CAN LLMS TAKE THE PLUNGE? FIRST, MAP THE STAIRS

Anonymous authors
Paper under double-blind review

ABSTRACT

Can large language models generate genuinely new hypotheses that may lead to scientific discoveries, or just combine the training data in a sophisticated way? We make this broad question testable by decomposing abduction (generation of explanatory hypotheses from observations) into a staircase of diagnostic steps. One path tests whether LLMs can detect and act on their own errors; the other climbs from selecting among supplied explanations, to recognizing structural correspondences, to constructing unsupplied hypotheses with testable consequences. For routine abduction, the tested LLMs underweight base rates relative to sample evidence, a pattern Grether (1980) documented in humans; across increasing model sizes, both fitted weights rise but their imbalance persists. Internal error signals become most informative where outputs are confidently wrong. In a toy model tracked across training, the internal signal becomes more informative than output uncertainty after a grokking-like transition. Across two frontier-model families, structural-recognition accuracy remains near chance in a single pass (0.50–0.54) but rises to 0.98–1.00 when hidden serial computation is permitted. In radical abduction without supplied candidates, when contradictory evidence signals a failure, models reach the correct hypothesis class in one world family, and occasionally produce correct final programs, but do not complete the full evidence-bound chain. Without such a trigger, two frontier-model families do not distinguish rule-bearing worlds from matched rule-free ones unless the hypothesis family is supplied. When we supply the method, a second pass repairs 110 of 114 first-pass errors. An interior reader can allocate retries but without a conclusive advantage over a trained output reader. Computation and supplied structure improve performance substantially, while constructing and testing explanations without those supports remains unreliable.

1 BEFORE THE PLUNGE

Every genuinely new idea in science begins as a guess that did not have to be true. Peirce, who reasoned this out alone and in poverty, gave that *guess* a name and its place in logic: *abduction*.

He first called it *hypothesis* and placed it beside induction and deduction, completing the trichotomy "Deduction, Induction, and Hypothesis" (Peirce, 1878), and refined the division across his career:

Deduction (rule + case → result): all beans in this bag are white; these are from the bag; so these are white. What *must* be.

Induction (case + result → rule): these beans are from the bag; these are white; so all beans in the bag are white. What, on repetition, *actually* is.

Abduction (rule + result → case): all beans in this bag are white; here are white beans; so perhaps these came from the bag. What *may* be, i.e., generating a hypothesis to explain an observation.

His claim was that *abduction is the only form of reasoning that introduces genuinely new ideas in science* (Peirce, 1934). Those ideas earn credibility only when deduction derives their consequences

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.

and induction tests them[1]. Without that second stage, a guess might be a conspiracy theory. Abduction is the well-educated wanderer-mind proposer, not the judge.

Abduction runs along a spectrum, and where a case falls on it decides what we should even expect of a model. At one end, **routine abduction** selects among a *small set of hypotheses already in the repertoire*: "the whole road is wet this morning, so it probably rained last night", or a doctor diagnosing a patient from symptoms. We term the other end **radical abduction**, where the candidate hypotheses are *not supplied* and the relevant one may have to be found in a huge space, so the explanation must be retrieved or constructed rather than merely selected. The division is a spectrum, not a partition, yet it shapes how we design our experiments.

Zahavy (2026) suggested[2] that LLMs cannot do abduction (in our terms, radical abduction) because they are not grounded in the real world through the felt experience of a body, and so cannot make groundbreaking scientific discoveries either. Their sharp example is Einstein's conviction that gravitational and inertial mass are one and the same, the equivalence principle, which he seems to have *felt* before he could derive it.

Zahavy's *felt experience* argument leads straight to what philosophers of mind call *qualia*, what redness or pain feels like from the inside as opposed to the functional description of the process. Chalmers (1995; 1996) argues that no functional or physical explanation accounts for *why there is something it is like to have an experience*, and calls this the "hard problem of consciousness". In contrast, Dennett (1991) argues that experience is built from many ordinary cognitive processes, with no extra ingredient left over. For the LLM cognition question, this fork is immensely important. If Dennett is right, any cognitive action, including felt insight like Einstein's, is made of processes shaped by past experience (nurture) and hardwired neurons (nature). Some of those processes may leave a signature we can look for. We cannot settle whether an LLM feels anything, and we do not try. We ask a narrower question that can be measured. Which parts of abductive inquiry can a frozen model perform on its own, and what changes when we supply computation, candidate explanations, or organizing structure?

The stakes are no longer hypothetical. LLMs have contributed a new lower bound on the proportion of zeta zeros on the critical line (Claude, 2026) in mathematics and designed functional bacteriophages (King et al., 2026) in biology, real results both in formal domains and out of them. Yet the results alone cannot tell us whether they are genuine abductive leaps or sophisticated recombination. Our working hypothesis is that *finding* an explanation is not one act but several, each able to fail on its own. Rather than ask whether LLMs can plunge, we try to map the uneven stairs.

Radical search has at least two starting points. A structural resemblance can invite the search: two distant ideas rhyme, as gravitational and inertial mass did for Einstein, or the Moon and the apple for Newton. Each domain is a web of similarities among concrete and abstract objects, and the search carries the relational structure of one across to the other, the move we picture as a *wormhole*, a similarity between similarities. We term this **resemblance-driven** abduction. The other case is **failure-driven**: a known theory meets a contradiction, forcing expansion in the dialectical *thesis-antithesis-synthesis* pattern. The two are not rival mechanisms, and a failure can itself send a model hunting for a resemblance.

Farmer (2026)[3] gives one special case of resemblance-driven abduction a concrete theoretical form. In this *identity abduction*, an LLM re-represents a problem to expose its invariants, retrieves a structurally corresponding domain, proposes an explicit mapping, and tests that mapping adversarially. This closely matches our wormhole path. Farmer reports a motivating episode in which a human supplied the decision to search and the broad strategy of searching for visually analogous images, though not the target correspondence. Farmer's paper specifies DAB-30 but does not report benchmark results; its arXiv record states that execution is reported in a companion paper, which we could not locate. We turn such support into a controlled intervention, *empirically* isolating component capacities of

[1]The divide between abduction and induction is contested. Harman (1965) treats enumerative induction as a special case of inference to the best explanation, and induction can itself yield new generalizations. Like other LLM work, we use Peirce's roles operationally, not as a claim about what induction can never do.

[2]The authors softened this claim in their OpenReview response.

[3]We became aware of Farmer (2026), Balani & Panda (2026), and Song et al. (2025) only after our experiments were underway, and treat the first two as concurrent.

that loop by supplying or withholding a cue, a candidate explanation, or a hypothesis family, rather than evaluating the complete agent.

Balani & Panda (2026) approached failure-driven abduction from a physicist's point of view. Their example is Planck's quantization hypothesis, prompted by the black-body problem that classical physics could not resolve. They argue that, at fixed weights, a wrong answer costs a model no more than a right one, so physical cost supplies no additional pressure to revise. As evidence, they cite Eldin (2026), who report nearly unchanged mean token entropy across task categories despite sharply different accuracies. Yet we argue that cost and entropy do not exhaust the information available in a model, which can still carry, somewhere inside, a decodable signal that its answer is likely wrong.

Failure comes in two kinds. When evidence contradicts the current explanation, as it did for Planck, we ask whether the contradiction triggers revision (Section 5). When a model simply answers wrongly, with no feedback, we ask whether an internal warning exists and can be used (Sections 3 and 6). Error detection matters since a failure cannot begin a search for a new theory until it is noticed. Most of the existing literature examines the model at its output. Kadavath et al. (2022) ask a model to grade its own answers, Salimi et al. (2026) survey the abduction landscape, and Steiglechner et al. (2026) use the Socratic method and find smoke but not fire. A few works open the hood. Burns et al. (2023) ask whether a model represents factual truth inside, and Marks & Tegmark (2023) whether geometric structure encodes it. We look inside for something narrower, a sign that the model's own answer is wrong. In several tested model-task pairs, a probe of the hidden state (Alain & Bengio, 2017; Belinkov, 2022) detects the model's errors better than the output's own uncertainty does. We call the difference the 'surplus', and in a toy model we watch it emerge through a grokking-like phase change (Power et al., 2022; Nanda et al., 2023). Related evidence is reported independently by Orgad et al. (2024); Azaria & Mitchell (2023); Zhang et al. (2025); Yuan et al. (2026); Sarfati et al. (2026). Still, a noticed failure is not a mended one. The remaining stairs ask what lies beyond detection.

Abduction can happen everywhere, in empirical sciences, architecture, even poetry. A realistic, data-oriented line of work asks whether a complete system can carry out such inquiry on real problems. Graph of States (Luo et al., 2026) wraps a frontier model in an external controller that keeps an explicit hypothesis-evidence graph and backtracks when evidence conflicts, with strong gains on real medical and systems diagnosis. Song et al. (2025) evaluate scenario-level questions and complete hypothesis, experiment, and refinement loops across scientific projects. On real cases, however, cramming and reasoning stay entangled. So we work in mathematics and logic, where claims can be checked exactly without a physical experiment; the diagnostic experiments in Section 3 also use trivia with known answers. In the procedural batteries each world is generated from a hidden secret, so its exact answer was never available for the model to memorize. A success therefore cannot be explained by cramming alone, though it still leaves open which reasoning route got the model there. The same design could in principle apply beyond mathematics and logic wherever an oracle can check a hypothesis's consequences, but we do not test transfer to empirical science in this article.

Our staircase begins with routine abduction, formulated as Bayesian updating over a supplied hypothesis set. Here we measure how models weight prior base rates against observed evidence. In radical abduction, one route begins with resemblance. Can a frozen model recognize a structural correspondence, retrieve one from a fixed synthetic library, or construct a mapping when given time, examples or parallel attempts? The other begins with failure. Does the model detect that an explanation has broken, act on that signal, and construct and test an alternative? Across fresh procedural worlds and several model families we combine behavioral tests, internal-state readers, serial computation, and controlled interventions on supplied information. Supplied structure and hidden serial computation change performance sharply, while constructing and verifying an explanation without explicit structural support remains unreliable in the tested settings.

2 THE STAIRCASE

The introduction named the steps. Table 1 turns them into experiments, showing what is given, what must be found, and which support, if any, we hand back to measure what it buys.

What do we mean by resemblance, and what does an explanation owe us? Newton looked at the falling apple and the orbiting Moon, and *recognized* that both might obey one law. He implicitly mapped one domain onto the other, and the map made a prediction: the pull measured on the ground

Step	Given	Must find	Handed back	Sec.
Routine abduction				
Select	candidate causes, base rates, evidence	most probable cause	none	§4
Resemblance route				
Recognize	system, two candidates	which is its twin	hidden serial computation, scratchpad, library, parallel tries	§4
Retrieve	system, library	twin in library	none	§4
Failure route				
Notice	no feedback	own error	none	§3
Act	error signal	answers to drop or retry	method for retry	§3, §6
Construct after failure	program language, evidence breaking first guess	new program fitting all evidence	four-program menu	§5
Construct, no trigger	examples only	forecasts that use the rule, if one exists	search cue, hypothesis family, LLM-authored guide	§5

Table 1: The staircase. In the primary Retrieve, Notice and Act measurements, an external probe, similarity measure or router reads the LLM's internal states instead of asking the LLM.

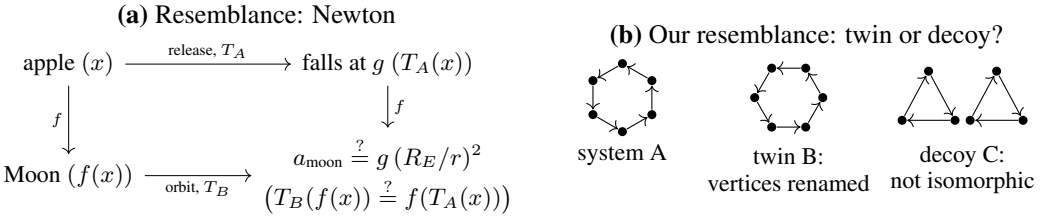

Figure 1: The resemblance route. **(a)** Newton's map: sending the apple to the Moon (f) predicts the Moon's acceleration from the g measured on the ground (R_E: Earth's radius; r: the Moon's distance): two paths around the square must agree. **(b)** Schematic love story: A is a hopeless hexagon of one-way loves, B renames the lovers, and C splits them into two love triangles. Everyone still loves exactly one and is loved by exactly one (edge types omitted).

should account for the Moon's orbit, and it did (Figure 1a). An explanation gains force when it predicts something it was not built from, and it has more teeth the more sharply those predictions separate it from rivals. The resemblance route asks whether a structure and its consequences can travel across domains. The failure route asks whether failed consequences make the LLM replace its explanation. In general, a map f from system A to system B must commute with the actions on each side, $f(T_A(x)) = T_B(f(x))$ (Figure 1a). When A and B sit far apart on the semantic surface, we reserve the name *wormhole* for such a map. Structure-mapping theory scores an analogy by the relations it carries across (Gentner, 1983; Falkenhainer et al., 1989). Farmer (2026) applies it as a test, rejecting a candidate mapping when its predictions fail.

On our resemblance route, the apple does not fall on our head. We replace the apple and Moon with small directed graphs whose structure can be checked exactly, shown only as alien sentences, one per edge ('dapo mirks feku' reads like 'Ann loves Bob'). B, the twin, is A with its vertices renamed. C, the decoy, rewires A but preserves every vertex's degree, so vertex, edge and degree counts do not separate B from C (Figure 1b). Separately, we rewrite B and C ever further from A's prose (new names → new sentence form → new word shapes, see Table D.2 for a love-story version). Every system also lists its sentences in random order, so no line can be matched by position. These rewrites never alter the graphs, so only the relations can tell twin from decoy. As the wording drifts, finding the twin becomes a search for a miniature wormhole (Section 4).

The *failure route* needs no second system, and our miniature (Section 5) follows Planck's case (Planck, 1900; Kuhn, 1978): an old theory, an observation that contradicts it, and a new theory explaining both old and new observations. Consider a hidden circuit with an input dial d_1, which has only a

few settings, and an output display D. Let $G_0 \subset G_1$ be the families of laws in which D depends on d_1 alone, or may also depend on a second dial d_2, respectively. In every observation the LLM is shown, d_2 is determined by d_1, so a G_0 law, the old theory, fits as well as the true G_1 law. Only an intervention that holds d_1 fixed while moving d_2 can discriminate G_0 from G_1. The LLM first writes down its current law, then chooses its own settings, so no one hands it the decisive test. Once G_0 is contradicted, the LLM must find a law in G_1 that fits every old and new observation, and a program checks it against the true circuit on every setting, not only the observed ones. We keep the program language fixed on purpose, so that every candidate can be checked exactly. The task thus mirrors Planck's step of finding a law that covers both old and new observations, not the invention of new concepts, which would need a different test. Newton's structure gained force when its consequences travelled; the structure Planck inherited lost force when its consequences failed.

Planck had a contradiction to start from. The last step removes it and asks whether an LLM notices a rule at all, unprompted. We call a world *deep* when a rule of a tested class generates its entries, and *flat* when a surface-matched world has no such rule. Depth means generative structure, not difficulty. We pair deep and flat worlds and ask the LLM to forecast entries we hold back.

Five design rules keep every experiment honest (Appendix A): paired contrasts, ground truth fixed independently of the model and checked mechanically, a positive control behind every null, predictions and thresholds frozen before each run, and access that ranges from a toy model's training trajectory through open-weight internal states to frontier behavior alone. "At chance" means not demonstrably above chance at the reported sample size.

3 What counts as seeing? Diagnostic experiments

A search for a better explanation starts with noticing. So our first question is narrow. Does a frozen model's hidden state carry a readable signal that the model's answer is incorrect, beyond what its output distribution reveals? Prior work shows that such signals exist (Kadavath et al., 2022; Azaria & Mitchell, 2023; Orgad et al., 2024; Kuhn et al., 2024; Cho et al., 2026; Villuri et al., 2026), and we do not claim them as new. We ask where an interior reader adds anything. Items are short prompts with one-token answers. We compare output entropy with a kernel ridge probe of the hidden state at the answer position, fitted out of fold and reported at its best layer. AUROC measures how well each reader ranks wrong answers above right ones, computed within item types (e.g., arithmetic, retrieval) before averaging. The interior's advantage is the *surplus*. Stronger output comparators and selection checks follow below and in Appendix C.

We first watched the wrong thing (Experiment 1). Our first guess was that an error would leave a trace in the hidden state's trajectory through the layers. On 225 prompts across Pythia up to 6.9B (Biderman et al., 2023) and Qwen-2.5 up to 7B (Yang et al., 2024), *trajectory summaries do not improve* on entropy in base Pythia. However, Qwen-7B shows us light: its hidden states carry additional error information, but the best single layer reveals more than the trajectory summaries.

The interior signal emerges during training (Experiments 2 to 4). To watch the signal form, we train a one-layer transformer on addition modulo 97, using half of all pairs. In each of three seeds it groks (Power et al., 2022; Nanda et al., 2023), jumping from memorization to generalization. Before the transition, entropy already separates right from wrong almost perfectly, and the hidden-state probe adds nothing. A memorizing model wears its errors on its sleeve. Afterward, accuracy occasionally collapses and recovers. These collapses are the only windows with enough errors to score, and in all seven of them the probe beats entropy.

A signal that can be decoded is not yet a signal that can be used. Experiment 3 turns each reader into an external gate that abstains on suspicious answers. Before the jump no checkpoint favors the interior gate. After the jump, at matched coverage, the interior gate wins decisively at 77 of 102 severe-collapse states from three retrained seeds. Milder collapses could still favor entropy. Module swaps in Experiment 4 implicate the input embedding more than the output map, while much of the hidden states' relational geometry survives. This offers a possible explanation for the probe's resilience, not a demonstration that the model notices its own failure (Appendix C).

Pretrained models refuse a simple taxonomy (Experiments 5 and 6). Does the post-grokking toy pattern hold in heavily trained models? Phi (Abdin et al., 2024) and Gemma-3 base up to 27B (Gemma Team, 2024) show no clear surplus on their test batteries, since entropy already flags the errors they make. Qwen-2.5 base at 14B and 32B differ. On the original 225-item battery, their surplus over entropy is about $+0.2$, smaller but still significant after removing arithmetic. Against the top-two probability margin, however, the surplus shrinks to a few points and is no longer significant. The striking difference is within tasks. On Qwen-14B arithmetic, entropy is lower on the errors than on correct answers (AUROC 0.22, below chance), while the probe separates them well (0.79). Across 19 base models, best-layer AUROC varies much less than entropy AUROC (Appendix C). Where the interior adds useful error information therefore depends on the model and the task, rather than on model family alone.

Balani & Panda (2026) argue that wrong answers incur no extra physical cost at fixed weights. We do not test that claim. But the inference drawn by Eldin (2026) from flat category means, that entropy cannot distinguish grounded answers from confabulations, is contradicted by our item-level results. Across 19 base models, within-type entropy AUROC ranges from 0.58 to 0.91, and in some model-task pairs an interior reader adds error information. Whether that information can steer action is the question of Section 6.

4 Choosing and recognizing

Noticing an error is a reason to look again, not an explanation. We next ask whether a model can choose among supplied causes, recognize a structural twin among supplied candidates, and finally find that twin in a fixed library.

Routine abduction: good until the clues disagree (Experiment 7). A doctor facing a fever weighs each candidate disease by how prevalent it is (the prior) and how well it explains the symptoms (the likelihood). Bayes' rule multiplies the two to update belief (posterior) about possible disease. Following Grether's model of human judgments (Grether, 1980), we fit a model of the LLM's answer probabilities with separate weights for the prior and likelihood. Up to answer-letter biases (Appendix D),

$$P_{\mathrm{LLM}}(h \mid e) \;\propto\; P(e \mid h)^{\alpha}\, P(h)^{\beta}, \tag{1}$$

where h is a cause and e the observation. With zero bias, Bayes' rule is $\alpha = \beta = 1$. The ratio $r = \beta/\alpha$ sets the relative weight of prior to likelihood. At fixed r and zero bias, increasing $\alpha > 0$ sharpens the distribution. We test five base models (Qwen-2.5 1.5B to 32B and Gemma-27B) on two independently generated sets (draws) of 405 vignettes, each stating the base rates and likelihoods of three causes. In *one-cue items*, we hold either the priors or the likelihoods equal, so the other cue alone determines the answer. In *conflict items*, the Bayesian winner is neither cue's favorite. Following either cue alone gives the wrong answer, so the model must combine them. Every model underweights the prior in both sets ($\hat{r}$ from 0.37 to 0.66; paired item bootstrap, Appendix D). Scaling Qwen increases both weights while their ratio remains below one. One-cue accuracy approaches ceiling, but conflict accuracy peaks at 0.52 against chance of one third. From 7B upward, at least three quarters of conflict errors choose the rarer, likelihood-favored cause.

Conflict is also where confidence becomes misleading. From 7B upward, entropy is *at or below chance* for detecting these errors in three of four models on the first set and all four on the second. At 14B, the interior probe exceeds entropy by $+0.25$ on the second set (95% interval excludes zero). Gemma-27B reverses its Section 3 pattern. Entropy exposes its retrieval errors but misses these cue-integration errors, whereas its whole-battery surplus replicates across both sets. Which channel reveals the error therefore depends on the task, not the model alone. On wrongly answered conflict items, a probe also decodes the correct cause above chance at 14B and 32B in both sets (32B: 0.74 and 0.70; chance one third). Yet it may expose the cue ordering rather than a corrected posterior, since its advantage over the output's own ranking did not replicate.

Recognizing a twin (Experiments 8, 8c and 8f). Choosing a structural twin requires a different comparison. The alien graphs of Section 2 supply no likelihoods. The model must recognize which candidate preserves the relations. We test the same five base models on two sets of 280 choices at three surface distances: relabel (new names), skin (also a new sentence form) and far (also new

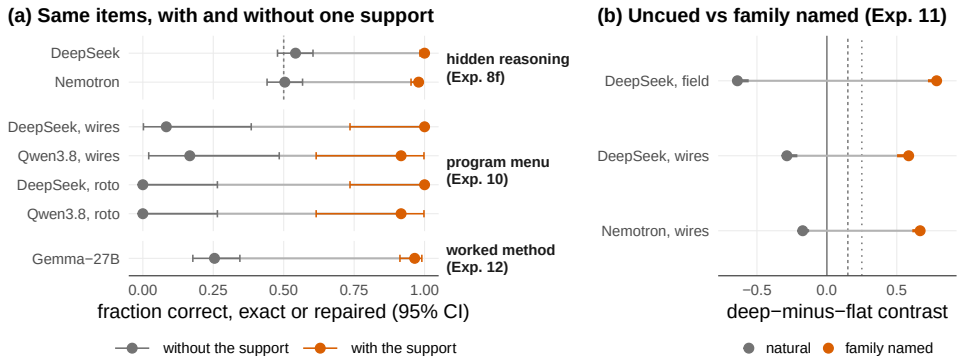

Figure 2: Effects of supplied support on matched items. **(a)** Accuracy without (grey) and with (orange) hidden reasoning (Experiment 8f), a four-program menu (Experiment 10), or a worked retry method (Experiment 12); the final grey control provides only room to think. Intervals are marginal 95% bounds; the dashed line marks chance for Experiment 8f. **(b)** Deep-truth probability minus maximum flat-twin probability under natural prompting or a named hypothesis family (Experiment 11). Whiskers are registered one-sided 95% bounds; horizontal lines mark the registered criteria. Exact values are in Appendices D to F.

word shapes). Easy controls have decoys with two edges missing. No model passing that control establishes above-chance structural recognition in one forward pass. Qwen-7B fails the control and remains descriptive. An interior probe recovers the easy count cue but meets no registered criterion for structural recognition either. Detecting an error does not by itself supply a correspondence.

Maybe one pass with no room to work is the wrong test. Experiment 8c gives Qwen-2.5-14B-Instruct deliberation, a scratchpad, worked examples or parallel tries. Forced-choice accuracy stays at chance in every arm, despite above-chance reading of local features and nine checker-certified complete mappings, all onto the twin. These partial successes leave open whether the supplied help was sufficient.

With more serial computation, the frontier models cross this boundary. Experiment 8f compares DeepSeek-V4-Pro and Nemotron-3-Ultra on the same fresh certified items, with reasoning disabled or a hidden reasoning budget before the same answer line. Accuracy on 240 structural items *rises from near chance to 1.000 and 0.979*, respectively (Figure 2a). Both pass the easy control even with reasoning disabled. The recognition wall is a one-pass wall at every scale we tested, and hidden serial computation crosses it. The prompt names the correspondence to seek, so the result concerns instructed recognition, not spontaneous discovery. Each family contributes one pinned configuration. Unusable routes are documented in Appendix D. Related one-pass failures are reported by Thushalika et al. (2026); Abdaljalil et al. (2026).

Retrieving a twin (Experiment 9). Recognition puts the candidates in front of the model. Analogical retrieval must find a candidate first (Forbus et al., 1995). We embed each system separately and ask whether hidden-state similarity ranks its twin above the decoy, then whether it retrieves the twin from a library of 480 systems. Neither similarity nor learned metrics establishes reliable retrieval. An initial Gemma positive did not replicate. Later relational readouts also had uneven control sensitivity (Appendix D). This is a limit of the rulers, not evidence of absence. The serial-computation result is compatible with constructing the correspondence on the fly, but does not establish that mechanism.

5 When the explanation itself is missing

Section 4 handed the model its candidate explanations. Here the candidate set is withheld. We test construction after a contradiction (Planck's case from Section 2), prediction without one, and guidance written by another instance of the same model.

Construction after a certified failure (Experiment 10). The circuit in Section 2 is the *wires* family, where one display secretly depends on a dial outside its declared panel. In a second family, *roto*, actions transform a five-state system, and their order can matter even though the passive records

admit an order-blind explanation. DeepSeek-V4-Pro and Qwen3.8 write executable explanations, choose interventions and revise on the returned results in 40 escape worlds per family. Forty adequate worlds per family check for needless rewriting. An oracle arm establishes solvability.

The registered chain requires an unsupplied amendment, a test separating live candidates, revision on its outcome, and a final program correct on old and fresh settings. *Neither model completes that chain in either family* (0 of 40 in each cell), although each produces truth-equivalent programs in 4 of 40 wires worlds and neither rewrites an unbroken mechanism. The criterion requires at least two live candidates at confirmation. Committing to one earlier fails it, even with a correct final program.

With a four-program menu, exact final programs *rise from at most 2 of 12 to at least 11 of 12* in both families and both models (Figure 2a). In wires, a replay shows the true program was the unique surviving match after the first batch of interventions in 22 of the 23 exact episodes. The menu changes information, representation and search together, so the experiment does not say which component produced the gap.

Does the contradiction help the model propose an alternative, or does another call suffice? A paired test branches from the same initial commitment into two matched continuations, with and without the contradiction, each with the model's own intervention. In wires, neither model emits the target dependency class beforehand. After the contradiction it appears in 12 of 26 and 11 of 33 exposed worlds (denominators differ because models commit differently), *never* in the matched control. The result does not extend to roto. The anomaly, not the extra call, makes the right class reachable, though what the class does remains uncertified. The trigger is only half supplied. The result is shown unlabelled, but the protocol asks for revised candidates and another test. Construction remains within a fixed program language. Appendix E gives the full chain audit and contrasts with other contradiction tests (Steiglechner et al., 2026; Shi et al., 2026).

Detecting structure with no cue (Experiment 11). Planck had a contradiction. Remove it. The model sees observations and must forecast sealed entries. In a deep world, a compact rule determines them, while its surface-matched flat twin violates that rule class and leaves the sealed outcomes undetermined. We score probability on the deep truth minus the largest probability assigned to any flat outcome. Identical forecasts on both twins cannot score above zero: success needs prediction where a rule exists and restraint where it does not.

We compare natural prompts, a placebo, a search cue, decoded observations (representation supplied, law withheld), and scaffolded observations (decoded plus the hypothesis family). DeepSeek-V4-Pro is tested on 224 fresh worlds from two families, an alien finite-field system and passive wires. Natural and scaffolded arms cover all 224 worlds, and the other arms use fixed subsets. The natural contrast is *negative in both*, with upper bounds below zero (Figure 2b). The model often concentrates on one outcome even in the flat worlds. Neither the search cue nor its placebo reliably rescues discrimination. Decoding alone helps partly in one family. Adding the family description changes the result. The model identifies *the deep truth in 111 of 112 valid packets* and spreads its forecasts on the flat twins. Nemotron-3-Ultra replicates the wires result, but its finite-field pilot produced too few valid answers within the token limit, so the two-family result rests on DeepSeek alone.

Can the model write useful guidance itself? In Experiment 11b, three DeepSeek instances studied separate solved examples and wrote guides, frozen before fresh recipients used them on six held-out wires pairs. The average gain over matched generic advice came from less concentration on flat outcomes, without better prediction on deep worlds. The human family description remained stronger. Reliable transfer from the model-written guides was not established. Appendix E reports the uncertainty, delivery failures and a one-guide Nemotron replication.

6 An alarm is not an address

Section 3 showed that an outside reader can use the error signal to abstain. We now supply a remedy and ask whether reading the interior allocates retries better than reading the output. The routing rule and remedy are supplied externally. The model chooses neither.

Routing a second pass (Experiment 12). We return to the routine-abduction battery with Gemma-27B and first retry all 405 items from the second set. A second pass with a worked prior-times-

likelihood example repairs 110 of 114 errors and breaks none of the 291 correct answers. Simply offering room to think repairs 29 errors (Figure 2a). That control usually re-emits an answer after only a few tokens, so it compares offered help, not equal amounts of computation.

Once the remedy works, the question is where to spend it. Routers trained on the first item set rank the second set without its labels. Using the captured second-pass answers, we evaluate policies that retry only the most suspicious quarter (101 items). Because repair is nearly universal, the useful router is the one that finds more errors within this budget. The interior beats generic uncertainty scores, but its advantage over a trained output reader remains inconclusive. Final error is 2.6 percentage points lower, with a paired 95% interval from 0 to 5.2 points. The second-pass intervention was predeclared after screening these first-pass results, rather than an independent confirmation (Appendix F).

Does the interior matter more when evidence is harder to recover? (Experiment 13). Experiment 13 states the same three-cause probabilities directly, through one or two linked lookup tables, or in a token-matched padded direct control. The correct answer never changes. One lookup reduces likelihood-following while accuracy changes little.

The registered test compares the interior's added routing benefit between the one-codebook and padded versions, against bounded output readers fitted without oracle-solved input features. The difference is -0.030 (95% interval: -0.063 to $+0.005$), negative on both evaluation sets and significantly negative against the MLP comparator. A post-hoc combination with the strongest output reader detects no additional routing gain on any version. The verdict remains unresolved, leaning against the predicted increase with harder evidence recovery. This tests one model in one-pass inference and changes access to evidence, not decidability (Appendix F).

Like Graph of States (Luo et al., 2026), we route externally, choosing which answers get a known remedy. Whether internal information can choose the remedy an error needs remains untested.

7 Discussion

Finding an explanation is not one act but several, and the steps respond to different kinds of help. Worked methods repair selection errors, hidden computation rescues prompted recognition, and supplied candidates or hypothesis families transform identification and forecasting. Success with that support does not establish that models can find it for themselves.

If Dennett is right, these steps should extend to empirical science. An inquiry must still propose an explanation, derive consequences and test them, and failure-driven inquiry must also notice the failure that begins the search. *The framework is not tied to text. Wherever a reliable simulator can check consequences, genomic or chemical models could face analogous tests.*

Whether science advances through innumerable routine abductions, or small horizon-expanding ones, or a few horizon-defying abductions is a debate we do not settle. These may not be the only stairs of abduction, but they are sturdy ones. Discoveries from Newton's to Planck's have needed each of them, and each is testable. The staircase already has suggestive evidence: a non-mathematician supplied the goal and only encouraging prompts, and a multi-agent system found a new mathematical route with long serial computation (Claude, 2026). Lamzouri (2026) subsequently simplified the proof and derived consequences the original had not stated.

We do not follow one model through every step, so the staircase shows what helps at each step, not which is hardest. We have no human comparison. Models construct within a supplied programming language and are asked to revise after failure; they do not start the inquiry themselves. We test whether interior signals help allocate a remedy, not choose one. Experiment 13 uses one model, and every task has a mechanically checkable answer.

Future LLMs may conjure hypotheses, but what makes *them* dream will be shaped by training data, corporate incentives, reinforcement mechanisms and over-optimization. Humans can just dream for the sake of dreaming, generate hypotheses, and wander in thought spaces for fun. So fellow humans, don't stop dreaming.

AI USE STATEMENT

This work is itself about language models and those were used significantly in this work. But the families that served as tools were never subjects. Under my supervision, Claude (Anthropic), Codex and ChatGPT (OpenAI) helped design the experiments, wrote and audited the generators, capture and analysis scripts, searched the literature, and edited this text. Claude and Codex conducted multiple rounds of adversarial review, which made the work much more rigorous.

The synthetic worlds were produced by those scripts from secrets that no model saw. The models under test (Pythia, Phi, Gemma-3, Qwen-2.5, Qwen3.8, DeepSeek-V4-Pro, Nemotron-3-Ultra, and our own toy transformers) took no part in task design. Two open-weight models outside the tested set, gpt-oss-120b and MiniMax-M3, served only as leakage auditors for Experiment 10, on worlds drawn from a separate audit seed rather than the scientific battery; MiniMax-M3 was also piloted as a subject in Experiment 8f and excluded as an instrument before any scientific run. Every prediction, threshold and kill criterion was frozen in a script header before its run; every probe analysis reproduces its frozen comparator exactly before a new number is reported; and the frontier and routing results (Experiments 8f and 10 to 13) were re-derived from the archived outputs by a second, independently prompted model.

However, the genesis of ideas, questions and final claims are mine, and I solely take full responsibility.

ETHICS STATEMENT

I declare that I have no affiliation with, and received no funding, compute, or other support from, any company that develops the language models studied or used in this work. Model access was obtained through standard commercial or free public services.

REPRODUCIBILITY STATEMENT

Predictions, thresholds and kill criteria for every experiment were frozen in the header of its script before the run, and later changes are recorded as dated amendments rather than silent edits. Every call to a hosted model pins the exact model, provider, quantization, date and reasoning setting, verified per call. Hidden worlds were generated from a secret that appears in no prompt and no public artifact, and the generators will be made public if this article is accepted. The appendices state every registered rule and gate. Code, configurations, prompts, item files and the complete call archives will be released with the camera-ready version.

ACKNOWLEDGMENTS

Acknowledgements withheld for double-blind review and will be visible if accepted.

REFERENCES

Samir Abdaljalil, Erchin Serpedin, and Hasan Kurban. IsoSci: A benchmark of isomorphic cross-domain science problems for evaluating reasoning versus knowledge retrieval in LLMs. *arXiv preprint arXiv:2607.01431*, July 2026.

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2017.

Amos Azaria and Tom Mitchell. The internal state of an LLM knows when it's lying. *arXiv preprint arXiv:2304.13734*, April 2023. Findings of EMNLP 2023.

Paras Balani and Subhrakanta Panda. LLMs don't pay for the jump: Thermodynamic decoupling in language models. *arXiv preprint arXiv:2608.14397*, August 2026.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: A suite for analyzing large language models across training and scaling. *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. *Proceedings of the 11th International Conference on Learning Representations (ICLR)*, 2023.

David J. Chalmers. Facing up to the problem of consciousness. *Journal of Consciousness Studies*, 2 (3):200–219, 1995.

David J. Chalmers. *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press, 1996.

Zirui Cheng, Zeyu Shen, Thomas L. Griffiths, and Peter Henderson. Using cognitive models to improve language model simulation of human persuasion games. *arXiv preprint arXiv:2606.17657*, June 2026. Princeton; applies Grether alpha-beta to LLMs by prompting.

Seonglae Cho, Zekun Wu, Kleyton Da Costa, and Adriano Koshiyama. The confidence manifold: Geometric structure of correctness representations in language models. *arXiv preprint arXiv:2602.08159*, February 2026.

Claude. More than two thirds of the zeros of the Riemann zeta function lie on the critical line. Technical report, Anthropic, August 2026. Technical report; not peer-reviewed.

Daniel C. Dennett. *Consciousness Explained*. Little, Brown and Company, 1991.

Sajib Acharjee Dip, Dawei Zhou, and Liqing Zhang. Think again or think longer? Selective verification for budget-aware reasoning. *arXiv preprint arXiv:2606.19808*, June 2026.

A. Gamal Eldin. Descriptive versus regulatory uncertainty in bounded predictive systems. *arXiv preprint arXiv:2605.18909*, 2026.

Brian Falkenhainer, Kenneth D. Forbus, and Dedre Gentner. The structure-mapping engine: Algorithm and examples. *Artificial Intelligence*, 41(1):1–63, 1989.

Michael W. Farmer. Abduction without a body? representational grounding and the abduction loop for scientific hypothesis generation. *arXiv preprint arXiv:2608.02505*, August 2026. URL `https://arxiv.org/abs/2608.02505`. v1.

Kenneth D. Forbus, Dedre Gentner, and Keith Law. MAC/FAC: A model of similarity-based retrieval. *Cognitive Science*, 19(2):141–205, 1995.

Gemma Team. Gemma: Open models based on Gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

Dedre Gentner. Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7(2): 155–170, 1983.

David M. Grether. Bayes rule as a descriptive model: The representativeness heuristic. *The Quarterly Journal of Economics*, 95(3):537–557, 1980.

Gilbert H. Harman. The inference to the best explanation. *The Philosophical Review*, 74(1):88–95, 1965. doi: 10.2307/2183532.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tyre, Zhengdong Wang, Tom Brown, Sam McCandlish, Jared Kaplan, and Chris Olah. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Samuel H. King, Claudia L. Driscoll, David B. Li, Daniel Guo, Aditi T. Merchant, Garyk Brixi, Max E. Wilkinson, and Brian L. Hie. Generative design of bacteriophages with genome language models. *Science*, 393(6811):eaec2657, 2026. doi: 10.1126/science.aec2657.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *Nature*, 625:659–664, 2024. arXiv preprint arXiv:2302.09664 (2023); published Nature 2024.

Thomas S. Kuhn. *Black-Body Theory and the Quantum Discontinuity, 1894–1912*. Clarendon Press, Oxford University Press, 1978.

Youness Lamzouri. A new proof that more than $2/3$ of the zeros of the Riemann zeta function are simple and on the critical line. *arXiv preprint arXiv:2609.02882v2*, 2026. Version 2, 8 September 2026.

Yu Luo, Rongchen Gao, Lu Teng, Xidao Wen, Jiamin Jiang, Qingliang Zhang, Yongqian Sun, Shenglin Zhang, Jiasong Feng, Tong Liu, Wenjie Zhang, and Dan Pei. Graph of states: Solving abductive tasks with large language models. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2603.21250. arXiv:2603.21250; accepted at ICML 2026.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *arXiv preprint arXiv:2310.06824*, 2023.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. In *Proceedings of the 11th International Conference on Learning Representations (ICLR)*, 2023.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. LLMs know more than they show: On the intrinsic representation of LLM hallucinations. *arXiv preprint arXiv:2410.02707*, October 2024. Reported ICLR 2025; venue not confirmed firsthand.

Charles Sanders Peirce. Deduction, induction, and hypothesis. *Popular Science Monthly*, 13:470–482, 1878.

Charles Sanders Peirce. *Collected Papers of Charles Sanders Peirce, Volume V: Pragmatism and Pragmaticism*. Harvard University Press, 1934. Edited by Charles Hartshorne and Paul Weiss.

Max Planck. Zur theorie des gesetzes der energieverteilung im normalspectrum. *Verhandlungen der Deutschen Physikalischen Gesellschaft*, 2:237–245, 1900.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets. In *First Mathematical Reasoning in General Artificial Intelligence Workshop, ICLR 2021*, 2022. arXiv:2201.02177.

Moein Salimi, Shaygan Adim, Danial Parnian, Nima Alighardashi, Mahdi Jafari Siavoshani, and Mohammad Hossein Rohban. Wiring the 'why': A unified taxonomy and survey of abductive reasoning in LLMs. *arXiv preprint arXiv:2604.08016*, April 2026.

Raphaël Sarfati, Pratyush Ranjan Tiwari, Siddharth Boppana, Christopher J. Earls, Srikar Varadaraj, and Eric Ho. What LLM forecasters know but don't say: Probing internal representations for calibration and faithfulness. *arXiv preprint arXiv:2607.08046*, July 2026.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? *arXiv preprint arXiv:2304.15004*, April 2023.

Dai Shi, Xiaoyu Li, and José Miguel Hernández-Lobato. When the canonical completion is wrong: Formalizing and measuring the jump in large language models. *arXiv preprint arXiv:2608.26187*, August 2026.

Zhangde Song, Jieyu Lu, Yuanqi Du, et al. Evaluating large language models in scientific discovery. *arXiv preprint arXiv:2512.15567*, 2025. URL https://arxiv.org/abs/2512.15567. v2, revised May 2026.

Julius Steiglechner, Lucas Mahler, and Gabriele Lohmann. LLMs can see the smoke but not the fire: Evaluating abductive reasoning with Elenchos. *arXiv preprint arXiv:2607.12733*, July 2026.

Kumar Thushalika, Sukumar Kishanthan, and Asela Hevapathige. Detecting differences is not understanding structure: Large language models fail at graph isomorphism. *arXiv preprint arXiv:2606.09484*, June 2026.

Gnaneswar Villuri, Hashmath Shaik, and Alex Doboli. When do internal probes beat reading the answer? Miscalibrated readouts and behavior-concealed knowledge in language models. *arXiv preprint arXiv:2609.04582*, September 2026.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.

Aojie Yuan, Zhiyuan Julian Su, Haiyue Zhang, Yi Nian, and Yue Zhao. Hidden error awareness in chain-of-thought reasoning: The signal is diagnostic, not causal. *arXiv preprint arXiv:2605.09502*, May 2026. Mechanistic Interpretability Workshop, ICML 2026.

Tom Zahavy. LLMs can't jump. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026. Position paper.

Anqi Zhang, Yulin Chen, Jane Pan, Chen Zhao, Aurojit Panda, Jinyang Li, and He He. Reasoning models know when they're right: Probing hidden states for self-verification. *arXiv preprint arXiv:2504.05419*, April 2025.

A DESIGN RULES, EXPANDED

Section 2 names five rules that hold across every experiment. Here we spell out the detail that the main text compresses.

Matched pair. Every contrast is paired in one of three ways. First, two look-alike items differ only in what the step tests. On the resemblance route, twin and decoy have the same numbers of vertices and edges and the same degree sequence for each edge type; exhaustive checking certifies that the decoy is not isomorphic to the source. On the failure route, rule-bearing (deep) worlds and rule-free (flat) worlds are surface-matched, and a pre-declared gate checks that a bounded surface reader cannot tell them apart. Second, one item is tested with and without a handed-back support: hidden reasoning, a scratchpad, a library, a menu, or the hypothesis family. The prompt and world remain fixed, so the contrast measures what the support buys. Third, the same answer is scored by an internal-state reader and by the output distribution. In the second and third designs, pairing removes item effects.

Hidden truth, exact check. In the procedural experiments, every world is generated from a random seed the LLM never sees, so the ground truth exists before the LLM touches the task. A deterministic program, not a language-model judge, checks every answer. The error-signal experiments (§3) use short arithmetic and trivia questions with known answers. Because some arithmetic labels are noisy, every error-signal result is also reported without them. Elsewhere the check is structural: a twin is certified by trying every vertex correspondence (§4), and a candidate program is run against every record and against fresh interventions (§5).

When the check is noisy, as it is whenever empirical evidence replaces a lookup, the firewall guarantee (Appendix B) bounds false acceptance: an outside propose-and-test loop cannot inflate it if every claim meets fresh evidence at significance levels fixed in advance. The proposer may be adversarial; the guarantee depends only on the checker seeing unused data and on a pre-committed spending schedule.

Positive control. Every battery carries a check that the instrument fired: easy items the LLM should get right (an easy-accuracy gate in §3; items whose decoy has fewer edges in §4), an oracle or scaffolded arm that shows the world is solvable (§5), and a repair oracle and an interior power probe (§6). We read a null result only where that check passes; an LLM that fails it is treated as an instrument that did not fire, not as evidence of absence.

Access. Three tiers of access appear across the experiments. A one-layer transformer trained from scratch on addition modulo 97 (§3) gives full access: we watch a signal form during training, checkpoint by checkpoint. Open-weight LLMs up to 32B parameters (mainly the Qwen-2.5 and Gemma-3 families) expose internal states at every layer, which we analyze with probes, similarity measures, and routers fitted offline. Frontier LLMs (DeepSeek-V4-Pro, Qwen3.8-A95B, Nemotron-3-Ultra), at pinned versions and providers, expose behavior only: output tokens and, where the API permits, reasoning traces. The access tier is a scope statement, not a hierarchy; each experiment uses the access needed to answer its question.

Frozen reading. Predictions and thresholds were fixed in each script's header before it ran, and later changes are dated in the header with a written account. This is the preregistration substitute for procedurally generated tasks, where a public registry entry would leak the task structure. A wrong answer is the positive class throughout, so an AUROC above 0.5 means the signal detects errors. Error signals are scored within item type, because pooling item types rewards a reader that merely tells hard types from easy ones without detecting errors within either. Behavioral measures use accuracy or a paired difference with a 95% confidence interval. The phrase "at chance" means not demonstrably above chance at the reported sample size, never that the true effect is exactly zero.

B THE FIREWALL GUARANTEE

We return to the division Peirce identified: abduction proposes, deduction validates. Section 3 finds a decodable error signal in some model-task pairs. Such a signal could trigger an external search, but

its validator must remain outside the proposal loop: the proposer cannot also serve as judge. The following result gives conditions under which the proposer's reliability, or lack of it, cannot inflate false acceptance.

The setup is sequential. An untrusted proposer submits claims one at a time, each depending arbitrarily on the full history of past proposals, tests, and failures. For each claim, a separate checker tests it on fresh evidence not used to construct the claim, producing a conditionally valid p-value (valid conditional on the full history available before the test draw) and spending a pre-committed level α_t with $\sum_t \alpha_t \leq \alpha$.

Result 1 (Firewall guarantee)**.** *Under these conditions, for any proposer policy, including adversarial ones,*

$$P(\textit{at least one false acceptance}) \leq \alpha.$$

The proof is a union bound: each term is bounded by α_t via conditional validity, and the sum is bounded by α via the spending constraint. No independence among proposal errors and no calibration model for the proposer is assumed. A bad proposer wastes budget and misses discoveries; it does not inflate false acceptance.

However, the firewall guarantee does not protect against a mistranslated scientific claim, a wrong simulator, leakage of acceptance data into proposal construction, or scope drift beyond the tested intervention distribution. These specification threats, not statistical bookkeeping, are arguably the harder problem in LLM-assisted science and define the conformance-testing targets for any implementation.

The surplus (Section 3) is a candidate trigger for an external loop, and the commutative-diagram criterion (Section 2) supplies an evaluation standard. The firewall guarantee controls false acceptance when the loop tests each proposal on fresh evidence under a pre-committed spending rule. It does not show that the signal identifies what to search for.

C Extended results for Section 3: error readers and grokking

Before a model can search for a better explanation, it has to register that the current one has failed. So we begin with the first step of the failure route: does a frozen model carry, somewhere inside, a readable signal that its answer is wrong, beyond what its output distribution already exposes? That such signals exist is by now established (Kadavath et al., 2022; Azaria & Mitchell, 2023; Orgad et al., 2024; Kuhn et al., 2024; Cho et al., 2026; Villuri et al., 2026), and we treat it as a precondition, not a contribution. What the detection literature leaves open, and what this section measures, is where in the computation the signal lives, when it forms during training, whether an outside reader can act on it, and how much of it the output channel already exposes across model families once the comparator is the strongest output reading rather than entropy alone.

Throughout this section an item is a short prompt with a one-token answer, correctness is decided by string match, and a wrong answer is the positive class. The output reader is the entropy of the next-token distribution (later also the top-2 probability margin and a probe trained on the full log-probability vector). The interior reader is a kernel ridge probe on the hidden state at the answer position, fitted out of fold, one probe per layer, with the best layer reported. Both are scored by within-type AUROC (Section 2, Frozen reading), and we call the interior's excess over the output reader the *surplus*. Arithmetic labels are noisy, so every number is also computed without them.

We first watched the wrong thing (Experiment 1). Our first guess was that the signal would live in the shape of the walk: how the hidden state moves from layer to layer (step lengths, turning angles, alignment with the final state). On 225 one-token prompts of six types (easy completions; common, rare and obscure facts; arithmetic; short reasoning), across a Pythia ladder up to 6.9B (Biderman et al., 2023) and a Qwen-2.5 ladder up to 7B (Yang et al., 2024), it does not. In base Pythia the ordered trajectory never beats output entropy, and neither does an order-free pooling of all 33 layers at once (6.9B: trajectory 0.826 vs entropy 0.865, difference -0.039 [-0.106, $+0.044$]). Within type, none of our interior readers improves on the output in that family. A reader of Qwen-7B's hidden state does recover error information beyond entropy (pooled layers 0.791 vs entropy 0.652, $+0.139$, $p = 0.0015$), which makes the Pythia null interpretable. But that recovery comes from the full hidden-state geometry rather than step lengths and angles: the trajectory features sit 0.11 below the

Exp.	Model	Cell and contrast	Measure	Result [95% CI]
2	toy, before grokking	probe vs entropy	AUROC	entropy 0.90 to 0.99; probe at chance
2	toy, collapses after grokking	probe vs entropy, seven collapses	AUROC	probe wins 7 of 7 (e.g. 0.947 vs 0.716)
3	toy, three retrained seeds	abstention gate on the interior vs on entropy, 102 severe-collapse states	decisive wins	interior 77 of 102; before grokking 0 of 32
6	Gemma-3-27B	probe vs entropy, whole battery	surplus	-0.016 [-0.071, $+0.024$]
6	Qwen-2.5-14B, 32B	probe vs entropy, whole battery	surplus	$+0.220$ [$+0.135$, $+0.309$]; $+0.200$ [$+0.101$, $+0.315$]; vs top-2 margin $+0.053$, $+0.040$ (n.s.)
6	Qwen-2.5-14B	arithmetic items only	AUROC	entropy 0.222; probe 0.787
7	Qwen-2.5 7B to 32B, Gemma-3-27B	conflict items, replication draw	AUROC	entropy 0.41 to 0.48 in all four models; probe 0.44 to 0.73; gap at 14B $+0.250$ [$+0.119$, $+0.385$]
7	Gemma-3-27B	whole battery, two draws	surplus	$+0.145$, $+0.111$ (both CIs exclude 0)
7	Qwen-2.5-14B, 32B	correct cause decoded from the hidden state on wrong conflict items	accuracy (chance $1/3$)	0.58 [0.44, 0.72], 0.74 [0.61, 0.86]; second draw 0.49, 0.70
12	Gemma-3-27B	101 routed retries: learned output router vs interior router	final error	0.136 vs 0.109; difference $+0.026$ [$+0.000$, $+0.052$]
13	Gemma-3-27B	probe over bounded output readers, one-join minus padded presentation	routing gain	-0.030 [-0.063, $+0.005$]

Table C.1: Interior reader against output reader: the comparisons the main text turns on (moved here from Section 3 for space). AUROC is within type with a wrong answer as the positive class, so values above 0.5 detect errors; a surplus is the best interior layer minus the output reader; n.s. marks an interval covering zero. The full model ladders, Phi, Experiment 1 and the recomputations without arithmetic items follow in the text of this appendix.

best single layer. Two further facts recur later. Without arithmetic prompts, the pooled Qwen surplus vanishes while a single-layer surplus of about $+0.07$ survives, so the surplus concentrates on one prompt type. Instruction tuning also moves the signal toward the output at 7B: relative to the base model, the best-layer interior AUROC falls by 0.066 [0.015, 0.122], while entropy improves at every size. One sentence for the rest of the paper: the signal lives in the shape of the cloud, not the shape of the walk.

The cloud learns to see (Experiments 2 to 4). Real models are trained once. To watch the signal form, we trained our own. A one-layer transformer (width 128, about 250k parameters) learns addition modulo 97 from half of all pairs under full-batch AdamW with weight decay 1.0, and groks (Power et al., 2022; Nanda et al., 2023): it memorizes first, then between steps 2200 and 2900 (three seeds) jumps to full generalization within a few hundred steps. Every 100 steps we freeze the model and read a 256-item probe set both ways. Before the jump, the registered reader detects no interior signal beyond the output: entropy separates right from wrong at AUROC 0.90 to 0.99 while the hidden-state probe sits at chance (0.42 to 0.61). To a linear reader, the memorizing interior is blank about correctness. After the jump, the relation inverts. Under this optimizer the grokked model suffers transient one-step collapses (accuracy falling to 0.2 to 0.5 and recovering within a step), the only post-transition windows with enough wrong answers to score. In all seven windows across three seeds, the probe beats entropy (for example 0.947 vs 0.716 at seed 103, step 3700). In a control collapse before grokking, the probe loses (0.681 vs 0.717). The flip is tied to the learned solution rather than collapse alone. Discontinuous metrics such as accuracy can make gradual change look abrupt (Schaeffer et al., 2023); here a task-blind statistic, the hidden cloud's participation ratio, begins to fall 300 to 1700 steps before the accuracy jump. The trajectory statistics we had bet on never lead.

Decodable is not yet usable. Experiment 3 turns each reader into a gate that abstains on the 64 most suspicious of the 256 items. We score both the risk-coverage curve and the errors caught at that budget, using a stratified paired bootstrap. Before the jump, the entropy gate wins at all 32 checkpoints, significantly at 31: reading the interior there would cost errors. After the jump, a rerun of all three seeds saves every collapse step (348 checkpoints). The interior gate is decisively better at 77 of 102 severe-collapse states and catches 704 additional errors when summed across those 77 states. It is perfect (64 of 64) at 30 states where entropy never is. At the deepest state (seed 103, step 3700), it closes 81 percent of the gap between entropy and an oracle. The advantage grows with collapse severity (correlation -0.61 between the number of wrong answers and the gate difference over 235 post-transition states): mild perturbations leave the output honest, whereas severe collapses scramble it into confident wrongness while the interior remains readable.

Experiment 4 asks what collapses. Swapping modules between a collapsed checkpoint and its healthy neighbors places the damage on the input side. Restoring the healthy embedding recovers 63 percent of the lost accuracy on average; attention comes second; restoring the unembedding recovers nothing in any of 48 episodes. The output map itself moves by only 0.3 percent and is causally inert. The hidden cloud moves by more than its own norm but retains its dimensionality, much of its relational geometry (linear CKA 0.87 to the pre-collapse cloud; Kornblith et al., 2019), and all of its correctness structure. One account is that the probe reads relational structure that survives the collapse, while the fixed output map reads absolute position; the displaced cloud therefore lands confidently in the wrong readout regions. The interior does not stay put. Its relational geometry survives.

Real models refuse a simple taxonomy (Experiments 5 and 6). Does the post-grokking regime describe heavily trained models? We run the 225-item battery, plus 300 trivia questions where errors are sparse, through Phi-1.5 and Phi-2 (Abdin et al., 2024), Gemma-3 base at 1B, 4B, 12B and 27B (Gemma Team, 2024), and Qwen-2.5 base at 14B and 32B, freezing predictions before each capture. Phi shows no surplus: entropy carries at least as much as the best layer (-0.016 and -0.018, both intervals crossing zero), and flags its arithmetic failures (Phi-2 arithmetic: probe 1.000 vs entropy 0.990). Gemma is negative at 1B and 4B, hints at a surplus at 12B ($+0.036$, not significant), and fails the pre-registered replication at 27B (-0.016 [-0.071, $+0.024$]). Gemma makes almost no arithmetic errors (1, 0, 0 and 1 of 30 across the four sizes); on its retrieval-dominated error mix, the output exposes the signal. Qwen differs. Its surplus over entropy on the 225-item battery is $+0.180$ at 7B, $+0.220$ [$+0.135$, $+0.309$] at 14B and $+0.200$ [$+0.101$, $+0.315$] at 32B. The surplus halves but remains significant without arithmetic and survives nested layer selection. At 32B, an interior gate catches 40 of 71 errors at the 64-item budget, against 24 for entropy.

In a post-hoc cross-model decomposition across the 19 base models measured so far, the pattern has a simple shape. The best interior layer sits in a narrow band (within-type AUROC 0.715 to 0.854) while entropy ranges from 0.584 to 0.905, and the two are uncorrelated (Spearman -0.003). Consequently the surplus tracks the weakness of the output channel almost perfectly (Spearman -0.90 between entropy AUROC and surplus, and the same against the probability margin). Across the tested models the interior reader varies much less than the output reader; families differ mainly in how much of the signal the output distribution exposes.

Two qualifications keep this honest. First, the comparator matters. Against the top-2 probability margin, the strongest output scalar on this battery, the broad Qwen surplus shrinks to $+0.053$ and $+0.040$ and is no longer significant, and under nested layer selection nothing on the whole battery favors the interior over the margin. Second, the per-type decomposition shows where the interior's genuine advantage lives. Entropy wins on trivia for every model, while the probe wins on Qwen's computation-flavored types, most sharply where the output is confidently wrong (Qwen-14B arithmetic: entropy 0.222, probe 0.787; Qwen-32B obscure facts: 0.451 vs 0.952). Phi fails arithmetic constantly and shows no surplus, so this is not a property of computation but of the pair: the surplus lives where a model's output confidence decouples from its correctness, and which (model, type) pairs decouple is an empirical fact, cheaply diagnosable from entropy alone. That is the regime claim we carry forward, and the one the abstract states: internal error signals are most informative where outputs are confidently wrong.

Balani & Panda (2026) argue that during fixed-weight inference a wrong answer incurs no extra physical cost. We do not measure that cost, so the claim remains untested here. They cite Eldin (2026), who report nearly flat token entropy across task categories of very different accuracy and

infer that entropy cannot distinguish grounded outputs from confabulations. Our item-level results contradict that inference: category means do not determine discrimination within a category, where entropy ranks errors with AUROC 0.58 to 0.91 across 19 base models. In some model-task pairs, an interior reader adds information where entropy fails, and an outside controller can act on that signal. Section 6 asks how much the interior adds to routing; first we climb the explanation path, from selecting supplied causes to constructing an explanation without them.

D Extended results for Section 4: routine selection, recognition and retrieval

Table 1 begins with the step every diagnosis takes: choose the most probable of a few supplied causes. The resemblance route then asks whether a model can recognize, and retrieve, a structural twin at surface distance. This section climbs those three rows.

Routine abduction: good until the clues disagree (Experiment 7). We formulate routine abduction as Bayesian updating over a supplied hypothesis set. Each short vignette gives three candidate causes, their base rates, and how often each produces the observed effect. The posterior is exact, its winner beats the runner-up by a factor of at least 1.5, and the balanced answer is one token: A, B or C. Five item types separate the cues: easy; prior (the base rates decide); likelihood (the evidence decides); conflict (the posterior winner is neither the most common cause nor the best explainer); and archive (a folk cause pitted against the stated numbers). Qwen-2.5 base from 1.5B to 32B and Gemma-3-27B base answer two independently generated draws of 405 items in one pass.

One-cue items are solved by scale: at Qwen-32B, prior and likelihood accuracy reach 0.98 and 0.99 on the first draw and 0.97 and 0.96 on the second. Conflict items are not: 0.37, 0.29, 0.44 and 0.52 for Qwen 1.5B to 32B and 0.42 for Gemma-27B on the first draw, and 0.44, 0.36, 0.43, 0.52 and 0.47 on the second, the worst non-trivial cell in every model against chance of one third. With three causes, a conflict error picks either the most common cause or the best explainer. From 7B up the model picks the best explainer in 78 to 98 percent of conflict errors across both draws; at 1.5B, in 54 and 62 percent.

We fit Grether's log-linear form (Equation 1) to the letter probabilities. For each item and each pair of answer letters i, j,

$$\log \frac{P_{\mathrm{LLM}}(i \mid e)}{P_{\mathrm{LLM}}(j \mid e)} \;=\; \alpha \log \frac{P(e \mid i)}{P(e \mid j)} \;+\; \beta \log \frac{P(i)}{P(j)} \;+\; c_{ij},$$

where the likelihoods and base rates are the numbers stated in the vignette and the pair intercepts c_{ij} absorb answer-letter bias. The weights are estimated by least squares over all items whose causes map to letters uniquely (403 of 405 on the first draw, 402 on the second). Intervals come from 1000 bootstrap resamples of items, and the interval for $\beta - \alpha$ uses the same resamples as the two weights. Table D.1 gives the ten fits. The prior weight is below the likelihood weight in every fit, with every paired interval for $\beta - \alpha$ below zero. Both weights grow along the Qwen ladder in both draws, from about 0.1 at 1.5B to about 2 for the likelihood and 1.2 to 1.4 for the prior at 32B, while $\hat{\beta}/\hat{\alpha}$ stays between 0.37 and 0.66. Scale sharpens the inference without rebalancing it.

Why do the errors point one way? In a conflict item, the gold cause g, the most common cause P, and the best explainer L are distinct. The best explainer is rarer than g, whereas the most common cause explains the evidence less well. Up to letter bias, the fitted rule chooses the cause maximizing $\log P(e \mid h) + r \log P(h)$, where $r = \beta/\alpha$. For any $r \leq 1$, the gold beats P; it beats L only when r exceeds the item's tipping ratio $\log\{P(e \mid L)/P(e \mid g)\}/\log\{P(g)/P(L)\}$. Underweighting the prior can therefore move an error only toward the rarer, better-fitting cause. Because the fitted ratios exceed the tipping ratio on most items (median 0.23), this form explains the direction of the errors, not their rate. Cheng et al. (2026) prompt models with specified Grether parameters; here we estimate the weights from responses without prescribing an updating rule.

The conflict cell is also where the output loses track of its own errors. From 7B up, output entropy on conflict items is at or below chance as an error detector in three of four models on the first draw (0.36, 0.50 and 0.44; 0.57 at 14B) and all four on the second (0.41 to 0.48): the model is more confident when it is wrong, and the probability margin does no better. The interior probe reads the same errors

	First draw			Second draw		
Model	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\beta}/\hat{\alpha}$ [95% CI]	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\beta}/\hat{\alpha}$ [95% CI]
Qwen-2.5-1.5B	0.15	0.09	0.58 [0.45, 0.71]	0.14	0.09	0.65 [0.49, 0.82]
Qwen-2.5-7B	0.95	0.40	0.42 [0.35, 0.49]	1.05	0.39	0.37 [0.31, 0.44]
Qwen-2.5-14B	1.22	0.77	0.64 [0.59, 0.69]	1.28	0.82	0.64 [0.58, 0.70]
Qwen-2.5-32B	2.04	1.22	0.60 [0.55, 0.65]	2.07	1.36	0.66 [0.60, 0.72]
Gemma-3-27B	1.47	0.61	0.42 [0.33, 0.50]	1.63	0.66	0.40 [0.34, 0.48]

Table D.1: Grether fits of Equation 1 to the letter probabilities of Experiment 7, full battery, both draws. Brackets are paired 95% bootstrap intervals (1000 item resamples); every paired interval for $\beta - \alpha$ lies below zero. R^2 ranges from 0.46 to 0.73. Bayes is $\alpha = \beta = 1$. Sources: exp7_beta_fit.py, exp7_beta_fit_run2.py and exp7_beta_ratio_ci.py, 2026-09-25.

at 0.59 to 0.77 on the first draw and 0.44 to 0.73 on the second (Qwen-7B falls below chance on the second draw). At 14B the gap on the replication draw is $+0.250$ [$+0.119$, $+0.385$]. On Gemma-27B, which errs broadly, the surplus is significant over the whole battery ($+0.145$ [$+0.092$, $+0.203$] on the first draw, $+0.111$ [$+0.044$, $+0.174$] on the second, at the same layers, surviving every stored output scalar and a nested layer choice). The model that wrote its retrieval errors out in Section 3 keeps its integration errors inside: which regime a model sits in depends on the task, not only on the family.

One step further. On conflict items answered wrongly, a probe decodes the correct cause above chance: 0.58 [0.44, 0.72] at 14B and 0.74 [0.61, 0.86] at 32B on the first draw, and 0.49 [0.35, 0.63] and 0.70 [0.56, 0.84] on the second, against chance of one third. At 32B, however, the correct cause is already the output's runner-up on 95 and 93 percent of errors. The first draw's 14B advantage in a two-way comparison (interior 0.68 against output 0.54) also does not replicate (0.70 against 0.68). The probe could therefore be reading the cue rankings rather than a corrected Bayesian answer: the construction puts the correct cause second on both cues, so cue order and posterior agreement are confounded. No advantage over the output's own ranking was established across both draws. The structural tasks below never show an interior correspondence signal, and the frontier models expose no interior states to this study.

Recognizing a twin (Experiments 8, 8c and 8f). Section 2 introduced three systems, each a directed graph on five or six parts with two relation types. One alien sentence renders each edge. B preserves all of A's relations under new names. The decoy C instead applies a few degree-preserving edge swaps, so it matches A in its vertex, edge and degree counts but not in relational structure. The generator tries every vertex correspondence and rejects any C that remains isomorphic to A; exactly one of B and C is therefore A's twin.

We ask the same question at three surface distances. Near (relabel), B and C keep A's sentence pattern and change only the alien names ('silu vroks tami'). Middle (skin) also changes the sentence pattern ('the vrok link runs from silu to tami'). Far introduces a third pattern and words with different shapes ('whenever tup pulses, kem answers on the vok channel'), so neither the pattern nor even a word shape links the systems. Only far is a wormhole in the sense of Section 2; the nearer rungs measure losses caused by surface change alone. Easy items, whose decoy has two fewer edges than A, are the positive control. We interpret a hard-item null only when easy accuracy clears the pre-declared 0.70 gate. Each draw contains 280 forced choices between B and C: 40 easy and 80 at each distance. Table D.2 places one real line from each rung beside the same idea told as a love story.

In one forward pass, no open-weight model distinguishes the twin from the decoy above chance. Qwen-2.5 base from 1.5B to 32B and Gemma-27B sit at chance at every distance in both draws (accuracy 0.475 to 0.537, every interval covering 0.5; 0 of 15 model-by-distance cells above chance on the second draw) while solving the easy items (0.78 to 0.98 where the gate passes). Qwen-7B fails the gate by answering B on 279 of 280 items and is reported descriptively. Nor does an interior reader find the answer: a probe trained to identify the twin meets the pre-declared criterion at no hard distance in any model. The same probe recovers the easy count cue (0.90 to 0.97), so the reader has power where that cue exists. No structural-correspondence signal meets the registered criterion.

Perhaps one pass with no room to work is the wrong test. Experiment 8c applies several supports to Qwen-2.5-14B-Instruct, which passes the easy gate (0.875), on the same 280 items: a chat-template

Rung	Real line (Experiment 8, first draw)	The same idea as a love story
System A	'beve gupos saso.'	'Ann loves Bob.'
Relabel	'tiku nuvis nidi.'	'Kim adores Lee.'
Skin	'the pebe link runs from dori to feto.'	'The adore link runs from Kim to Lee.'
Far	'whenever nuk pulses, tal answers on the lam channel.'	'Whenever Kip sends a text, Rod's heart jumps.'

Table D.2: The surface-distance ladder of Experiment 8: one real line per rung, beside a love-story version for intuition. Each line states one directed edge: names are vertices, and the verb or its alien substitute marks one of two edge types. Relabel changes only the names; skin also changes the sentence form; far changes both the form and the shapes of the alien words. Nothing on the far surface says "loves"; a second edge type would read "Rod's hands shake". Real lines are the first line of system A and of the twin in items relabel_001, skin_001 and far_001.

Structural items, $n = 240$ per arm	DeepSeek-V4-Pro	Nemotron-3-Ultra
reasoning off	0.542 [0.478, 0.604]	0.504 [0.441, 0.567]
hidden reasoning	1.000 [0.984, 1.000]	0.979 [0.952, 0.991]
paired difference	$+0.458$ [$+0.396$, $+0.521$]	$+0.475$ [$+0.412$, $+0.537$]
easy items, off / on	0.950 / 1.000	0.900 / 1.000

Table D.3: Twin recognition at frontier scale (Experiment 8f). Accuracy on the 240 structural items (relabel, skin and far pooled; chance 0.5) with Wilson 95% intervals; the paired difference is a 5000-replicate item bootstrap. Same items, same prompt, same one-line answer; the arms differ only in whether the provider ran a hidden reasoning pass first.

forced choice, generic deliberation, an explicit node-and-relation scratchpad, a worked example from the same or a different class, and five ensembles that divide one token budget across parallel tries. Structural accuracy remains at chance in every arm (0.52 to 0.57). The model nevertheless reads local features: hubs, degrees, directions and relation kinds are above chance (0.85, 0.76, 0.59, 0.79), though two-step paths and cycles are not. Nine checker-certified complete isomorphisms appear across six arms, every one mapping onto the twin and never the decoy, but too rarely to change forced-choice accuracy. By the declared fence, these nulls can mean either a capacity limit or insufficient support; they do not establish absence.

The frontier changes the answer, but only with time to think. Experiment 8f runs a fresh certified draw through DeepSeek-V4-Pro and Nemotron-3-Ultra under pinned providers. The two arms differ only in hidden serial computation: reasoning is either disabled, leaving room for one answer line, or given a requested 6000-token hidden budget before that same line. Table D.3 gives the result. With reasoning off, both models sit at chance on the 240 structural items and collapse to a default letter, even though they read the easy count cue in one pass. DeepSeek answers C on 219 of 238 parsed structural items; Nemotron answers B on 235 of 240. With hidden reasoning, the same models on the same items reach 1.000 and 0.979. The paired gains are $+0.458$ [$+0.396$, $+0.521$] and $+0.475$ [$+0.412$, $+0.537$].

Three fences matter. First, the prompt names the correspondence to seek, so this is instructed recognition rather than spontaneous discovery. Second, each model family contributes one pinned configuration (fp8 and fp4 quantizations on different hosts), not a general claim about frontier models. Third, Qwen3.8 and MiniMax were excluded as instruments: Qwen's thinking could not be switched off, while MiniMax's reasoning traces truncated. An initial DeepSeek run used arm-blocked execution and had eight censored outputs; we repeated it on a fresh draw with interleaved arms and report that first run below. Related one-pass failures are reported by Thushalika et al. (2026) and Abdaljalil et al. (2026). What Experiment 8f adds is the within-model crossing from chance to near ceiling when hidden serial computation is enabled.

Experiment 8f initial run (seed 80; the fresh-draw run in Table D.3 is the registered result). The initial DeepSeek-V4-Pro run used the original seed-80 items, a 12,288-token total allowance, and arm-major ordering: all 280 direct calls preceded all 280 reasoning calls, confounding arm with clock time. Structural accuracy was 0.529 [0.466, 0.591] without reasoning and 0.967 [0.936, 0.983] with it, a paired gain of $+0.438$ [$+0.371$, $+0.504$]. The predeclared decisions match the fresh-draw run:

the direct arm does not clear chance, while hidden reasoning does. All eight reasoning-arm errors lacked a parsed final answer; seven reached the token cap and one was terminated by the provider at 7175 tokens. Every parsed answer was correct. Only 5 of 280 direct calls carried reasoning, below the predeclared 0.10 adherence limit; all five reached the 128-token cap and were unparsed. The run cost $4.95 and took 62 minutes. A disjoint pilot nine minutes earlier showed the same split and direct-arm default.

Retrieving a twin (Experiment 9). Recognition places both candidates in the prompt. Retrieval must find the candidate first. We ask two questions: when each system is embedded alone, does it sit closer to its twin than to its decoy, as MAC/FAC assumes for human analogical reminding (Forbus et al., 1995); and can it retrieve that twin from a library of 480 systems? We embed all 840 systems separately and choose the comparison layer on a held-out half.

Every Qwen model is null at every distance. Gemma-27B places the twin above the decoy on far items in the first draw (0.700 [0.592, 0.789]), but the pre-registered replication at the same layer returns chance (0.487), so we retire the positive. In library retrieval, the twin reaches the top five for 0 of 80 far queries in every model, with median rank near 390 of 480. Surface clustering dominates the tested similarities. A supervised metric and a second-order readout of within-system relational geometry do no better, and their power controls fire inconsistently across scales. This is a limit of the rulers, not evidence that no structural index exists. Hidden serial computation may construct a correspondence on the fly, but Experiment 8f does not establish that mechanism.

E Extended results for Section 5: construction, uncued detection and authored guides

Section 4 handed the model candidate explanations. Here the candidate set is withheld. We test construction after a contradiction, prediction with no announced structure, and guidance written by another instance of the same model.

Construction after a certified failure (Experiment 10). The circuit worlds of Section 2 are generated from a secret and checked by a program. They come in two families. In *wires*, dials are grouped into panels that feed displays; in an escape world, one display secretly also reads a dial outside its declared panel. In *roto*, named actions apply hidden affine transformations to five states, so sequences containing the same actions can end differently when their order changes. Passive records in each escape world fit a restricted explanation: a one-dial law in wires and an order-blind rule in roto. That explanation fails on a setting the model must choose. Forty adequate worlds per family, where no amendment is needed, guard against gratuitous rewriting; twelve noisy-but-adequate worlds test whether noise is mistaken for an anomaly.

DeepSeek-V4-Pro and Qwen3.8 each receive 40 escape worlds per family with a 48k-token budget. They write executable explanations, choose interventions, read the results, and revise. A certified chain must begin with the model's restricted explanation, introduce an amendment absent from the records, test it against a live rival with a discriminating intervention, revise on the result, and seal a program that reproduces both the records and fresh interventions. An oracle arm solves every world (40 of 40 in every cell).

Neither model completes the certified chain in either family: 0 of 40 in all four cells (upper bound 0.088). On the adequate worlds, neither ever rewrites an unbroken mechanism (0 of 40), so the null is not an instrument failure. With a four-program menu, however, exact final programs on matched twelve-world subsets rise from 1 to 12 for DeepSeek and from 2 to 11 for Qwen in wires, and from 0 to 12 and 0 to 11 in roto. Across both families, all 172 complete first-batch prediction vectors are arithmetically correct; 16 incomplete vectors are excluded. In wires, replaying every candidate against the first intervention batch leaves the true program as the unique match in 22 of 23 episodes. Given candidates, selection by intervention is nearly perfect.

Without the menu, the wires models emit a structural amendment in 9 and 11 of 40 worlds and produce a truth-equivalent final program in 4 of 40 each, but never within a chain that also contains a discriminating test. The chain criterion requires at least two live candidates at confirmation, whereas these models usually commit to one beforehand (33 and 36 of 40 episodes). We therefore report

the null with its conjunct. The menu result establishes that evidence-based identification usually succeeds once candidates are supplied; it cannot isolate why open-ended search fails because the menu changes information, representation, and search together. Roto says less about this gap because the models reach the certified-failure state in only 8 and 4 of 40 worlds.

The reachability arm validates this interpretation. One neutral commitment branches into two matched continuations using the model's own intervention: one reveals the contradiction and the other does not. Before the anomaly, neither model emits the target structural class (0 of 39 and 0 of 40, at the registered resolution). Afterward, it appears in 12 of 26 and 11 of 33 exposed wires worlds, but never in the matched control. The anomaly, rather than an extra call alone, prompts the correct class; what that class does remains uncertified.

Two descriptive textures qualify the result. DeepSeek rewrites 83 percent of noisy-but-adequate wires worlds, so noise is often mistaken for anomaly. The models also amend silently: they change the program in 9 to 11 of 40 worlds without ever declaring the frame inadequate. The trigger is half supplied. The contradictory observation is shown without a label, but the protocol asks for revised candidates and another intervention. Construction also remains inside a supplied program language; no new representation is required. Related tests study detection under a supplied ontology (Steiglechner et al., 2026), override of a provably wrong default (Shi et al., 2026), and end-to-end symbolic program search (Song et al., 2025). Here the model's own executable program is made to fail, and the effect of revealing that failure is measured directly.

Detecting structure with no cue (Experiment 11). Planck had a contradiction. Remove it. Each world is now an archive of observations with sealed entries to forecast. In a deep world, a compact rule from a registered class determines those entries. Its surface-matched flat twin has no rule from that class, so the sealed outcomes are undetermined. The honest policy predicts on deep and spreads probability on flat. We score the probability on the deep truth minus the largest probability on any flat outcome, averaged over two relabelings of each latent pair. Identical forecasts on both twins cannot score above zero.

The two families are a multi-view field (degree-2 to degree-5 polynomials over an alien presentation of $\mathbb{F}_{11}$, viewed through four coordinate charts; 11 outcomes) and quiet wires (the panel from Experiment 10 viewed passively; 4 outcomes). We compare five arms: natural, with every reference to hidden structure banned; a length-matched placebo; a generic search cue; decoded, which supplies the representation but withholds the law; and scaffolded, which also names the hypothesis family. A bounded surface reader cannot distinguish deep from flat, and the pre-declared gate passes in both families.

DeepSeek-V4-Pro, at medium reasoning effort, is tested once on a fresh battery of 224 worlds. Without a cue it does not distinguish rule-bearing worlds from their rule-free twins: the natural contrast is -0.640 (one-sided 95% upper bound -0.559) in the field and -0.286 (upper bound -0.211) in wires, against a registered bound of 0.15. This is not an abstaining null. Among valid packets, the natural arm puts its top probability on the deep truth in 11 of 61 field worlds but assigns at least half its mass to one undetermined outcome in 55 of 59 flat worlds. It concentrates where it should spread and misses the rule where one exists.

Naming the hypothesis family changes the result. DeepSeek finds the deep truth in 64 of 64 and 47 of 48 valid packets and spreads probability on flat, giving contrasts of $+0.785$ (lower bound $+0.723$) and $+0.585$ (lower bound $+0.502$). The generic search cue and its placebo do not improve reliably; decoding alone helps partly in the field ($+0.313$) and not in wires (-0.041). Nemotron-3-Ultra replicates the wires comparison: natural -0.173 (upper bound -0.127), decoded -0.018, and scaffolded $+0.667$ (lower bound $+0.611$), with all 24 deep worlds correct. Its natural forecasts are nearly uniform rather than concentrated, but they fail the same bound. Nemotron's field pilot produced too few valid scaffolded answers within the 48k-token allowance, so the two-family result rests on DeepSeek and the two-model result on wires.

A guide the model wrote itself (Experiment 11b). If the family must be named, can the model name it? Three fresh DeepSeek instances each study separate solved quiet-wires examples and write a bounded guide. We freeze those guides before fresh recipients forecast held-out matched pairs under four conditions: decoded observations alone; decoded plus the human family sentence; decoded plus one model-authored guide; or decoded plus a length-matched generic note.

Across the three guides, the average improvement over generic advice ($+0.105$ [$+0.023$, $+0.187$], six latent pairs) comes from less concentration on rule-free worlds, not better prediction on rule-bearing ones. The gain over decoding alone is imprecise ($+0.260$ [-0.093, $+0.613$]); only one guide beats its generic control; and the human sentence remains far stronger (0.424 against 0.083). Treating two undelivered packets adversarially makes the guide-versus-generic interval include zero. A Nemotron-authored guide prescribes a linear family over $\mathbb{Z}_4$ that fits none of its worked examples; recipients given it return uniform forecasts on every world, while the human sentence gives the same model a contrast of 0.75. In two archived author traces, the model derives contradictions to candidate rule classes and then recommends those classes in its final guide. Reliable transfer from model-authored guides is not established. What transfers, when anything does, is restraint.

F Extended results for Section 6: retry routing and Experiment 13 sensitivities

Section 3 showed that an outside reader can abstain on the interior's error signal. Abstention is the weakest action. The two paths now meet: can a detected error select useful computation, or repair the answer, and does the interior buy anything a reader of the output does not? In both experiments the routing loop is ours; the model is never asked to read itself.

Routing a second pass (Experiment 12). We return to the routine-abduction battery with Gemma-27B, the model whose interior reader best detects these errors, and add an external loop. A kernel-ridge error router is fitted on the first draw and applied without labels to the second draw's 405 items. It sends the most suspicious quarter (101 items) through the same weights for a second pass.

The remedy matters more than the routing. A worked prior-times-likelihood example repairs 110 of 114 first-pass errors and breaks none of the 291 correct answers. Merely asking the model to think again repairs 29 errors and breaks 7; the model usually repeats its answer after three tokens. The difference in repair fractions is $+0.711$ [$+0.632$, $+0.789$]. As in Experiment 8c, generic deliberation does little, though the tasks and contrasts differ.

Because the supplied method repairs almost every retried error, routing becomes a precision-at-budget problem. Among 101 selections, the interior router finds 73 of 114 errors; a learned label-free router on the output distribution finds 59; generic uncertainty scores find fewer; and a router given item type as privileged metadata finds more than any of them. Final error is 0.109 with the interior router and 0.136 with the learned output router, a difference of $+0.026$ [$+0.000$, $+0.052$]. The lower bound is exactly zero, so the advantage is borderline. The interior clearly beats generic uncertainty ($+0.057$ [$+0.030$, $+0.084$]) but loses to the type-aware output router (-0.039 [-0.059, -0.020]). Its second-draw within-type AUROC is 0.835, against 0.739 for the best label-free output reader.

Thus an outside reader can use the interior signal to allocate a second pass better than naive output uncertainty, but it does not demonstrably beat a trained reader of the output. The intervention was predeclared after screening first-pass results on the same draw, so it is not an independent confirmation. Dip et al. (2026) route second passes from the output to save computation. Villuri et al. (2026) likewise find little advantage for internal probes over the answer margin.

Does the interior matter more when evidence is harder to recover? (Experiment 13). Experiment 13 renders the same three-cause problem in four forms while holding its probabilities and correct answer fixed. The transparent form states every probability directly. A padded control adds unused lookup entries to match length. In the shallow form, each cause points through one printed codebook to its probability; the deep form uses two successive codebooks. The manipulation changes how evidence is recovered, not whether the problem is decidable. It changes Gemma-27B's strategy more than its accuracy: the likelihood-favored answer falls from 69 percent of padded items to 56 percent of shallow items, while first-pass accuracy moves from 0.51 to 0.45.

We compare the interior probe with bounded output readers using surface features and answer probabilities, the full logit vector, or a capped MLP, without oracle-solved input features. On shallow items, the interior's increment is small and not established: held-out log-loss gain $+0.023$ [-0.001, $+0.047$] and routing gain $+0.015$ [-0.010, $+0.043$] at a 25 percent budget. The registered contrast subtracts its padded-item routing gain from its shallow-item gain. It is negative in point (-0.030

[-0.063, $+0.005$]) and significantly negative against the MLP reader. Across transparent, padded, shallow, and deep forms, the gains decline from $+0.045$ and $+0.045$ to $+0.015$ and $+0.005$, opposite our pre-registered prediction.

Read alone, the probe catches about as many errors as the strongest output reader: 75, 80, 70 and 65 per 100 routed, against 66, 77, 68 and 63. A declared post-hoc two-score stack detects no additional routing benefit over that reader on any form. This is not evidence that the interior contains nothing; it is compatible with redundancy, limited power, score-scale shift, or model mismatch. The registered verdict is unresolved with a negative lean, for one model in one-pass inference.

Three analysis deviations and sensitivities matter. *Layer selection.* The executed code selected the probe layer using ordinary and relabeled training items, whereas the plan specified ordinary items only. The executed rule chose layer 38 (all-skin out-of-fold AUROC 0.6950); the plan-conforming rule chose layer 42 (ordinary-only AUROC 0.6925). Both give an unresolved overall verdict, but the log-loss increment changes from established at layer 38 (P13-1 letter (a)) to unresolved at layer 42 (letter (u)). We treat layer 42 as registered and layer 38 as a sensitivity.

Full-logit reader. R3, a kernel reader on 262,262 features built from the final-token logits, has strong held-out routing but a training-CV log-loss of 4.06 nats. Near-interpolation produces a calibration failure, and the registered joint learner gives the single interior column almost no weight (-0.00047). Its zero increment therefore shows that this learner failed to use the probe; it is not a structural result about where the information resides.

Post-hoc stack. A declared sensitivity combines each reader's scalar score with the interior score in a two-feature calibrated ridge, fitted on training data and evaluated on held-out draws. Against R3, the stack detects no additional routing benefit (75/80/69/65 errors caught per 100 routed, versus 66/77/68/63 for R3 alone). This bounded result remains compatible with redundancy, limited power, score-scale shift, or model mismatch.

Across these experiments, the value of the interior depends on both model and task. Sometimes the output already exposes the error signal, as for base Pythia and Phi on our batteries, Gemma on retrieval, and the toy model before grokking. Elsewhere an interior reader adds information, as for selected Qwen computation and cue-conflict cells and for the grokked toy during severe collapses. An outside controller can use that signal to abstain or allocate a supplied second pass, but Experiment 13 finds no established routing advantage over the strongest tested output reader.

The model is never asked to read itself; every loop here is external. Graph of States (Luo et al., 2026), which maintains an explicit hypothesis-evidence graph and backtracks on conflict, is one implemented example of such a controller. The alarm can be heard from outside. Whether it identifies which remedy is needed remains untested.

G THE WORMHOLE DESIGN

Section 2 uses a wormhole as an image for structural resemblance across surface distance. Two domains may share almost no vocabulary or presentation and still organize their entities, transformations, and consequences in the same way. The wormhole is that relational bridge. It is not a new matching algorithm: structure-mapping theory, SME, and MAC/FAC already formalize related ideas (Gentner, 1983; Falkenhainer et al., 1989; Forbus et al., 1995), and Farmer's identity-abduction loop gives a recent scientific formulation (Farmer, 2026). Our use is diagnostic. We ask which parts of finding, checking, and using such a bridge are present in a frozen model.

Write two domains as

$$\mathcal{D}_A = (V_A, \mathcal{R}_A, \mathcal{T}_A), \qquad \mathcal{D}_B = (V_B, \mathcal{R}_B, \mathcal{T}_B),$$

where V contains entities or states, $\mathcal{R}$ contains typed relations, and $\mathcal{T}$ contains transformations with observable consequences. A candidate wormhole consists of a partial entity map $f : V_A \dashrightarrow V_B$ and a map ϕ between relation or transformation types. Its defining test is whether the diagram commutes:

$$f\big(T_A^i(x)\big) \;\approx\; T_B^{\phi(i)}(f(x))\,.$$

Apply corresponding transformations in the two domains, and the mapped outcomes should agree. Equality is exact in our synthetic worlds and approximate in scientific analogies. A resemblance may

suggest an explanation before it has been tested, but it earns evidential credit when consequences not used to construct the map also travel through it.

Traversing such a wormhole separates into five operations:

1. *Trigger*: register that the present explanation deserves reconsideration, or receive a resemblance worth pursuing.
2. *Retrieval*: find a promising foreign domain among many possibilities.
3. *Alignment*: construct the correspondence between entities and relations.
4. *Validation*: test the correspondence on withheld consequences.
5. *Enactment*: let a validated correspondence change the next inference or action.

The experiments isolate pieces of this sequence rather than demonstrating one autonomous traversal. Experiments 1–7 ask when an outside reader can detect and use an error signal; Experiment 9 tests bounded retrieval and finds no reliable structural index under the tested rulers; Experiments 8 and 8f test alignment when both candidate domains and the question of correspondence are supplied; Experiment 10 tests construction and validation after a certified failure; and Experiments 3 and 12 let an external controller enact a supplied response. The firewall in Appendix B controls false acceptance when validation uses fresh evidence.

This decomposition also marks the boundary of the claim. Routine Bayesian selection in Experiment 7 is not a wormhole. Nor do the failure-route experiments themselves require a cross-domain jump: a contradiction can demand a search without specifying where to search. A decodable alarm therefore does not imply a reachable explanation, and we do not infer a hierarchy of hypothesis sets from probe surplus. The missing link is a mechanism by which the structure of a problem or contradiction becomes an address into a large conceptual archive. None of our experiments shows a model deciding for itself when to launch that search, retrieving an unrestricted domain, validating the mapping, and allowing it to redirect inquiry as one continuous act.