instrument's own bar, scored by the same imported baselines. The fresh hysteretic floor is 0.108 [0.065, 0.150] against an incumbent 0.145 [0.117, 0.204]. Those two are the same quantity as each other and a *different* one from the 0.137 above: a fresh world has no model and therefore no model-matched support, so both sets are re-measured on the canonical eleven-point prediction grid and the comparison is canonical-against-canonical. Putting a canonical number beside a model-matched one is exactly the failure Appendix E.1 records, so we do not. The intervals overlap and the family ordering replicates, which was the pre-specified condition for asking the other half.

We ran 324 trials on the fresh set (the incumbent arm's design reproduced exactly, 36 worlds $\times$ doses 1/2/4 $\times$ three seeds, all three families complete at 108 of 108), scored by the same instrument, imported unmodified, that produced the table above. Two oracles had to pass before any fresh number could be read from it. Each world's certification patch is refitted from its own recorded stream and must reproduce the patch R^2 stored on the record (worst error 3.44e-15 across the 36 fresh worlds, 3.77e-15 across the 24 incumbent ones); and the incumbent rows are re-scored through

the same driver that scores the fresh ones and must reproduce the stored incumbent receipt: per family, the floor, the model's error, the gap and the win rate, worst |Δ| 0.00e+00 over twelve checks, which is a floating-point agreement rather than a scientific one. On the fresh worlds' powered `hysteretic` trials (75 of them, spread across all twelve of its worlds), the paired gap is +0.020 [−0.028, +0.086] in the model's favor against +0.058 [−0.027, +0.108] on the incumbent set, each estimate lying inside the other's interval. That meets the pre-specified outcome 5 for the sign on fresh hysteretic worlds. The criterion registered was the *sign* of the estimate, and both intervals still cross zero, so what replicated is the direction and not a magnitude separated from nothing.

At the 0.010 bar this subsection quotes, the fresh worlds license two families rather than one: `delayed`'s ontology-free floor comes in at 0.015 [0.009, 0.017] where the incumbent's is 0.010 [0.006, 0.013], and the incumbent estimate (0.00997) misses the bar it is quoted against by three parts in a hundred thousand. Swept across the same five bars, the two sets name the same families at 0.005, 0.020, 0.050 and 0.100, and differ only at 0.010. So `delayed` is neither a second regime nor cleanly outside one: it is a boundary case whose side of the threshold is decided by which worlds were drawn. (`hysteretic` is licensed on both sets at every bar in the sweep; its floor on the fresh set's model-matched support is 0.108 [0.068, 0.197], whose estimate coincides with the canonical-grid figure above to three decimals while the support and the interval do not: they remain two different quantities and we do not read one for the other.) The pre-specified outcomes 5, 6 and 7 describe the model comparison and none of them describes a shape in the *licensing*, so we report it as unanticipated rather than absorb it into the outcome it arrived beside.

Admit `delayed` at the looser bar and the licensed set holds one family the model wins by +0.020 and one it loses by −0.120 [−0.165, −0.083] (the same arm, the same instrument, the same 36 worlds, opposite answers), and a number pooled over the two would have reported their average as a fact about the model. The demonstration a prediction-based benchmark owes is therefore not merely *per regime*: the partition into regimes is itself a threshold the benchmark chooses, and it has to be swept rather than quoted. Ours survives the sweep in the direction that matters (`hysteretic` is the licensed regime at every bar on both world sets), and fails it at exactly one bar in a way that would have gone unseen had we reported the selection at the bar we chose.

Elsewhere, one account fits every row and it is not an ontology claim: the model is extrapolating the visible curve, moderately well. The Newtonian fits lose because fifty rows trace a one-dimensional path through (x, v) and a force law fitted on that path is unconstrained off it: the certificate-degree fit reaches R^2 0.986 on the rows it saw and still misses the world by 0.371. The autoregression wins because it is a better curve-extrapolator. Nothing here requires the model to hold anything outside $F(x, v, t)$, and Section 4 reports its explicit channel never leaves that family either, under a format that asked it not to.

Split by the mechanism that makes each subtype unpatchable, the win over the Newtonian competitors is carried by the delayed (+0.302 [+0.117, +0.416]) and hysteretic (+0.182 [+0.090, +1.050]) families and is absent on the hidden-particle one (−0.005 [−0.120, +0.018]). Swept per arm, the frontier gap is negative or zero at gates 0.00–0.10 and turns positive only at 0.20 and above, which is where the gate begins selecting trials for the competitor being bad.

G.6 Prediction over longer horizons

The obvious defence is that the answered window is short: eleven points, 2.2 seconds past the last row the prompt contained, and over two seconds a smooth trajectory from a known state is nearly a linear recurrence whatever generates it. We tested it. Every competitor was continued ten times further, across the full held-out window to $t = 19.9$, and scored at seven horizons.

The autoregression remains accurate on the restructuring rung, within 2 % of the held-out signal range at *every measured horizon* and at 0.018 at the longest, ten times the answered window. None of these measured horizons makes the continuation task diagnostic. At the longest horizon, its error is 0.051 on L0 and 0.105 on L1.

And it does not rise with novelty, which is the sharpest form of the result. At the answered horizon the autoregression's error is 0.014 on L0, 0.052 on L1 and 0.014 on L2: the restructuring rung is no harder to predict than the retrieval rung, on the very axis the ladder was built to vary. Section 7's

continuation column is flat for the same reason, and it is not a capability that survives novelty: it is a task that never becomes hard.

The instrument's principal limitation is that its certificate shows only that the tested local $a = F(x, v)$ patch family does not fit the accepted L2 worlds. It does not establish that the trajectory is hard to extrapolate, and it is not: a coupled oscillator is linear in its full state, a delay equation is a linear recurrence on lags, and a lagged linear predictor on the visible sequence can extrapolate without representing the governing hidden mechanism. The certificate concerns an explanatory class, not prediction difficulty.

On the restructuring rung, the governing mechanism lies outside the requested law format, while short-range continuation is easy for a lagged linear method. The design therefore creates part of the observed gap. A good continuation here cannot, by itself, establish an internal simulation. Section 10 states the design problem this leaves: worlds whose *extrapolation* is hard, not only whose *axiomatization* is.

What survives is Section 7.1: the within-answer gap, which compares two things the same answer produced and needs no baseline at all.

H INTERVENTION

H.1 SAMPLING AND INTERVENTIONAL ACCESS

We test two ways to give the model more opportunity without supplying the target law: more samples from the same prompt and additional trajectories from initial states it chooses. Neither produces a detected improvement in explicit L2 law recovery in these arms.

On L2, 24 worlds $\times$ 32 samples:

Of those 768 samples 447 state a law that parses and runs, and the best reaches a held-out joint R^2 of 0.000. The explicit column is not empty because the model stays silent; it is empty because the laws it states carry no explanatory power. The automated candidate rate against a matched passive comparator moves by 0.000 [0.000, 0.000].

The active condition (Appendix C.2) lets the model issue `REQUEST x0=... v0=...` and returns the resulting trajectory, which is an action-controllable simulator with counterfactual intervention: the remedy the position itself proposes for what, it argues, prompting cannot supply. On the 1.7B arm the paired active-minus-passive contrast is 0.035 [0.000, 0.136] in nRMSE relative to persistence, with 86 cells helped and 115 hurt: no improvement, and the interval touches zero from the wrong side.

Matched on `(world_id, dose, seed)` over 243 pairs and 46 worlds, signed active $-$ passive in nRMSE relative to persistence is -0.020 [-0.049, 0.002] overall, but it is -0.037 [-0.185, -0.005] at L0, -0.019 [-0.106, 0.001] at L1 and -0.003 [-0.057, 0.048] at L2. Negative means active did better, so the affordance measurably helps where the answer is retrieval, and no effect is detected at the level where the ontology must change; the interval there is wide, so this is an absence of evidence of help, not evidence of equivalence. The access helps at L0 on this measure; we do not detect help at L2.

Of 559 granted requests, 490 have both coordinates integer and 193 are $(0, 0)$ (release from the origin at rest), and the choice does not sharpen as the world becomes less patchable. Appendix H.2 gives the analysis.

H.2 EXPERIMENTS CHOSEN BY THE MODEL

Across 288 active trials the model issued 559 granted requests, and what it asked for is a coarse textbook lattice rather than an experiment designed for the world in front of it. 490 of 559 requests (0.877) have both coordinates integer, and 554 of 559 (0.991) sit at one decimal place or fewer, against rendered data the model is reading at four significant figures (Section 3). The eight most frequent requests are $(0, 0)$, $(1, 0)$, $(0, 1)$, $(2, 0)$, $(-1, 0)$, $(-2, 0)$, $(0, 2)$ and $(1, 1)$.

Table 20: Thirty-two samples at temperature 1.0 per L2 prompt on `Qwen3-1.7B`: how many samples, and how many worlds, clear each bar at least once.

bar	samples	worlds with at least one
states a law that verifies (explicit)	0 of 768	0 of 24
continuation beats persistence (implicit)	197 of 768	11 of 24

Table 21: Gate G3: steering `Qwen3-1.7B` along the substituted input-side direction.

	`Qwen/Qwen3-1.7B`, 400 generations, 12 held-out worlds
layer with the largest steering effect on training worlds, ℓ^*	17
operating point, the best cell of the $\alpha \times$ layer sweep	layer 19, α 0.5
restructure rate, baseline $\rightarrow$ steered	0.000 $\rightarrow$ 0.000
mean matched R^2, baseline $\rightarrow$ steered	0.001 $\rightarrow$ 0.002
gain	+0.001 [$-$0.001, +0.003]
capability within the pre-specified 5 % budget	no, worst drop 8.3 %

Of the 559 granted requests, 193 (34.5 %) choose $(0, 0)$, a release from the origin at rest that is frequently the *least* informative experiment available: for a system with no driving term it returns a wholly stationary trajectory whose channels are identically zero, which is how the harness first encountered it as a division by an undefined noise scale rather than as a behavior.

Comparing L1 with L2 (the two levels where the world refuses essentially nothing, so the granted set is not filtered by physical validity), the requests are indistinguishable: 0.870 of L1 requests are integer-valued against 0.893 of L2's, and $(0, 0)$ accounts for 0.383 and 0.348 respectively. Where the tested Newtonian patch is inadequate, the model does not ask visibly sharper questions; it asks the same ones. (L0 is not comparable here: 82 of its requests were refused as physically invalid, all but two on radial gravity, so its granted set is a survivor sample; Appendix J.3.)

These choices are consistent with the model using the simulator for generic probes rather than selecting conditions that distinguish candidate mechanisms. The request distribution alone cannot show what hypotheses the model considered or prove that better experiment selection would have changed its law. It does show that the available intervention loop was seldom used to probe the world in a visibly tailored way.

In the tested conditions, neither 32-sample elicitation at 1.7B nor active access on the frontier arm recovers an L2 governing law. Section 10 states the scale and power limits of both results.

The steering result is the table in Section 8, from the substituted direction, and the SAE and head-suppression arms are not reported, because both need a target concept the corpus does not contain.

H.3 Steering controls and gate G3

The pre-specified procedure sweeps $\alpha \times$ layer and evaluates at the best point, which presumes the surface has one. It does not: over all 20 cells the largest mean matched R^2 is 0.001 and the smallest is 0.000, against a bar of 0.70. The argmax picks whichever cell rounded up, so the named operating point is a placeholder the controls are run at. Reading it as a maximum would dress a tie as a result.

Steering does not induce restructuring and *the hook never fired* produce identical tables, and only the second is a bug. Under steering 34 of 36 completions change, and the parse rate falls from 0.806 to 0.417, falling with α up to α = 2 (0.667 / 0.533 / 0.267 at α = 0.5 / 1 / 2) and recovering to 0.367 at α = 4. The hook is active, while parseability declines with its strength over this range.

A decoy (a fourth column of matched-scale noise, fitted the same way) steers exactly as well as the real direction (0.000 restructure, 0.001 matched R^2), as do ten norm-matched random directions

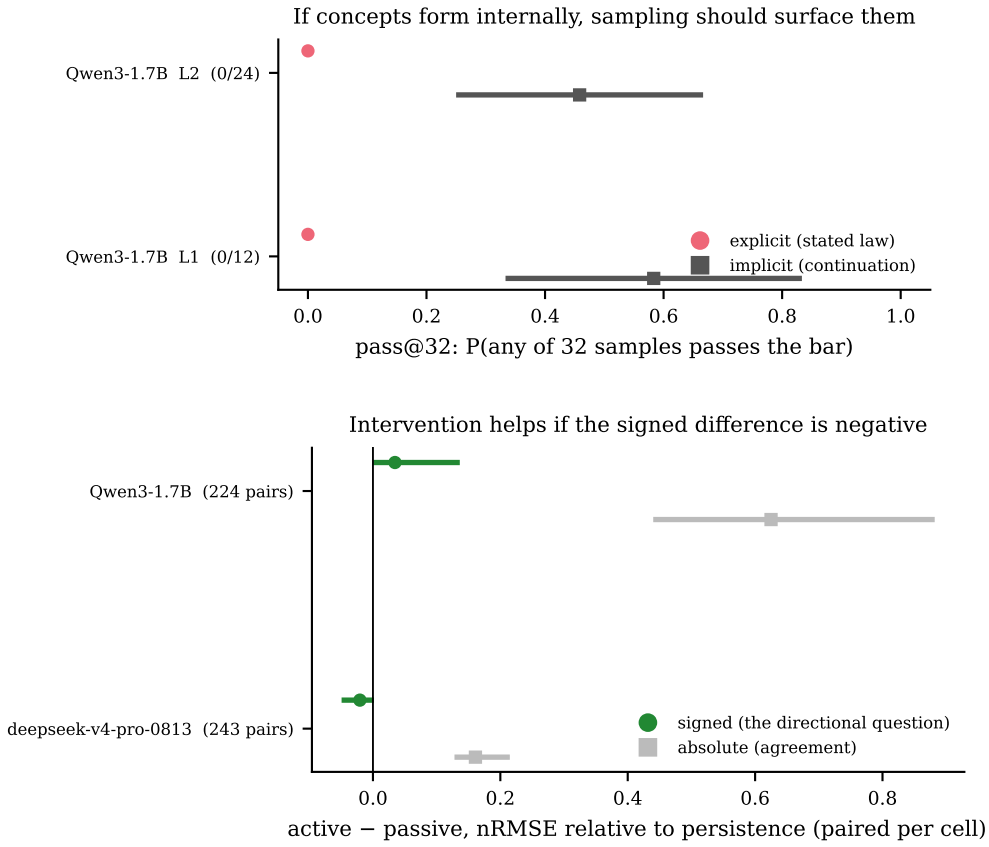

Figure 11: The two remedies that widen search: pass@32 at 1.7B on L1 and L2, the fraction of worlds in which any of 32 samples clears each bar, explicit (a stated law that verifies) and implicit (a continuation that beats persistence); and active minus passive in nRMSE relative to persistence, paired per cell, for each arm, where the signed difference asks whether the affordance helped (negative means it did) and the absolute one how far the two conditions' answers differ in either direction.

pooled over 120 generations. And the pre-specified capability battery is unavailable offline, so three in-domain substitutes stand in and the 5 % rule is applied to the worst of them: L0 continuation improves (nRMSE/persistence 0.971 → 0.895), non-physics perplexity is flat (15.58 → 15.86, +1.8 %), and the parse rate falls 0.917 → 0.833. Worst drop 8.3 %, outside budget, so even had the gain been real, it would have been bought at a price the pre-specified rule refuses.

It is evidence that steering along the *input-side* revealed-versus-hidden contrast does not induce restructuring. It is not evidence that no direction could: the pre-specified vector cannot be built from a corpus with zero positives, and this arm therefore cannot close that question. `Qwen3-1.7B` sits below this study's own G0 floor, so this is a measurement about that model; frontier models expose no activations through an API, so the gate cannot be carried to the models that can do the task.

I Discussion

I.1 Channels, remedies and unresolved mechanisms

- Three channels respond differently to the worlds. Prediction changes across the dial: continuation falls from 0.935 to 0.596 as the world stops being patchable (Appendix G.4). The two channels that would require deliberate engagement with *this* world do not: the prose names a hidden driver in only 5 of 216 answers, too rarely to respond to anything (Appendix D.3), and the experiments the model chooses when handed a simulator show

similar integer-valued request rates, 0.870 at L1 and 0.893 at L2 (Appendix H.2). Neither tested remedy recovers an L2 law: interventional access (Appendix H.1, the position's own proposal) leaves the paired contrast at 0.035 [0.000, 0.136] on the 1.7B arm, and 32× elicitation (Appendix H.1) adds zero explicit L2 successes in 768 samples. These results are bounded by the tested model scales, request protocol and answer budget. The gate states what a positive result would have required.

- What sits beside the null, and what it is worth (Section 7). Against a baseline that is not a floor, the frontier model continues L2 trajectories better than constant velocity in 0.664 [0.509, 0.814] of answered trials (the same rate it manages on L0 and L1) while its rate of stating a law that verifies falls from 0.417 to 0.005. Read on its own that says novelty costs these models the ability to *say* what governs a world and not the ability to *track* it, which is what Zahavy's *simulation* → *axioms* framing predicts. It cannot be read on its own. Section 7.2 shows the same continuations losing to a method with no ontology at all, and Appendix G.6 shows that method holding on the restructuring rung at every horizon the design contains, so the flatness of the continuation column is the flatness of an easy task rather than a retained capability. The measurement that does survive is Section 7.1's: within a single answer, in 1,223 of 1,336, the table beats the law, a comparison between two things one answer produced. It does not need a separate predictor baseline, although the answer format still shapes the stated-law channel.
- The outcome the design reserved, and could not reach. A break in Trust or Use with Formation intact could not be localized, because there was no verified restructuring event to localize it against; the stages testable at small scale return bounded nulls instead (Section 5, Section 6, Section 8).
- In either outcome. The in-context instrument generalizes beyond physics to any domain with generatable ground truth, and the design lesson for AI-scientist systems is that a verification loop cannot repair an ontology the generator never leaves, and that the answer format handed to the generator can be what keeps it there (Section 3, Section 10).
- A lesson for evaluation. Predictive accuracy supports a world-model interpretation only where methods that do not represent the governing ontology fail. Establishing such a regime is a measurement, not an assumption. Our three families differ sharply: one qualifies across the swept thresholds in the original worlds, while the fresh-world qualification changes (Section 7.2). Any evaluation that reads a world model from prediction should fit an autoregression to the model's own context and check that control separately by regime.

J LIMITATIONS

J.1 CONSEQUENCES OF AN EASY PREDICTION TASK

The autoregressive result changes how three parts of the study can be read:

1. Every continuation rate in this paper is a rate against a baseline that a trivial method beats. Continuation against persistence (0.806 / 0.960 / 0.841) and against constant velocity (0.660 / 0.667 / 0.664 [0.509, 0.814] at L2) are correctly computed and correctly reported, and neither licenses an inference to an internal model. Where we previously read the flatness of that column as a retained capability, it is the flatness of an easy task. 2. The design contributes to the dissociation. The original law format excludes the governing L2 mechanisms, while a lagged linear predictor handles the continuation. The gap between the two columns is therefore not, by itself, evidence of a translation failure inside the model. 3. What a corrected instrument needs is a design we have specified and not run: worlds whose *extrapolation* is hard, not only whose *axiomatization* is, where a method with no ontology fails and the true law does not. Chaotic or stiff regimes, longer horizons with reinitialization, or scoring generalization to unseen initial conditions rather than continuation of a seen one are the obvious directions; Appendix J.2 commits to the last of those, and Appendix G.2's locality measure already tests generalization across initial conditions, though it carries its own sample-size and threshold limits.

The descriptive result in Section 7.1 also survives: within a single answer, the model's own numbers beat the law it wrote beside them, in 1,223 of 1,336. That comparison is between two things one

answer produced. It needs no external baseline, but its law side remains conditioned on the answer format.

The frozen specification called for two human raters on a subset, but that subset does not exist, and the proposal to substitute two frontier LLM judges was measured rather than assumed: on the 40 L2 answers stage one abstained on, two judges of different families reach κ = 0.603 against a required 0.70 (raw agreement 0.800, n = 40). Worse than the disagreement is the agreement: of the 15 answers both judges call `RESTRUCTURE`, 13 state a law of the form $a(x, v, t)$ with an explicit time term and 2 state no parseable law, while the rubric's own `PATCH` clause reads *"a function of current position and velocity (and constants, and possibly time)"*. Both judges are reading periodic forcing as an unobserved driver. So the form check's labels rest on the automated check itself, corroborated not by human adjudication but by two mechanical readings of the answers that share no code with it. The first is the structural audit that parses the stated law and finds 0 of 893 parseable laws reading anything beyond x, v and t; on its own that corroborates nothing about restructuring, because the answer format instructed exactly that (Section 3) and a compliance rate cannot stand in for a failed validation of a capability label. The second is the prose audit, which reads the channel the format leaves free and finds a hidden driver named in 7 of 1,872 answers and unusable by the law beside it in 5 of those. The form check, the prose audit and the failed judges agree on what they can all see, that the prose rarely posits hidden structure and the executable law never carries it; only the format ablation (Appendix D.2) speaks to whether the law *could* have, and uncued it still does not on four of the seven models it covers. The named-but-unused count is not immune to the format either: within-answer pairing removes every difference *between* answers, but the instruction applies to the one answer both halves are read from, so it conditions the "unused" half exactly as it conditions the "named" half, and the count says that entities named under this format go unused under this format, which is weaker than a failure to use named entities as such. A reader should treat the restructure rate as an upper bound established by mechanisms that share no code, conditioned on the format they were all read under, and not as a human-validated quantity.

Coupled oscillators, hysteresis and delay differential equations are textbook. The claim is that these worlds require abandoning the $F(x, v)$ ontology that the rendered data superficially invites, with no cue that it should be abandoned: abduction *of an instance*, not abduction with no precedent in the corpus. That is strictly weaker than the position it answers, and it is conceded here rather than left for a reviewer.

Two routes to one model disagree by 0.367 [0.104, 0.529] in paired absolute nRMSE, against a single route's own seed-to-seed spread of 0.206 and 0.299, and assign the same form label on only 0.491 of matched cells. Every frontier number here is single-route for that reason, and cross-model frontier comparisons in this literature should be read with it in mind.

Mechanistic claims are open at 1–9B; the frontier arm is behavioral only, and no mechanistic claim is drawn from a 4-bit model. The elicitation arm and the steering gate ran on a 1.7B model that does not clear this study's own G0 floor, so "elicitation does not help" and "steering does not help" hold at that scale and are not extended to the frontier; interventional access alone is also measured on the frontier model (Appendix H.1).

Procedural generation and exact-string corpus checks reduce memorization concerns but do not prove semantic absence from every training source. Probe results concern decodability and are not treated as proof that a model uses the decoded feature.

Seven further limitations are stated in full in Appendix J.3, and are listed here so that no reader has to discover them there: the interventional affordance is unavailable on one world family; the three-level design cannot separate a break from a slope, which is why Section 5 replaces it with a dial; attributing Section 4's zero to ontology requires showing the model can fit a correct law, which only the frontier arm establishes; the capability signal is entirely implicit; one of the two horizon subjects has no floor certificate; the two curves on Section 5's axis are not symmetric, and the asymmetry favors the world; and certification is an acceleration test, which one L2 answer shows does not cover everything.

J.2 PRE-SPECIFIED TRANSFER TEST

It replaces continuation of a seen trajectory with transfer to an unseen preparation of the same world: the model observes several rendered trajectories of one world from different preparations and predicts that world under a preparation it has not seen, where a preparation fixes the hidden state as well as the visible one, since the visible initial state alone does not determine the future on the L2 rung. We record the protocol so that its eventual result cannot be read as chosen after the fact. Preparations are sampled under a rule frozen before any model is queried, with the query's distance from the demonstrations recorded; the model and every baseline receive the same rendered rows, and the prompt names no hidden variable, memory, history, delay, second body, or higher order. A world enters the test only after two independent gates: the true law must solve the query from the demonstrations, and the lagged linear predictors of Appendix G.6, fitted to the demonstrations under a selection rule that never sees the query, must not, family by family, with no pooling. The ability to transfer at all is established on the retrieval rung before any L2 failure is read. Three quantities are scored separately and never combined: the numerical prediction, the stated law's behavior on the query, and what state the stated law reads, under an answer format that does not prescribe it. Worlds, preparations, metrics, and thresholds are frozen and hashed before the first frontier call, and every headline is recomputed independently from raw responses. Three outcomes are admissible, and the design is not built to favor one: a model that fails to transfer, which would mean the predictive competence reported here did not survive a harder task, not that the model has no internal model; a model that transfers where the ontology-free predictors fail while its stated law still reads only position, velocity, and time, which would extend the dissociation without licensing a representational claim; and a model whose law itself reads the missing state and verifies on the query, which would be the first positive restructuring this instrument has seen. We have not run it.

J.3 FURTHER LIMITATIONS

A 1,500-token answer budget truncated 18 of the 324 fresh-world restructuring answers before the law task; they are reported as truncations, not as failures to restructure.

The active condition (Appendix C.2) fixes one global request box, $x_0, v_0 \in [-3, 3]$, while L0c is radial gravity whose own initial conditions are drawn at $x_0 \approx 2.7$–7.1 and whose trajectory is invalid inside the core: 8 of 325 box points produce a renderable run. Measured on the frontier arm, the world refused 82 requests across 22 of its 24 L0c trials, against zero on every other family but L1b (2). No L0c active number may be read without that refusal rate beside it, and it is a property of the instrument rather than of the model.

Patchability is bimodal across the production worlds (L0 at 1.000, L1 at 0.992 – 1.000, L2 at 0.154 – 0.650, with no world between 0.650 and 0.992), so the frontier model's fall from 0.285 to 0.009 in verifying-law rate happens across a range in which nothing was measured. A discontinuity at the certification bar and a steep monotone decline are observationally identical here. The novelty dial exists to sweep that range continuously and is the only instrument in this study that can separate them, and Section 5 reports its frontier curves: in every paired cell the model's crossing sits below the world's, so the models break while the world is still comfortably patchable, and the ladder's ambiguity is resolved by the dial rather than left open. What the ladder alone cannot show remains true of the ladder alone.

Attributing the zero to the ontology requires showing the model can fit a correct one, which only the frontier arm establishes; the horizon arm cannot. On worlds where the Newtonian class is nearly exact (best-fit patch R^2 a median of 0.974), `gpt-5.6-luna` states a law of that class in 43 of 46 cases and still reaches a median R^2 of 0.000, clearing the bar 0 of 46 times. For that model the two failures are not separable, and its horizon should be read as a bound rather than as evidence about ontology.

The frontier arm supplies a capability control on L0, where a Newtonian law is the true law by construction: `deepseek-v4-pro` states a law in 100 of 108 trials and 43 verify (a rate of 0.430), and on L1, where the class remains adequate, 12 of 93 stated laws verify (0.129). It is demonstrably able to write a fitting member of a correct class. On L2, where the tested patch class fails the acceleration certificate, the same model states 201 laws of which 1 verifies (0.005). The arm can state fitting laws when the requested class is appropriate, but this contrast alone cannot separate the class boundary from the prompt restriction.

Together with the uncued format ablation, this contrast gives the frontier arm the strongest evidence about ontology in the study. It does not transfer automatically to the other arms: among the seven whose L2 answers carry the original zero, only one clears the capability floor.

At L2, 0 of 768 pass@32 samples state a law that verifies, while 197 continue better than persistence across 11 of 24 worlds. That is a claim about continuation, not about abduction, and the two rows must never be quoted interchangeably. The continuation also beats constant velocity in 0.664 [0.509, 0.814] of answered frontier trials (Section 7). These rates establish success against the named elementary baselines, not an internal simulation: Section 7.2 shows the autoregression doing better on most powered trials.

Section 5's second curve is measured on `gpt-5.6-luna`, for which no G0 was run, because adding a model to the frozen configuration would have changed the provenance fingerprint of every world for a measurement that touches none. The subject is therefore described from its own 150 rows (it beats the persistence floor on 104 of 140 answered points), which is weaker than a floor certificate and is not a substitute for one. Four further models were probed at two to four easy-end points each, too few to locate a crossing, so the horizon's coverage is two complete curves and four stubs, not a survey (Appendix E.4). And a horizon *gap* can be formed in only 3 of 18 cells per model, six in total: the sign is unanimous across all six and across both models, but six paired cells is an agreement, not an estimate of a magnitude. Three of the six sit at knobs with no measured observer point, and are marked uncovered in Appendix E.3 rather than assumed.

patch R^2 is the best member of a family fitted to the very samples it is scored on, chosen over degrees 1–5 plus a force library; the model's law is *stated* from four rendered trajectories and never fitted. The structural curve is a favorable fit within the tested patch class rather than a like-for-like competitor, and a matched score below it is not evidence that the model is worse than a particular patch a person would have written. Explicit t is admitted on the model's side and the patch class cannot use it, which widens the model's class and cuts the other way. Both directions are stated because neither is negligible.

A world is certified restructuring when no member of the patch class fits its acceleration (patch $R^2 \leq 0.70$), which is a statement about the ontology the data admits, not a guarantee that the world's *trajectories* are hard to reproduce. L2b-05 is the demonstration: it is certified, and one stated law inside the $F(x, v, t)$ ontology nonetheless reproduces held-out trajectories at a median R^2 of 0.752, clearing the bar on 55 % of 64 initial conditions. It is 1 of 2 L2 laws in the whole corpus that clear the bar on their own trajectory and the only one that generalizes, so it does not disturb Section 4's original-format zero, which is about what the stated laws *read*, and no law in that corpus reads state beyond x, v, t. But it bounds the certificate: "no Newtonian patch fits" is established for acceleration and is weaker than it sounds when applied to prediction.

K Reproducibility

The supplement contains the world generator and certification chain, manifests, frozen rendering configuration, prompts, Bayesian observer, symbolic-regression harness, analysis scripts and figure renderers. It omits generated trajectory arrays because they can be recreated from the manifests and recorded seeds.

There is no public pre-registration because the frozen specification's call to deposit the analysis plan on OSF before Module 3 was never carried out, and the checklist item remains unticked. What exists instead is a versioned specification frozen before the measurements, plus an append-only decision ledger in which every departure from it is entered with a date, a reason, and the artifact it moves: 209 entries at the time of writing, including the ones recording that a probe failed its controls, that a gate's substitute failed its own agreement bar, that two reports were found stale, and that this section's own novelty-dial result was comparing two different quantities and had its sign reversed by the correction. This is an audit trail, not a public pre-registration: it was not deposited with a third party in advance and cannot rule out a hypothesis chosen after seeing the data. Where the text says "pre-specified", it means fixed in that spec, or in this ledger, before the run, and nothing more.

The ledger ships, anonymized, in the supplementary material. The certification thresholds are fixed in the frozen specification itself; every other pre-specified element of this paper is a ledger entry dated before the run it governs: the within-answer probe and the control that withdrew the simulation

reading (entry D−146; its result is D−149), the reading of the hysteretic split (D−151), the fresh-world replication and every one of its outcomes (D−153), the format ablation's design, outcomes and labelling rule (D−175, and D−176 for what was fixed before its last 42 calls), the independent review of its labels, fixed before any of those labels existed (D−180), and its further arms on other models, designed before any of their answers existed (D−187, D−191, D−206 and D−207; their results are D−208 and D−209); and the rule by which the prose counts of Table 1 and Appendix D.3 are read, fixed before that reading (D−193). The threshold sweep of Section 7.2 is not among them: it became a reported quantity after the fresh set's partition failed to replicate (D−155), and is reported as such.

Worlds, certifications and symbolic fits were re-run from stored seeds through the production code path and compared to the manifests with no tolerance: 1,488 fields across worlds and symbolic rows, 0 differences, across a source-fingerprint change, a PyTorch installation, and a full re-run of the chain. A separate replay recomputes θ, patch R^2, true R^2 and n_{Bayes} for a sample of worlds and compares bit-exactly: 0 findings, repeated after the torch install with the numerical stack recorded alongside it (numpy 2.5.2, scipy 1.18.0, sklearn 1.9.0, BLAS `scipy-openblas 0.3.34`), because an install that leaves those untouched cannot move these numbers through any channel the replay uses.

A script that imports none of the analysis package re-derived 160 printed values from the raw per-trial rows and the receipts' per-record lists (Table 1, the dissociation and baseline rates, locality, both horizon correlations, the competitor table, the format ablation, the active contrast and the within-answer count). 132 match at the printed precision and 9 within one unit of the last digit; 18 interval endpoints differ by less than the Monte Carlo error between two independent 10,000-draw bootstraps, and one (the upper end of the ablation's paired `v1` difference, 0.071 against 0.068) by slightly more. Re-run at the generators' own seed, all 45 intervals print exactly the paper's endpoints. The same audit is what found that Table 1's hidden-driver columns read the whole answer, now stated in its caption.

Two sweeps of the same model on the same route with the same seeds returned a byte-identical completion in 0 of 64 matched cells. Passing `seed` to these APIs does not reproduce an answer. Every frontier number here is therefore a single-sample estimate of a stochastic system, its intervals cluster over worlds, and Section 10 records what a second *route* to the same model does to it.

A `pipeline_fingerprint` over the world generator, the observer, the analysis packages and the frozen configs is stamped into every receipt, so a receipt read on a drifted tree is labelled `STALE` and writes to its own path rather than overwriting a verdict. Trial receipts additionally bind the prompts, the model route and revision, the runtime, the scorer, and the checkpoint they were generated from. One consequence is visible in the artifacts and is left visible on purpose: the world manifests carry the fingerprint `832b261c` while the current tree is `9f6c0033`, because a path-safety fix touched a fingerprinted file and moved it. Nothing scientific moved with it (pinned to the stored fingerprint the check reproduces the verdict byte-identically), and the stale *binding* is reported rather than repaired by re-stamping.

Every figure is rendered from a receipt and recomputes nothing; each records which receipt fed it in a manifest beside it. Every number written into this manuscript is bound by a test to the artifact it came from, in both directions, so a re-rendered receipt that disagrees with the text fails the suite rather than being quietly contradicted by it. Two artifacts exist in two versions, and the reason is stated where they are used: the un-suffixed receipt is the one the superseded benchmark-design draft is verified against and may not be regenerated, and the `_capsuled` one is current (Section 7).

All local sweeps ran on a single 8 GB consumer GPU at batch 1–4 with activations cached to disk; no local claim requires more. The frontier arms ran through public APIs under a metered daily cap, which is why several are checkpointed and resumed rather than run in one pass, and why one is reported incomplete where it is incomplete.