OUTSIDE THE DICTIONARY: LLM-GUIDED EQUATION DISCOVERY IS LIMITED BY WHAT IT PROPOSES

Anonymous authors
Paper under double-blind review

ABSTRACT

Symbolic regression, including its recent language-model-guided variants, searches a fixed dictionary of primitives, and fails when the governing law needs a primitive outside that dictionary or a quantity that was never measured. Such failures are invisible to in-distribution evaluation: we prove that on a bounded observed domain, at every sample size and every positive noise level, some in-dictionary expression is statistically indistinguishable from the true law. We therefore score hypotheses by extrapolation, and use this to measure what limits a standard discovery loop in which a language model proposes closed-form laws that are then fitted, selected, and tested by active probing. We contribute CRISIS-60, 60 procedurally generated environments in which a low-order polynomial fits the observed data almost perfectly but a law counts as recovered only if it extrapolates to a disjoint region with $R^2 > 0.999$; its control category, whose law is itself a polynomial, is solved at 100.0% by polynomial regression against 0.0% elsewhere. We also contribute the Epicycle Index, a ratio of extrapolation error to description length anchored at the incumbent theory, and a decomposition of recovery into proposal coverage, whether a correct law was ever proposed, and selection efficiency, whether it was then chosen. Selection is nearly saturated: it returns a correct law on 98.0% of the instances where one was proposed, so a perfect selector would gain only 1.1 points, whereas on 3 of 15 structural families no correct law is ever proposed. This accounts for why the loop's parsimony penalty and active probing contribute nothing measurable (95% upper limits 0.6 and 6.4 points), and the same limit holds for LLM-SR (Shojaee et al., 2025), whose representation and search differ from ours. Under observational noise the limit moves to selection, whose efficiency falls to 14.3%. The loop recovers 53.3% of laws (95% CI 42.8 to 63.9) against 23.9% for genetic symbolic regression and 31.1% for LLM-SR at matched budget; with five times the candidates LLM-SR reaches 51.7%, so the loop's advantage is efficiency rather than attainable recovery. Because the decomposition needs only the candidates a search already scores, it tells the builder of any such system which half is worth improving: here, the proposal distribution. Code, benchmark, and all logged runs are available at `https://anonymous.4open.science/r/outside-the-basis-1867`.

1 INTRODUCTION

Automated discovery systems recover closed-form descriptions of data by searching expressions built from a fixed dictionary of primitives (Langley et al., 1987; Cranmer et al., 2020; Udrescu & Tegmark, 2020; Petersen et al., 2021), and given the right dictionary they are strong, recovering a large fraction of ground-truth expressions on standard benchmarks (La Cava et al., 2021). The dictionary is supplied by the practitioner, and this is where the framing quietly does much of the work: Kepler could not have found an elliptical orbit inside a dictionary of circles, and no amount of search over epicycles would have produced one.

We separate two operations that symbolic discovery usually conflates. Parameter interpolation searches for coefficients within the span of a dictionary $\mathcal{B}$ and can always improve the fit on observed data by admitting more terms. Basis extension introduces a primitive $\phi^* \notin \mathrm{span}(\mathcal{B})$, or postulates a

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.

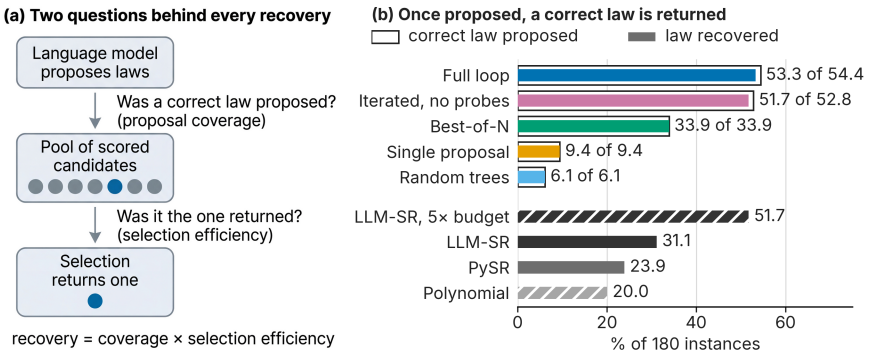


Figure 1: (a) Recovery requires that a correct law be proposed and that selection then return it, so it factorises into proposal coverage and selection efficiency (Equation 4); schematic. (b) Measured over the 180 noise-free instances of Section 6. The frame is coverage and the fill recovery, so a fill short of its frame is a correct law proposed but not returned. Every configuration that records its candidate pool is shown, and selection efficiency across them ranges from 96.9% to 100.0%; baselines record none and show recovery alone. Table 1 gives intervals for all.

latent coordinate that was never measured, and so changes the hypothesis space rather than moving within it. The two are indistinguishable on interpolation data and diverge sharply outside it, because a sufficiently flexible in-basis expansion reproduces any smooth relation on a bounded observed domain. The scope this creates is narrower than the framing may suggest and we state it at the outset: these environments test selection between the two operations under observational equivalence, plus one category requiring a constant in a structurally consequential position rather than the coefficient position where constants ordinarily appear. Inventing a primitive that no grammar supplies is not tested here, and Section 8 gives the boundary in full.

The resulting failure is specific and recognisable. A method adds terms until the observable regime is explained, achieves a near-perfect in-distribution fit, and extrapolates catastrophically, because the added complexity encoded the sampling window rather than the mechanism. We call this the Ptolemaic failure mode, and it is not hypothetical: genetic symbolic regression here produces a saturating kinetics hypothesis of complexity 27 attaining $R^2 = 0.999939$ on held-out validation data and extrapolating at $R^2 = -4.41 \times 10^9$ (Appendix L). Diagnosing it needs a metric that charges for complexity and credits only generalisation, which accuracy on an in-distribution split cannot do.

We make four contributions, and the discovery loop we run is not among them: coupling a language-model proposer to an evaluator is established practice (Romera-Paredes et al., 2024; Shojaee et al., 2025; Xia et al., 2026; Wang et al., 2025), and complexity-penalised selection is standard in symbolic regression (Cranmer et al., 2020). First, Theorem 1 shows that at any sample size and positive noise level, no test reading only observable-region data can separate an in-basis fit from the law it imitates. Second, CRISIS-60 (Section 4) comprises 60 procedurally generated environments over 15 structural families, with a control category inside the polynomial basis. Third, the Epicycle Index (Section 3) is an extrapolation diagnostic anchored so that the incumbent theory scores one. Fourth, we decompose recovery into proposal coverage and selection efficiency (Figure 1; Section 7) and find selection nearly saturated, for our loop and for LLM-SR, which accounts for all four component measurements, including the two null results. The same instrumentation shows that iteration pays through conditioning on what failed rather than through repetition.

2 Problem Formulation

Let $\mathcal{B} = \{\phi_1, \ldots, \phi_k\}$ be a dictionary of primitives and let $\mathcal{H}(\mathcal{B})$ denote the set of finite expressions over $\mathcal{B}$, the observable variables $\mathbf{x} \in \mathbb{R}^d$, and real constants. Let $\mathcal{D}_{\text{obs}} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$ be measurements drawn from an observable region $\mathcal{X}_{\text{obs}} \subset \mathbb{R}^d$, generated by $y = f^*(\mathbf{x}) + \xi$ with $\xi \sim \mathcal{N}(0, \sigma^2)$, and let $\mathcal{X}_{\text{ood}}$ be disjoint from $\mathcal{X}_{\text{obs}}$.

Definition 1 (In-basis interpolation and basis extension)**.** *A hypothesis H is an* in-basis interpolation *if $H \in \mathcal{H}(\mathcal{B})$. It is a* basis extension *if it employs a primitive $\phi^* \notin \mathrm{span}(\mathcal{B})$, or if it introduces a free coordinate q_L that appears in no column of $\mathcal{D}_{\mathrm{obs}}$ and whose value must be inferred jointly with the remaining parameters.*

The second clause is weaker than it may appear. A term such as $B/(r - r_0)^2$ is expressible in any grammar that composes variables with constants, so a dictionary-restricted search is not barred from writing it; the clause identifies a constant in a structurally consequential position, inside the argument of a term where it relocates a pole. PySR sometimes reaches such a position (Section 6), so the question is how often a search does, not whether it can.

The reason an evaluation restricted to $\mathcal{X}_{\mathrm{obs}}$ cannot separate the two operations is not that the separation is hard but that the evidence for it is absent. The following is proved in Appendix A.

Theorem 1 (In-distribution indistinguishability)**.** *Let $\mathcal{X}_{\mathrm{obs}}$ be compact, let f^* be continuous on it, let $\mathbf{x}_1, \ldots, \mathbf{x}_N \in \mathcal{X}_{\mathrm{obs}}$ be any design, and write P_f^N for the law of $\mathcal{D}_{\mathrm{obs}}$ when the mean function is f. For every N, every $\sigma > 0$, and every $\delta \in (0, 1)$ there is an in-basis $P_d \in \mathcal{H}(\mathcal{B})$ with*

$$\mathrm{TV}\big(P_{f^*}^N,\, P_{P_d}^N\big) \;\le\; \frac{\sqrt{N}}{2\sigma} \sup_{\mathbf{x} \in \mathcal{X}_{\mathrm{obs}}} \big|f^*(\mathbf{x}) - P_d(\mathbf{x})\big| \;\le\; \delta. \tag{1}$$

Consequently any test of $f = P_d$ against $f = f^*$ that reads only $\mathcal{D}_{\mathrm{obs}}$ and holds level α has power at most $\alpha + \delta$.

More data does not break the tie. If f^* is analytic on a neighbourhood of $\mathcal{X}_{\mathrm{obs}}$, its best degree-$d$ approximation satisfies $\sup |f^* - P_d| \le C\rho^{-d}$ for some $C > 0$ and $\rho > 1$ by Bernstein's theorem, so Equation 1 holds at level δ as soon as $d \ge \log(C\sqrt{N}/2\sigma\delta)/\log\rho$ (Corollary 1). The description length an in-basis expansion must spend to stay statistically indistinguishable therefore grows only logarithmically in the number of observations: quadrupling the data costs the epicycle a fixed number of further terms, not a change of status. The bound needs $\sigma > 0$; the main experiments are noise-free, where it reduces to its approximation half, that some in-basis expansion matches the law on the observed window within any tolerance a finite-precision fit can resolve, and Section 7 also measures the noisy regime the bound addresses directly.

The theorem is about the observable window rather than any particular criterion, so it is not answered by a better in-distribution selection rule, by cross-validation inside $\mathcal{X}_{\mathrm{obs}}$, or by more data. Extrapolation costs little by comparison: at mean-square separation Δ on the extrapolation region a likelihood-ratio test errs with probability at most $\exp(-N_{\mathrm{ood}}\Delta^2/8\sigma^2)$, so $N_{\mathrm{ood}} = \mathcal{O}(\sigma^2\Delta^{-2}\log(1/\eta))$ points suffice (Proposition 1), and the separation is bounded below: Proposition 2 shows that against a screened law no polynomial of degree at least one stays within any finite error on an unbounded domain, while every one of them is uniformly ϵ-accurate on the compact observed window. Uniform approximation inside is always available, so in-distribution accuracy carries no information about whether the basis was extended; divergence outside is unavoidable, so extrapolation does.

3 The Epicycle Index

Detecting the Ptolemaic failure mode requires charging a hypothesis for the description length it consumes and crediting it only for the generalisation it buys. Neither quantity is scale-free on its own, so both need a reference, and we take it to be the incumbent theory H_0 fitted on the same data.

Definition 2 (Epicycle Index)**.** *Let $\mathrm{Comp}(H)$ be the number of nodes in the abstract syntax tree of H and let $\mathrm{NRMSE}_{\mathrm{ood}}(H)$ be its root-mean-square error on $\mathcal{X}_{\mathrm{ood}}$ normalised by the standard deviation of the target there. Define*

$$\rho_C(H) = \frac{\mathrm{Comp}(H)}{\mathrm{Comp}(H_0)}, \qquad \rho_E(H) = \frac{\mathrm{NRMSE}_{\mathrm{ood}}(H)}{\mathrm{NRMSE}_{\mathrm{ood}}(H_0)}, \qquad \mathcal{E}_I(H) = \rho_C(H)\,\rho_E(H). \tag{2}$$

Equation 2 is a ratio of ratios and therefore invariant to the units and magnitude of the underlying law, and its anchor is exact rather than conventional: $\mathcal{E}_I(H_0) = 1$ by construction, so values below one mark hypotheses that bought more extrapolative power than the description length they cost and values above one those that did not, a hypothesis doubling in size while halving extrapolation error

being exactly break-even. A definition whose value instead tends to zero as a hypothesis becomes correct has no interpretable reference point and must import a threshold from outside; anchoring on H_0 removes that freedom. The index reads the extrapolation split and is therefore a diagnostic rather than a selection rule: no method here may use it to choose, and Equation 3 sees only held-out data from the observable region.

4 THE CRISIS-60 BENCHMARK

CRISIS-60 contains 60 environments in 5 categories, each realised by two to four structural families so that solving a category cannot be reduced to memorising one functional form (Appendix B). Constants are drawn procedurally per environment and seed, giving 180 instances over 3 seeds. Each environment exposes an observable region, from which training and validation data are drawn and within which all active queries must fall, and a disjoint extrapolation region used only for scoring.

The screened, saturating and relativistic categories require a primitive outside the polynomial basis, namely an exponential cutoff, a rational saturation or a square-root singularity. The latent category requires positing a coordinate offset that appears in no measured column; Proposition 4 establishes that it is identifiable in the noise-free limit for the pole family. The in-basis category is a control whose law is a polynomial in the measured variable: a method that discovers structure at all should solve it, and a method that solves only it has not extended any basis. Across the suite a polynomial whose order is selected on validation data attains a median R^2 of at least 0.9992 in the observable regime in every category, so no category is distinguishable from the others by in-distribution difficulty. Nothing we ran saturates the suite: pooling every method in this study, 66.7% of instances are recovered by at least one and 60 by none, so 33.3% of the benchmark remains open, concentrated in the latent and relativistic categories rather than spread thinly.

We report contamination directly rather than asserting its absence. The forms are textbook physics and claiming they are absent from pretraining corpora would not be credible; what is randomised is the parameterisation, so the claim is not that the forms are unseen but that identifying which one governs a particular unlabelled environment, and estimating a latent coordinate within it, is not retrievable from memory. Appendix E measures what a model produces with no data at all.

5 THE DISCOVERY LOOP WE INSTRUMENT

The loop alternates proposal, verification and experiment (Algorithm 1). We give it no name: each configuration is identified by the components it contains, and the one with all four enabled is the full loop. We claim no novelty for the architecture, whose parts are drawn from existing practice: language-model proposal from FunSearch and LLM-SR (Romera-Paredes et al., 2024; Shojaee et al., 2025), complexity-penalised selection from symbolic regression (Cranmer et al., 2020), and disagreement-based querying from Bayesian experimental design (MacKay, 1992). Each part is specified because Section 7 removes them one at a time. Nothing in the loop sees ground truth; its only channel to the environment is a probe inside the observable region, which cannot leak the extrapolation split.

The proposer receives the observations, the incumbent law and its fit quality, the Pareto frontier in structural form, and a description of where the leading hypothesis is wrong, and returns expressions with unknown constants written symbolically. The prompt states the whole permitted grammar, never names a primitive identifying a particular family, and is the grammar the enumerative control also receives. Constants may stand for quantities never measured, which is how a latent coordinate enters: `c0/r**2 + c1/(r - c2)**2` postulates an unobserved locus `c2` that the fitter estimates. Candidates are parsed and fitted by multi-restart least squares; those failing to parse, not depending on an observable, or producing non-finite predictions are discarded and counted (Appendix K). Selection minimises a description length,

$$L(H) = n \log\left(1 - R^2_{\mathrm{val}}(H)\right) + \lambda\, \mathrm{Comp}(H), \tag{3}$$

the Gaussian negative log-likelihood on held-out data up to a constant plus λ nats per syntax-tree node, at $\lambda = 5 \approx \log n$. Log-likelihood space is what makes the trade-off well posed: a criterion linear in $1 - R^2$ is dominated by the penalty as soon as the fit is good, since the error term is bounded by one while the penalty is not, and a search using it returns the incumbent regardless of the evidence,

Table 1: Recovery on CRISIS-60 over 180 instances, with 95% cluster bootstrap intervals over environments. Coverage is the fraction of instances whose pool held a qualifying hypothesis and selection efficiency the fraction of those the method returned, both blank where the run recorded no pool. "Struct." is candidate expressions per instance, and for PySR its internal evolutionary budget.

Method	**Recovery (%)**	**Cover.**	**Sel. eff.**	**Symbolic**	**Med. $\mathcal{E}_I$**	**$\mathcal{E}_I > 1$**	**Struct.**
Polynomial epicycles	20.0 [10.0, 30.0]	–	–	20.0	5.33	79.4	6
PySR (genetic SR)	23.9 [15.0, 33.3]	–	–	19.4	0.91	46.8	19800
PySR ($5\times$ budget)	25.0 [15.0, 36.7]	–	–	25.0	0.67	38.6	250000
Random trees + identical search	6.1 [2.8, 10.0]	6.1	100.0	6.1	1.23	72.5	32
Proposer, single shot	9.4 [3.3, 16.7]	9.4	100.0	9.4	1.02	52.8	1
Proposer, best-of-N	33.9 [23.3, 45.0]	33.9	100.0	29.4	0.31	25.7	31
LLM-SR (Shojaee et al., 2025)	31.1 [20.6, 42.2]	–	–	26.7	0.31	36.1	32
LLM-SR ($5\times$ budget)	51.7 [38.3, 65.0]	–	–	45.0	0.05	30.0	160
Proposer + iteration	51.7 [41.1, 62.2]	52.8	97.9	46.7	0.03	21.2	28
Proposer + iteration + probes	**53.3** [42.8, 63.9]	54.4	98.0	48.3	0.01	21.8	27

which we observed before adopting Equation 3. When the frontier holds competing hypotheses the loop queries the simulator where they disagree most inside the observable region, which Proposition 3 shows maximises expected evidence per query.

6 EXPERIMENTS

Setup and baselines. The proposer is Qwen3.5-35B-A3B unless stated otherwise, at temperature 0.9 and top-p 0.95, with every setting in Table 11. Every method receives $C_{\mathrm{abduct}} = 32$ structures, $C_{\mathrm{deduct}} = 6$ restarts per candidate, and the identical scorer. A hypothesis recovers the law when its extrapolation R^2 exceeds 0.999, the accuracy-solution criterion standard in this literature; a stricter symbolic-equivalence rate is reported alongside, and Appendix H repeats the analysis at other thresholds. Intervals are a cluster bootstrap over environments, since the 180 instances are 60 environments crossed with 3 seeds. We report exact p-values from 13 paired tests without a multiplicity correction, and Appendix K gives the accounting under one. Polynomial regression selects its order on validation data, PySR (Cranmer, 2023) receives the same grammar at its default budget and five times it, LLM-SR (Shojaee et al., 2025) is executed from its reference implementation on the same proposer and scorer at both budgets (Appendix J), and the enumerative proposer samples random expression trees over the identical grammar inside the identical loop.

The benchmark measures basis extension, and genetic search produces Ptolemaic hypotheses. Polynomial regression recovers 100.0% (36/36) of the in-basis control and 0.0% of the four out-of-basis categories (Table 4), despite fitting the observable regime of all five at a median R^2 of at least 0.9992. The separation is total and it is the point of the design: the out-of-basis categories are not harder to fit, they are impossible to extrapolate from within the incumbent basis. PySR reaches 23.9%, concentrated in the same control at 88.9%, and its failures elsewhere are not poor fits but hypotheses spending their nodes on the sampling window rather than the mechanism, one of them fitting at $R^2_{\mathrm{val}} = 0.999993$ while extrapolating at $R^2 = -5.47 \times 10^3$ (Appendix L); 46.8% of them carry $\mathcal{E}_I > 1$. At five times the budget it reaches 25.0% (CI 15.0 to 36.7, one seed), overlapping its default-budget interval, so more search effort does not close the gap.

Against the closest published system, at two budgets. LLM-SR (Shojaee et al., 2025) shares this architecture and reports gains concentrated out of domain, so it carries the most information. Executed from its reference implementation on the same proposer, scored by the same function, and given the same distinct-structure budget, it recovers 31.1% (CI 20.6 to 42.2), leading PySR by +7.2 points, a margin this evidence does not establish ($p = 0.0854$). Two comparisons are more informative. It is not distinguishable from drawing the same number of independent proposals and keeping the best, at -2.8 points ($p = 0.302$), so at this budget its island-structured experience buffer does not measurably outperform sampling. And it sits +22.2 points below the full loop (CI [+13.3, +31.1], $p = 2.29 \times 10^{-8}$, 47/7 discordant).

That last gap is a difference in cost and not in ceiling. LLM-SR targets thousands of samples, so 32 structures is far below its intended regime; given 160 it reaches 51.7% (CI 38.3 to 65.0, one seed), nominally *above* the full loop by -3.3 points ($p = 0.774$). Neither difference is distinguishable from zero and we claim nothing from the sign. Arriving takes the baseline 160 structures and 74 generation calls per instance against 27 and 7: it does not search worse, it arrives later.

The ladder, and where it stops paying. The full loop recovers 53.3% of instances (Table 1; 95% CI 42.8 to 63.9), against 23.9% for PySR (+29.4 points, CI [+20.0, +39.4], $p = 1.66 \times 10^{-12}$, 58/5 discordant pairs) and 20.0% for polynomial regression (+33.3, CI [+22.8, +44.4]); its symbolic-equivalence rate is 48.3% of instances and 90.6% of its recoveries, so most are the correct structure rather than accurate extrapolations, and only 21.8% of its hypotheses exceed the $\mathcal{E}_I = 1$ anchor against 79.4% for the polynomial. Within the ladder, a single proposal recovers 9.4%, below PySR; drawing 31 independent proposals and keeping the best raises this to 33.9%; residual feedback between rounds raises it further to 51.7% (+17.8 points, CI [+10.0, +26.1], $p = 9.43 \times 10^{-7}$), and does so while evaluating *fewer* distinct structures, 28 against 31. The gains are not free: at matched structure budget the full loop issues 7 generation calls per instance against 4 for best-of-N.

Where it fails. The relativistic category is hardest for every method, with the full loop at 22.2% (8/36) and every baseline at or near zero, and Appendix D describes what each of the resistant families demands and what the loop writes down instead. Parse failures account for 0.1% of proposals.

The Epicycle Index tracks extrapolation failure. Evaluated on all 20418 hypotheses scored during search, with rank correlations taken within each environment before summarising because the index is normalised by each environment's own incumbent while raw error is not, its median within-environment correlation is $\rho = 0.815$ over 60 environments, with a minimum of 0.067, and the separation at the anchor is total: 0.0% of hypotheses with $\mathcal{E}_I > 1$ recovered the law against 12.2% of those at or below it (Figure 3 (a)), the anchor being what the incumbent takes by construction rather than a fitted threshold. Appendix I shows what the complexity charge adds over extrapolation error alone, and the bound on that claim: a threshold tuned with hindsight beats the untuned anchor.

7 Which Components Contribute, and Why

7.1 Two of four components do not earn their cost

Table 6 removes one component at a time at matched proposal budget. Two of the four carry the result and two do not measurably contribute, and publishing the negative half is the point: a loop of this shape is easy to assemble and the cost of its parts is not otherwise visible. The enumerative control is the most informative row: given the identical grammar, selection, probing and budget, random expression trees recover 6.1% against 53.3% for the same loop driven by a language model (+47.2, CI [+37.8, +56.7]), so the search machinery is not sufficient alone; and one proposal without any search recovers 9.4%, below genetic symbolic regression, so the proposer is not either.

The other two components measure nothing, and a null result is worth only the bound that comes with it. Setting $\lambda = 0$ leaves recovery at 52.8% on 1/0 discordant instances of 180, and removing active probing costs +1.7 points at $p = 0.69$. The same conditioning turns each into a 95% upper limit: the parsimony term cannot be contributing more than 0.6 points and probing more than 6.4 (5.6 by the clustered bootstrap), so neither is merely unresolved at this size. On a suite half this one the probing arm looked suggestive at the ten per cent level and we reported it as such; doubling the benchmark resolved it against the component. Re-selecting offline from the recorded pools across three orders of magnitude leaves recovery flat at 53.3% through $\lambda = 20$, falling to 52.8% at $\lambda = 50$ and 4.4% at $\lambda = 500$ (Figure 3c). The rest of this section shows both null results have one cause, not specific to either component.

7.2 Recovery is limited by what gets proposed

A run assembles a pool $\mathcal{P}$ of scored hypotheses and returns one, so recovery factorises exactly:

$$\Pr[\text{recover}] \;=\; \underbrace{\Pr\left[\exists\, H \in \mathcal{P} : R^2_{\text{ood}}(H) > \tau\right]}_{\text{proposal coverage } \kappa} \;\times\; \underbrace{\Pr\left[R^2_{\text{ood}}(\hat{H}) > \tau \;\middle|\; \exists\, H \in \mathcal{P} : R^2_{\text{ood}}(H) > \tau\right]}_{\text{selection efficiency } \eta}, \tag{4}$$

where $\hat{H}$ is what Equation 3 returns. The first factor belongs to the proposal distribution and the second to the selection rule. Coverage is the recovery an oracle selector would reach from the same proposals, so $\kappa - \Pr[\text{recover}]$ is the entire headroom available to any better selection criterion, and both factors are read from the logs at no compute, since each run logs every hypothesis it scored.

Figure 1 (b) shows both, and the second factor is nearly saturated everywhere: every configuration selects at between 96.9% and 100.0% of its own oracle. The full loop covers 54.4% of instances and returns a recovering hypothesis on 53.3%, losing 2 of 98, so perfect hindsight over the identical proposals would gain 1.1 points, with an interval reaching 2.8. That is a fact about the rule and not about pools that offer it no choice: on covered instances a median of 4 of 25 scored candidates meet the criterion, and only 12.2% hold exactly one. Both null components follow, and the decomposition says which factor each acts through. Removing the parsimony term moves coverage by +0.0 points on 0 discordant instances, so it is purely selectional and the headroom above is its ceiling. Probing is not selectional at all: efficiency is 98.0% with it and 97.9% without, while coverage moves +1.7 points (CI [-3.3, +7.2]), which is the whole of its effect. Its null is a fact about proposals, which Proposition 3 anticipates.

Why the rule is nearly oracle-optimal is measurable rather than arguable, and Appendix I derives it: the penalty weight at which the objective would first reverse a correct hypothesis has median 462 over the 23 instances where that hypothesis is also the larger, against an operating point of $\lambda = 5$, so the rule runs two orders of magnitude below where parsimony would begin to cost it anything.

Where coverage fails it fails completely. Of the 15 structural families (Appendix H), 3 have coverage of exactly zero: across 33 instances and every hypothesis the loop scored in them, a law meeting the criterion was never once proposed, and reporting them as merely unrecovered would leave open whether the search found the right structure and rejected it. Zero coverage is a fact about this proposal distribution at this budget and not about the problem, since a differently biased search recovers 4 of those instances and a larger budget reaches most of the families outright: what is missing is not reachability but the prior that would make these forms worth writing down early. The diagnostic is not specific to this loop and costs nothing: any search that scores its candidates already holds the information, and the verdict says which half of a system is worth working on. We ran it on the published baseline rather than asserting that it transfers. With LLM-SR re-executed at matched budget and its candidate programs recorded, coverage is 31.7% against recovery of 30.0%, so it selects at 94.7% of its own oracle and leaves 1.7 points on 60 instances. On the same seed and environments the loop covers 51.7% and selects at 93.5%, so the +18.3 points of recovery between them decompose into +20.0 of coverage and -1.7 of selection, the second favouring the baseline. A system whose representation, search and experience buffer are not ours is limited in the same factor, and is not worse than ours in the other one.

7.3 Compute axes, budget and reach

Only one of the two test-time compute axes moves coverage: averaging over the other, C_{abduct} moves it 62.7 points from 3.3% to 66.0% while C_{deduct} spans 4.7 and is not monotone, and recovery follows at 60.7 points against 4.0. That is the cause of the recovery asymmetry rather than a restatement of it, since optimisation improves the fit of a structure it is handed and cannot add one: if proposals carry a per-instance hit probability $p \sim \nu$, then $\kappa(C_{\text{abduct}}) = 1 - \mathbb{E}_\nu[(1-p)^{C_{\text{abduct}}}]$ increases with supremum $\nu(p > 0)$, while C_{deduct} enters only through a fit-success probability bounded by one.

Whether the residual difficulty is budget or reach is a question about that supremum, which extending the proposal axis bears on and, as it turns out, does not settle. Fitting κ over the fitted budgets with ν a Beta distribution, so that every instance is eventually reachable, and again with an unreachable mass added, gives two curves indistinguishable over that range; Proposition 5 shows they differ in kind outside it, the miss rate decaying as C^{-a} under the first and onto a floor under the second. We ran the larger budget on the same seed and C_{deduct}, withheld from both fits, and filled the fitted column in

on the environments the grid had skipped, so that fit and test both rest on 60 environments. What the budget bought needs no law: coverage moves from 50.0% at 32 to 76.7% at 128, reaching 16 of the 30 instances it had missed with 0 going the other way, and 2 of the 3 zero-coverage families above are reached at 128. Coverage and recovery coincide there, so selection lost nothing at four times the pool. What the fitted laws say about it is not reportable, and the reason is worth more than the verdict would have been. Fitting each half of the environments separately, each carrying every budget, returns shape parameters of 1.96 and 0.11, a factor of 18, and reachable fractions of 78.0% and 40.0%. The two halves then disagree about what they predict: on one the measurement 83.3% (25/30) falls below the reachable-everywhere prediction of 93.4% ($p = 0.0454$) and is consistent with the saturating 77.9% ($p = 0.324$), while on the other 70.0% (21/30) lies *above* both, at $p = 0.0173$ and $p = 8.30 \times 10^{-4}$. A two-parameter law fitted to five proportions over a range that has not begun to saturate does not survive refitting on a different sample of the same benchmark, and Appendix H shows the same failure on a curve with no data in it. We report what the budget bought and claim nothing about whether the residual difficulty is budget or reach.

7.4 Iterating pays by conditioning, and only above a proposer scale

Iteration confounds two things that a comparison against best-of-N cannot separate: drawing proposals over several rounds rather than at once, and conditioning each round on what the earlier ones got wrong. Withholding the frontier and the residual context and changing nothing else separates them, on 30 instances measured under one configuration. Rounds without conditioning recover 26.7% against 36.7% for the same budget drawn at once, -10.0 points ($p = 0.508$) with the sign against the rounds and an exact upper limit of 9.3 points on what re-sampling alone could be contributing. Restoring the conditioning takes the same loop to 53.3%, +26.7 points above it (CI [+6.7, +46.7], $p = 0.0386$, 10/2 discordant), and the decomposition puts that gain where the rest of this section puts everything else: coverage moves +30.0 points ($p = 0.0225$). What iteration buys is the conditioning and not the repetition.

Varying the proposer from 0.5B to 35B parameters under an otherwise identical loop gives the clearest structural result here, and it lies in the gap rather than in either curve. Best-of-N does not improve measurably with scale at all, at 35.0%, 35.0%, 24.2% and 30.0% from 4B to 35B with every interval overlapping every other, while the full loop rises from 26.7% to 55.0% and the difference widens monotonically from -8.3 points at 4B ($p = 0.0414$) to +25.0 at 35B ($p = 2.27 \times 10^{-7}$); below 8B the machinery is worthless or mildly harmful. The decomposition says what the gap is made of: over the same range the loop's coverage rises from 28.3% to 56.7% while matched-budget sampling moves from 35.0% to 30.0%, so conditioning the proposer on what has failed enlarges the set of structures it writes down, and only once it is large enough for that conditioning to change what it writes. Appendix C gives the full curve and the parse-failure rates that make its low end readable. Whether the trend continues at the capability frontier we cannot answer with open-weights models alone.

7.5 Under noise the binding constraint moves

Robustness to observational noise, in full in Appendix C, is the clearest negative result here and where the account above stops applying. With the criterion held fixed, recovery falls from 55.0% at $\sigma = 0$ to 29.2% at $\sigma = 10^{-2}$ and 2.5% at $\sigma = 10^{-1}$; the ordering against PySR survives but the absolute capability does not. The decomposition says that the binding constraint moves as it falls. Coverage declines from 56.7% to 38.3% and 17.5% while selection efficiency collapses from 97.1% to 76.1% and 14.3%: at the highest level the correct structure is still proposed on 17.5% of instances and is almost never the one returned, because noise corrupts the held-out fit the selection rule reads. What is proposal-limited without noise becomes selection-limited with it, which is the regime in which a better selection criterion would pay, and it is the one this work does not address.

8 Limitations

What "outside the dictionary" means here is narrower than the phrase suggests. Definition 1 is relative to the incumbent's polynomial span, while the proposer's grammar already contains exponential, logarithmic, square-root and hyperbolic primitives, so three categories test selecting a supplied

primitive against an equally good in-basis fit rather than inventing one, and PySR's 16.7% on the latent category shows that a structurally placed constant is within reach of ordinary search. Genuine ontological extension is not tested.

The environments are synthetic, algebraic and low-dimensional, with at most three observables, and the loop is not robust to observational noise, so nothing here is evidence about discovery from measured data. Every proposer is an open-weights Qwen model of at most 35B parameters, so the capability frontier is untested and the scale trend of Section 7 rests on a single model family. The decomposition is an oracle over the pool a run realised rather than over the hypothesis space, and it holds only without noise: selection efficiency falls to 14.3% at $\sigma = 10^{-1}$. Contamination is bounded rather than excluded, since the forms are textbook physics and a zero-shot probe (Appendix E) recovers laws without data in 1 of 5 categories for the main proposer.

9 Related Work

Symbolic regression searches a primitive dictionary using genetic programming (Cranmer et al., 2020; Cranmer, 2023), physics-informed decomposition (Udrescu & Tegmark, 2020) or learned policies (Petersen et al., 2021), benchmarked by La Cava et al. (2021), and is strong when the dictionary contains the required primitives. Definition 1 does not describe a case these methods are formally barred from: it describes a region of the search space they reach infrequently, which our measurements quantify rather than assume. Language models have recently been coupled to program search for discovery, by FunSearch (Romera-Paredes et al., 2024), AlphaEvolve (Novikov et al., 2025), and closest to our setting LLM-SR, which treats equation discovery as program synthesis and reports gains concentrated out of domain (Shojaee et al., 2025), with agentic variants adding tool use and experience reuse (Xia et al., 2026; Wang et al., 2025) and in-context (Merler et al., 2024) and knowledge-informed (Taskin et al., 2026) variants pursuing related aims inside a fixed dictionary. We run the LLM-SR implementation itself on CRISIS-60 rather than only citing it. Our contribution is not the coupling, which these works establish, but the isolation of basis extension as a measurable capability, a metric for its failure mode, and an account of what limits such a loop.

Section 7 borrows the coverage-and-selection split that repeated-sampling work uses for reasoning, where coverage grows log-linearly with samples while majority voting and reward models plateau, so what binds is selecting a correct sample rather than producing one (Brown et al., 2024), and the best allocation of test-time compute depends on the difficulty of the instance (Snell et al., 2024). Equation discovery inverts that relation: extrapolation into a disjoint region is an unusually strong verifier, so selection is nearly free and coverage binds. That literature also fits coverage curves in order to extrapolate them, and Section 7 is a caution about the practice rather than an application of it. Active querying connects the loop to Bayesian experimental design (MacKay, 1992; Settles, 2009; Foster et al., 2021) and the distinction we operationalise is long-standing (Peirce, 1934; Hanson, 1958; Magnani, 2009); that language models struggle with parsimony and formal abduction is reported directly (Xu et al., 2023; Liu et al., 2024; Sun & Saparov, 2025), and Zahavy (2026) argues they cannot perform such leaps unaided, which our enumerative and single-shot controls address.

10 Conclusion

We study what limits language-model-guided equation discovery when the governing law lies outside the dictionary. We prove that interpolation data cannot distinguish searching a fixed dictionary from extending it, and we introduce CRISIS-60 and the Epicycle Index to make that distinction measurable, holding in-distribution fit constant across categories that differ only in whether the law lies inside the basis. We then decompose recovery into proposal coverage and selection efficiency and find it proposal-limited: selection returns a correct law on 98.0% of the instances where one was proposed in our loop, and on 94.7% in LLM-SR. A better parsimony weight, verifier or stopping rule can therefore recover at most the 1.1-point oracle gap, so effort belongs on the proposal distribution, although under observational noise the verdict reverses. Because the diagnostic needs only the candidates a search already scores, any discovery system can run it at no cost. What decides which structures get written down remains the open problem.

AI USE STATEMENT

Language models appear in this work in two distinct roles, and we separate them because only the second bears on research conduct. As an *object of study*, open-weights models serve as the hypothesis proposers that the paper evaluates; their versions, decoding parameters, and prompts are documented in Appendix K and are part of the measurement rather than of its production.

As a *research tool*, large language models were used to help draft and edit the manuscript prose and to help implement and refactor the analysis and plotting code.

ETHICS STATEMENT

This work studies discovery of closed-form laws in synthetic physical environments. It involves no human subjects, no personal data, and no dual-use hazard.

REPRODUCIBILITY STATEMENT

The benchmark generator, discovery loop, baselines, scoring function, analysis, and figure and table generation are provided in the supplementary material and at `https://anonymous.4open.science/r/outside-the-basis-1867`, together with the logged result files themselves, one JSON record per evaluated instance carrying every hypothesis that instance scored. Every number in this paper, in text and in tables alike, is emitted directly from those files by the included generation script, so the manuscript cannot disagree with the experiments it reports, and the decomposition of Section 7 is recomputable from them without a GPU. Seeds $(42, 123, 456)$, environment parameters, prompts, and budgets are documented in Appendix K.

An included test suite asserts the properties this evaluation depends on rather than leaving them to prose: that replacing the extrapolation split with noise leaves the selected hypothesis bit-identical, so nothing in the search can be reading it; that probe regions are disjoint from extrapolation regions in every environment; that the incumbent scores $\mathcal{E}_I = 1$ exactly, as Definition 2 requires; that polynomial regression fits every environment's validation data above $R^2 = 0.99$ while extrapolating successfully in the in-basis control and nowhere else, which is the benchmark's design property; and that the coverage decomposition of Equation 4 is exact on every logged row, with selection efficiency above a floor, which is the mechanism the paper's account rests on.

REFERENCES

Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V. Le, Christopher Ré, and Azalia Mirhoseini. Large language monkeys: Scaling inference compute with repeated sampling. *arXiv preprint arXiv:2407.21787*, 2024.

Miles Cranmer. Interpretable machine learning for science with PySR and SymbolicRegression.jl. *arXiv preprint arXiv:2305.01582*, 2023.

Miles Cranmer, Alvaro Sanchez-Gonzalez, Peter Battaglia, Rui Xu, Kyle Cranmer, David Spergel, and Shirley Ho. Discovering symbolic models from deep learning with inductive biases. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:17429–17442, 2020.

Adam Foster, Desi R. Ivanova, Ilyas Malik, and Tom Rainforth. Deep adaptive design: Amortizing sequential Bayesian experimental design. In *International Conference on Machine Learning (ICML)*, pp. 3384–3395, 2021.

Norwood Russell Hanson. *Patterns of Discovery: An Inquiry into the Conceptual Foundations of Science*. Cambridge University Press, 1958.

William La Cava, Patryk Orzechowski, Bogdan Burlacu, Fabrício Olivetti de França, Marco Virgolin, Ying Jin, Michael Kommenda, and Jason H. Moore. Contemporary symbolic regression methods and their relative performance. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2021.

Pat Langley, Herbert A. Simon, Gary L. Bradshaw, and Jan M. Zytkow. *Scientific Discovery: Computational Explorations of the Creative Processes*. MIT Press, 1987.

Emmy Liu, Graham Neubig, and Jacob Andreas. An incomplete loop: Deductive, inductive, and abductive reasoning in language models. In *Conference on Language Modeling (COLM)*, 2024. arXiv:2404.03028.

David J. C. MacKay. Information-based objective functions for active data selection. *Neural Computation*, 4(4):590–604, 1992.

Lorenzo Magnani. *Abductive Cognition: The Epistemological and Eco-Cognitive Dimensions of Hypothetical Reasoning*, volume 3 of *Cognitive Systems Monographs*. Springer Science & Business Media, 2009.

Matteo Merler, Katsiaryna Haitsiukevich, Nicola Dainese, and Pekka Marttinen. In-context symbolic regression: Leveraging large language models for function discovery. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pp. 427–444. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-srw.49.

Alexander Novikov, Ngân Vũ, Marvin Eisenberger, Emilien Dupont, Po-Sen Huang, Adam Zsolt Wagner, Sergey Shirobokov, Borislav Kozlovskii, Francisco J. R. Ruiz, Abbas Mehrabian, M. Pawan Kumar, Abigail See, Swarat Chaudhuri, George Holland, Alex Davies, Sebastian Nowozin, Pushmeet Kohli, and Matej Balog. AlphaEvolve: A coding agent for scientific and algorithmic discovery. *arXiv preprint arXiv:2506.13131*, 2025.

Charles Sanders Peirce. *Collected Papers of Charles Sanders Peirce*, volume 5 of *Pragmatism and Pragmaticism*. Harvard University Press, 1934.

Brenden K. Petersen, Mikel Landajuela, T. Nathan Mundhenk, Claudio P. Santiago, Soo K. Kim, and Joanne T. Kim. Deep symbolic regression: Recovering mathematical expressions from data via risk-seeking policy gradients. In *International Conference on Learning Representations (ICLR)*, 2021.

Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Matej Balog, M. Pawan Kumar, Emilien Dupont, Francisco J. R. Ruiz, Jordan S. Ellenberg, Pengming Wang, Omar Fawzi, Pushmeet Kohli, and Alhussein Fawzi. Mathematical discoveries from program search with large language models. *Nature*, 625:468–475, 2024.

Burr Settles. Active learning literature survey. *Computer Sciences Technical Report 1648, University of Wisconsin–Madison*, 2009.

Parshin Shojaee, Kazem Meidani, Shashank Gupta, Amir Barati Farimani, and Chandan K. Reddy. LLM-SR: Scientific equation discovery via programming with large language models. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2404.18400.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024.

Yunxin Sun and Abulhair Saparov. Do language models follow Occam's Razor? an evaluation of parsimony in inductive and abductive reasoning. *arXiv preprint arXiv:2509.03345*, 2025.

Bilge Taskin, Wenxiong Xie, and Teddy Lazebnik. Knowledge integration for physics-informed symbolic regression using pre-trained large language models. *Scientific Reports*, 16(1):1614, 2026. doi: 10.1038/s41598-026-35327-6.

Silviu-Marian Udrescu and Max Tegmark. AI Feynman: A physics-inspired method for symbolic regression. *Science Advances*, 6(16), 2020.

Runxiang Wang, Boxiao Wang, Kai Li, Yifan Zhang, and Jian Cheng. DrSR: LLM based scientific equation discovery with dual reasoning from data and experience. *arXiv preprint arXiv:2506.04282*, 2025.

Shijie Xia, Yuhan Sun, and Pengfei Liu. SR-Scientist: Scientific equation discovery with agentic AI. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2510.11661.

Fangzhi Xu, Qika Lin, Jiawei Han, Tianzhe Zhao, Jun Liu, and Erik Cambria. Are large language models really good logical reasoners? a comprehensive evaluation and beyond. *arXiv preprint arXiv:2306.09841*, 2023.

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

A PROOFS

A.1 PROOF OF THEOREM 1, AND THE LOGARITHMIC COROLLARY

Corollary 1 (More data does not break the tie)**.** *Suppose in addition that f^* is analytic on a neighbourhood of $\mathcal{X}_{\text{obs}}$, so that its best degree-d approximation satisfies $\sup_{\mathcal{X}_{\text{obs}}} |f^* - P_d| \le C\rho^{-d}$ for some $C > 0$ and $\rho > 1$, as Bernstein's theorem provides. Then Equation 1 holds at level δ as soon as*

$$d \;\ge\; \frac{1}{\log \rho} \log \frac{C\sqrt{N}}{2\sigma\delta}. \tag{5}$$

Proof. Fix $\varepsilon > 0$. Because $\mathcal{X}_{\text{obs}}$ is compact and f^* continuous, the Stone-Weierstrass theorem supplies a degree d and a $P_d \in \mathcal{H}(\mathcal{B})$ with $\sup_{\mathcal{X}_{\text{obs}}} |f^* - P_d| \le \varepsilon$. Under either mean function the observations are independent Gaussians of common variance σ^2 at the same design points, so

$$\mathrm{KL}\big(P^N_{f^*} \,\|\, P^N_{P_d}\big) \;=\; \frac{1}{2\sigma^2} \sum_{i=1}^{N} \big(f^*(\mathbf{x}_i) - P_d(\mathbf{x}_i)\big)^2 \;\le\; \frac{N\varepsilon^2}{2\sigma^2}. \tag{6}$$

Pinsker's inequality gives $\mathrm{TV} \le \sqrt{\mathrm{KL}/2} \le \varepsilon\sqrt{N}/(2\sigma)$, which is the first inequality of Equation 1; choosing $\varepsilon \le 2\sigma\delta/\sqrt{N}$, which Stone-Weierstrass permits, gives the second. The power bound is the standard consequence: for any test φ with $\mathbb{E}_{P_d}[\varphi] \le \alpha$, $\mathbb{E}_{f^*}[\varphi] - \mathbb{E}_{P_d}[\varphi] \le \mathrm{TV}(P^N_{f^*}, P^N_{P_d}) \le \delta$.

For the corollary, substitute the approximation rate $C\rho^{-d}$ for ε and solve $C\rho^{-d} \le 2\sigma\delta/\sqrt{N}$ for d. The rate itself is Bernstein's theorem: on a compact interval, a function analytic in a neighbourhood is approximated by polynomials of degree d with error $O(\rho^{-d})$, where ρ is the parameter of the largest Bernstein ellipse in which the function is analytic. All but two families of the benchmark are univariate and this is the applicable statement; in d_x variables the same geometric rate holds for a tensor-product construction with constants depending on d_x, which changes the constant C but not the logarithmic dependence on N that the corollary asserts. □

The argument is elementary and we claim nothing for its machinery: Stone-Weierstrass, the Gaussian Kullback-Leibler identity and Pinsker's inequality are standard, and the statement is the specialisation of a familiar indistinguishability bound to the dictionary-extension setting. Its role is to replace an assertion the discovery literature usually leaves informal with something that can be checked, and to fix what the benchmark must therefore score.

The theorem is stated for a fixed design so that it covers the benchmark's construction exactly, in which the observable sample is drawn once and reused. Nothing changes if the design is random on $\mathcal{X}_{\text{obs}}$: conditioning on the design and averaging preserves both the Kullback-Leibler bound and the total-variation bound, since the bound holds uniformly over designs.

A.2 EXTRAPOLATION SEPARATES AT BOUNDED COST

Proposition 1. *Let H_1, H_2 be evaluated at N_{ood} points of $\mathcal{X}_{\text{ood}}$ with mean-square separation $\Delta^2 = N_{\text{ood}}^{-1} \sum_j (H_1(\mathbf{x}_j) - H_2(\mathbf{x}_j))^2$, and let observations there carry the same Gaussian noise of scale σ. Under equal priors the likelihood-ratio test between them has error probability at most $\exp(-N_{\text{ood}}\Delta^2/8\sigma^2)$, so $N_{\text{ood}} \ge 8\sigma^2 \log(1/\eta)/\Delta^2$ observations suffice for error at most η.*

Proof. The two hypotheses are Gaussian mean vectors separated by squared Euclidean distance $S = N_{\text{ood}}\Delta^2$ in $\mathbb{R}^{N_{\text{ood}}}$ with isotropic covariance $\sigma^2 I$. The Bayes error of the likelihood-ratio test between two such alternatives with equal priors is $\Phi(-\sqrt{S}/(2\sigma))$, and the Gaussian tail bound $\Phi(-t) \le e^{-t^2/2}$, valid for $t \ge 0$, gives $\exp(-S/(8\sigma^2))$. Rearranging for S gives the sample size. □

Read against Theorem 1 the two statements bracket the problem. Inside the observable window, discriminating an in-basis expansion from the law it imitates requires a separation that the expansion is free to make arbitrarily small, at a description-length cost that grows only logarithmically in

the sample size. Outside it, the separation is bounded below by Proposition 2 and the number of observations needed to act on it is logarithmic in the confidence and independent of how well either hypothesis fitted. This is the sense in which the benchmark's extrapolation split is not merely a harder test but a different kind of evidence.

A.3 In-basis fits do not extrapolate

Proposition 2. *Let $f^*(r) = Ar^{-2}(1+\mu r)e^{-\mu r}$ with $A, \mu > 0$, and write $P_d(r) = \sum_{j=0}^{d} w_j r^j$. Then no polynomial approximates f^* uniformly on unbounded domains: for every d and every $\mathbf{w}$,*

$$\lim_{R\to\infty} \sup_{r\geq R} |P_d(r) - f^*(r)| \;=\; \begin{cases} \infty & \text{if } w_j \neq 0 \text{ for some } j \geq 1, \\ |w_0| & \text{otherwise,} \end{cases} \tag{7}$$

which is bounded away from zero in every case except $P_d \equiv 0$, and $P_d \equiv 0$ has error $\sup_{r\geq R} f^*(r) > 0$ at every finite R. By contrast, on any compact $\mathcal{X}_{\text{obs}}$ and for any $\epsilon > 0$ there exist a finite d and coefficients with $\sup_{r\in\mathcal{X}_{\text{obs}}} |P_d(r) - f^*(r)| < \epsilon$.

Proof. Since $f^*(r) = A\mu r^{-1}e^{-\mu r} + Ar^{-2}e^{-\mu r}$, we have $f^*(r) \to 0$ as $r \to \infty$, faster than any inverse power. Let $m = \max\{j : w_j \neq 0\}$ when $\mathbf{w} \neq \mathbf{0}$. If $m \geq 1$ then $|P_d(r)| \sim |w_m| r^m \to \infty$, so $|P_d(r) - f^*(r)| \to \infty$ and the supremum over $[R, \infty)$ is infinite for every R. If $m = 0$ then $P_d \equiv w_0 \neq 0$ and $|P_d(r) - f^*(r)| \to |w_0|$, giving the second branch of Equation 7: finite, but bounded away from zero. If $\mathbf{w} = \mathbf{0}$ the error is $\sup_{r\geq R} f^*(r)$, which is positive for every finite R because f^* is strictly positive. No polynomial is therefore uniformly consistent on $[R, \infty)$, and every polynomial of degree at least one diverges.

For the approximation claim, f^* is continuous on the compact set $\mathcal{X}_{\text{obs}}$, so by the Stone-Weierstrass theorem the polynomials are dense in $C(\mathcal{X}_{\text{obs}})$ under the supremum norm, giving a finite d and coefficients achieving error below any $\epsilon > 0$. □

The two halves together state that in-distribution accuracy is uninformative about basis extension while extrapolation accuracy is informative, which is the property the benchmark is built to exploit.

A.4 Evidence rate of disagreement probing

Proposition 3 (Evidence rate of disagreement probing)**.** *Let H_1, H_2 be two hypotheses and let observations at $\mathbf{x}$ be $y = H_{\text{true}}(\mathbf{x}) + \xi$, $\xi \sim \mathcal{N}(0, \sigma^2)$. The expected log-likelihood ratio in favour of the true hypothesis from a single query at $\mathbf{x}$ equals $\|H_1(\mathbf{x}) - H_2(\mathbf{x})\|^2/(2\sigma^2)$. Consequently, if every query achieves separation at least δ, then the number of queries needed to drive the posterior odds against the false hypothesis below η is at most $\lceil 2\sigma^2 \log(1/\eta)/\delta^2 \rceil$.*

Proof. Write $\Delta(\mathbf{x}) = H_1(\mathbf{x}) - H_2(\mathbf{x})$ and suppose without loss of generality that $H_1 = H_{\text{true}}$. For a single observation $y = H_1(\mathbf{x}) + \xi$ with $\xi \sim \mathcal{N}(0, \sigma^2)$, the log-likelihood ratio is

$$\Lambda(\mathbf{x}) = \log \frac{p(y \mid H_1)}{p(y \mid H_2)} = \frac{1}{2\sigma^2}\Big(\|y - H_2(\mathbf{x})\|^2 - \|y - H_1(\mathbf{x})\|^2\Big) = \frac{1}{2\sigma^2}\Big(\|\Delta(\mathbf{x}) + \xi\|^2 - \|\xi\|^2\Big). \tag{8}$$

Taking expectations over ξ and using $\mathbb{E}[\xi] = 0$ gives $\mathbb{E}[\Lambda(\mathbf{x})] = \|\Delta(\mathbf{x})\|^2/(2\sigma^2)$, which is the Kullback-Leibler divergence between the two predictive distributions at $\mathbf{x}$. The rule $\mathbf{x}^* = \arg\max_{\mathbf{x}} \|\Delta(\mathbf{x})\|^2$ therefore maximises expected evidence per query. Over N_q independent queries each with separation at least δ, the expected accumulated log-odds is at least $N_q\delta^2/(2\sigma^2)$, so $N_q \geq 2\sigma^2 \log(1/\eta)/\delta^2$ suffices in expectation to reach log-odds $\log(1/\eta)$. □

We state this as a proposition rather than a theorem because it is a direct specialisation of a standard sequential-testing argument; its role is to justify the probe rule, not to claim a new result. It also makes the limitation explicit: the guarantee is in expectation and requires a separation floor δ, which fails exactly when two hypotheses agree throughout the observable region and differ only outside it. That case is realised in our benchmark by the latent-halo family, where the unobserved centre lies beyond the probe domain, and it is where active probing cannot help.

A.5 IDENTIFIABILITY OF A LATENT COORDINATE

Proposition 4. *Let $F(r) = c_1 r^{-2} + c_2 (r - r_0)^{-2}$ with $c_1, c_2 > 0$ and $r_0 > 0$. The parameters (c_1, c_2, r_0) are uniquely determined by the restriction of F to any open interval $\mathcal{I} \subset (r_0, \infty)$ on which it is observed without noise.*

Proof. F is a rational function, hence analytic on $\mathbb{C} \setminus \{0, r_0\}$. If two such functions agree on an open interval $\mathcal{I}$, they agree on a set with an accumulation point, so by the identity theorem for analytic functions they agree on the whole connected domain of definition and in particular have the same poles and principal parts. The pole locations recover $\{0, r_0\}$ and hence r_0, since $r_0 > 0$ is the unique nonzero pole; the residues of the order-two poles recover $c_1 = \lim_{r \to 0} r^2 F(r)$ and $c_2 = \lim_{r \to r_0} (r - r_0)^2 F(r)$. The continuations to $r = 0$ and $r = r_0$ are legitimate because analytic continuation from $\mathcal{I}$ determines F globally, even though neither point lies in $\mathcal{I}$. □

Identifiability in the noise-free limit does not imply recoverability at finite noise, and the sensitivity of the pole location to perturbation grows as the observation window recedes from r_0. Section 6 reports the empirical consequence.

A.6 WHAT A PROPOSAL BUDGET CAN BUY

Proposition 5. *Let an instance carry a probability p that a single proposal meets the recovery criterion, let $p \sim \nu$ across instances with proposals drawn independently, so that coverage at budget C is $\kappa(C) = 1 - \mathbb{E}_\nu[(1-p)^C]$, and write $f = \nu(p > 0)$ for the reachable mass. If ν restricted to $(0, 1]$ is $\mathrm{Beta}(a, b)$, the miss rate $M(C) = 1 - \kappa(C)$ satisfies*

$$M(C) \;=\; (1 - f) \;+\; f \, \frac{\Gamma(a+b)}{\Gamma(b)} \, C^{-a} \big(1 + O(C^{-1})\big), \qquad C \to \infty. \tag{9}$$

Consequently the budget attaining coverage $1 - \varepsilon$ is $C(\varepsilon) = \big(f \, \Gamma(a+b)/\Gamma(b)\big)^{1/a} \, \varepsilon^{-1/a} (1 + o(1))$ when $f = 1$, and is infinite for every $\varepsilon < 1 - f$ when $f < 1$.

Proof. On the reachable mass the Beta integral is exact for every C,

$$\mathbb{E}\big[(1-p)^C \,\big|\, p > 0\big] \;=\; \frac{B(a, b+C)}{B(a, b)} \;=\; \frac{\Gamma(a+b)}{\Gamma(b)} \cdot \frac{\Gamma(b+C)}{\Gamma(a+b+C)}, \tag{10}$$

by $B(x, y) = \Gamma(x)\Gamma(y)/\Gamma(x+y)$. Stirling's series gives $\Gamma(z+\beta)/\Gamma(z+\alpha) = z^{\beta-\alpha}\big(1 + O(z^{-1})\big)$ for fixed α, β, so with $z = C$, $\beta = b$ and $\alpha = a + b$ the ratio is $C^{-a}(1 + O(C^{-1}))$. On the complementary mass $(1-p)^C = 1$ for every C, and adding the two parts gives Equation 9. Setting $M(C) = \varepsilon$ with $f = 1$ and solving for C gives the stated budget; when $f < 1$, $M(C) \geq 1 - f$ for every C. □

Only the behaviour of ν near the origin enters. If ν restricted to $(0, 1]$ has a density regularly varying at zero with index $a - 1 > -1$, a Tauberian argument replaces the exact integral and returns the same C^{-a} decay with a different constant, so the dichotomy is a statement about whether an unreachable mass exists and not about the two-parameter family fitted in Section 7.

This is what makes the budget extension a test rather than a description. The two candidate laws are not two curves of similar shape through the same five points: they differ in the limit, one approaching complete coverage along a power law whose exponent is the shape parameter and the other stopping at $1 - f$, so the quantity an extended budget estimates is an exponent against a floor. The proposition also prices the optimistic reading. At the exponent the refit returns once the held-out budgets are included, $a = 0.62$, Equation 9 puts the budget for nine tenths coverage at 512 distinct structures per instance against the 27 the loop evaluates at its operating point, so the reading on which the benchmark is eventually exhausted by proposal budget alone does not describe a practical route to exhausting it. That is the sense in which the distinction between budget and reach is a practical one and not only a limiting one.

Table 2: Structural families in CRISIS-60. Each category is realised by several ground-truth forms; H_0 is the incumbent theory whose failure defines the anomaly in that family.

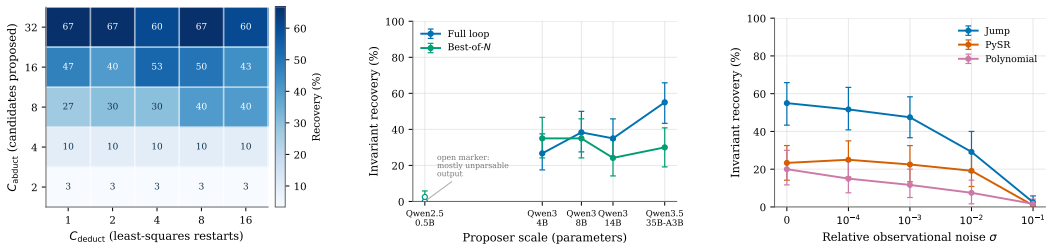
Category	Family	Ground-truth form	Incumbent H_0
Screened Potential	`yukawa`	$A/r^2\,(1+\mu\,r)\,\exp(-\mu\,r)$	c_0/r^2
	`exp_screen`	$A\,\exp(-\mu\,r)/r^2$	c_0/r^2
	`gauss_screen`	$A\,\exp(-(\mu\,r)^2)/r^2$	c_0/r^2
Latent Coordinate	`dark_pole`	$A/r^2 + B/(r-r_0)^2$	c_0/r^2
	`dark_halo`	$A/r^2 + B\,\exp(-(r-r_0)^2)$	c_0/r^2
	`dark_shift`	$A/(r-r_0)^2$	c_0/r^2
Relativistic Bound	`vel_add`	$(u+v)/(1+u\,v/c^2)$	$c_0\,(u+v)$
	`rel_mass`	$m_0/\sqrt{1-(v/c)^2}$	c_0
Saturating Kinetics	`michaelis`	$Vm\,x/(K+x)$	$c_0\,x$
	`hill`	$Vm\,x^2/(K^2+x^2)$	$c_0\,x^2$
	`logistic_sat`	$Vm\,\tanh(x/K)$	$c_0\,x$
In-Basis Control	`cubic_drag`	$-g\,v - b\,v^3$	$c_0\,v$
	`quad_drag`	$-g\,v - b\,v^2$	$c_0\,v$

B BENCHMARK SPECIFICATION

Table 2 lists the structural families. Every environment draws its constants uniformly from family-specific ranges under the environment seed. Training and validation sets contain 80 and 60 points respectively, sampled independently and uniformly from the observable region; the extrapolation set contains 150 points on a uniform grid over the disjoint extrapolation region. Observational noise, where applied, is Gaussian with standard deviation proportional to the signal standard deviation of the environment, so that a single noise level is comparable across environments whose outputs differ by orders of magnitude.

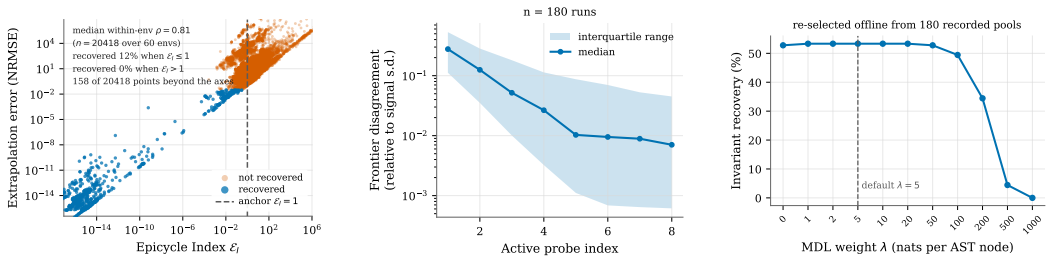
C COMPUTE ALLOCATION, PROPOSER SCALE, AND NOISE

Figure 2 collects the three sweeps: panel 2a reports recovery over a measured grid of the two compute axes, panel 2b against proposer parameter count, and panel 2c against relative observational noise with the full pipeline re-run at each level. The grid is 2 to 32 proposed structures crossed with 1 to 16 optimisation restarts, at 30 instances per cell. The extension of the proposal axis to $C_{\mathrm{abduct}} = 128$ used for the test in Section 7 is a separate block, run on the same environments and seed as the grid and at $C_{\mathrm{deduct}} = 4$ throughout, since the grid establishes that the second axis moves nothing; Appendix H tabulates it against the fitted laws, and it re-runs $C_{\mathrm{abduct}} = 32$ so that the extension and the fit can be compared on a shared cell.



(a) Compute allocation. (b) Proposer scale. (c) Observational noise.

Figure 2: (a) Recovery over the two test-time compute axes, proposals against optimisation restarts. (b) Recovery against proposer parameter count, intervals clustered by environment. (c) Recovery against relative observational noise.



(a) Epicycle Index validation. (b) Measured probe trajectories. (c) MDL weight sensitivity.

Figure 3: (a) The Epicycle Index against measured extrapolation error over every hypothesis scored during the language-model and enumerative searches, with the $\mathcal{E}_I = 1$ anchor. (b) Frontier disagreement against probe index, median and interquartile range over measured runs. (c) Recovery as a function of the description-length weight, re-selected offline from the recorded candidate pools.

The scale sweep in detail. The full loop rises from 26.7% at 4B to 55.0% at 35B while best-of-N stays flat, and the gap between them is -8.3 points at 4B ($p = 0.0414$), +3.3 at 8B, +10.8 at 14B ($p = 0.0106$) and +25.0 at 35B ($p = 2.27 \times 10^{-7}$). We read the trend in the gap because neither curve is monotone on its own, intervals at $n = 60$ per cell are wide, and the 14B point sits below 8B for the full loop. Table 3 also reports the fraction of proposals that never became a scorable hypothesis, which is what makes the low end readable: the 0.5B proposer emits an unparsable expression for 95.2% of its proposals against 0.6% at 4B, so its 0.0% measures instruction-following rather than discovery, and this experiment cannot separate the two.

Table 3: Recovery against proposer scale under an otherwise identical loop, with intervals clustered by environment.

Proposer	**Params**	**Full loop (%)**	**Best-of-N (%)**	**Unparsable (%)**
Qwen2.5 0.5B	0.5B	0.0 [0.0, 0.0]	2.5 [0.0, 5.8]	95.2
Qwen3 4B	4B	26.7 [17.5, 37.5]	35.0 [24.2, 46.7]	0.6
Qwen3 8B	8B	38.3 [27.5, 50.0]	35.0 [24.2, 45.9]	1.4
Qwen3 14B	14B	35.0 [24.2, 45.8]	24.2 [14.2, 34.2]	0.2
Qwen3.5 35B-A3B	35B	55.0 [43.3, 65.9]	30.0 [19.2, 40.8]	0.0

D Recovery by Category

The relativistic families demand an exact rational or square-root form whose competing polynomial fit is excellent throughout the observable regime, and the proposers reliably produce saturating surrogates such as $c_0 \tanh(c_1(u+v)) + c_2 uv$ that are right in-distribution and wrong at the asymptote. The families the loop never solves fail for three different reasons: the halo's unobserved centre lies outside the probe domain and so beyond the reach of Proposition 3; the image family's latent radius enters a numerator and a squared denominator at once; and in the equation of state the loop corrects the ideal gas law by an excluded volume of the wrong sign and never finds the attraction term. Each is a failure to write the structure down rather than a failure to choose it, which is what a coverage of exactly zero means.

E Contamination Probe

Each proposer was asked to state the corrected law for every environment with no measurements supplied, given only the incumbent theory and the permitted grammar. Each of its samples is then scored through the same pipeline used everywhere else in this study: parsed, its free constants fitted to the environment's observable data, and evaluated on the extrapolation split under the same $R^2_{\text{ood}} > 0.999$ criterion. A hit therefore means the model produced the correct structure without seeing any data, which is the quantity the contamination concern is about. We deliberately do not

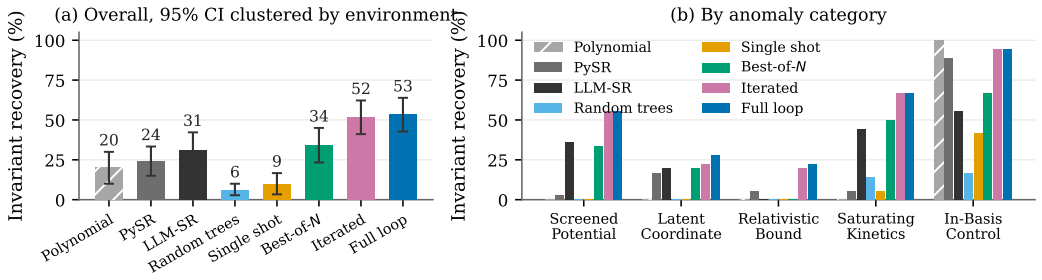


Figure 4: Recovery on CRISIS-60. Intervals are 95% cluster bootstrap over environments. The in-basis control separates methods that fit from methods that extend the basis. Panel (a) presents Table 1 graphically and panel (b) presents Table 4.

Table 4: Recovery by category; the bracketed figure is the number of instances recovered, out of 36 in each cell. Figure 4 (b) presents the same data graphically.

Method	**Screened**	**Latent**	**Relativ.**	**Saturat.**	**In-basis**
Polynomial epicycles	0.0 (0)	0.0 (0)	0.0 (0)	0.0 (0)	100.0 (36)
PySR (genetic SR)	2.8 (1)	16.7 (6)	5.6 (2)	5.6 (2)	88.9 (32)
PySR (5× budget)	0.0 (0)	0.0 (0)	8.3 (1)	16.7 (2)	100.0 (12)
Random trees + identical search	0.0 (0)	0.0 (0)	0.0 (0)	13.9 (5)	16.7 (6)
Proposer, single shot	0.0 (0)	0.0 (0)	0.0 (0)	5.6 (2)	41.7 (15)
Proposer, best-of-N	33.3 (12)	19.4 (7)	0.0 (0)	50.0 (18)	66.7 (24)
LLM-SR (Shojaee et al., 2025)	36.1 (13)	19.4 (7)	0.0 (0)	44.4 (16)	55.6 (20)
LLM-SR (5× budget)	75.0 (9)	33.3 (4)	8.3 (1)	66.7 (8)	75.0 (9)
Proposer + iteration	55.6 (20)	22.2 (8)	19.4 (7)	66.7 (24)	94.4 (34)
Proposer + iteration + probes	55.6 (20)	27.8 (10)	22.2 (8)	66.7 (24)	94.4 (34)

score by keyword presence: a test for a minus sign or a division bar is satisfied by nearly any algebraic output and reports contamination that is not there.

Table 5 gives the result. For the proposer used throughout the paper, zero-shot recovery is 5.0% overall, confined to the screened category at 25.0% and zero elsewhere, so the reported recovery on the remaining four categories cannot be attributed to retrieval. One comparison against an earlier version of this suite deserves emphasis, because it is what the probe is for. That version contained a latent family whose law was exactly $c_0/(r + c_1)^2$, a natural generic guess, and the smaller 4B model named it correctly 66.7% of the time with no data at all. It was replaced by the family whose latent constant enters both a numerator and a squared denominator, which no single-offset template can absorb, and the same probe now recovers 0.0% of that category. Recovery on the latent category fell correspondingly, which is the substitution working rather than the benchmark becoming arbitrarily harder. The probe establishes an upper bound on retrieval, not its absence.

Table 5: Zero-shot contamination probe: percentage of environments a model recovers with no data supplied, scored by the same $R^2_{\text{ood}} > 0.999$ criterion as the main results. Higher values indicate more scope for retrieval rather than discovery.

Category	`Qwen3-4B-Instruct-2507`	`Qwen3.5-35B-A3B`
Screened Potential	0	25
Latent Coordinate	0	0
Relativistic Bound	0	0
Saturating Kinetics	0	0
In-Basis Control	0	0

F THE DISCOVERY LOOP IN FULL

Algorithm 1 The instrumented discovery loop

Require: Observations $\mathcal{D}$, simulator $\mathcal{S}$, incumbent H_0, budgets $C_{\text{abduct}}, C_{\text{deduct}}$, weight λ
1: $\mathcal{P} \leftarrow \{H_0\}$ fitted on $\mathcal{D}$
2: **while** proposal budget remains **do**
3: Sample structures $\{H'_j\} \sim \pi(\cdot \mid \mathcal{D},\ \text{ParetoFront}(\mathcal{P}),\ \text{residuals of best})$
4: Fit each H'_j by multi-restart least squares (C_{deduct} restarts); $\mathcal{P} \leftarrow \mathcal{P} \cup \{H'_j\}$
5: **if** the frontier holds two or more hypotheses **then**
6: $\mathbf{x}^* \leftarrow \arg\max_{\mathbf{x} \in \mathcal{X}_{\text{obs}}} \|H_1(\mathbf{x}) - H_2(\mathbf{x})\|^2$; append $(\mathbf{x}^*, \mathcal{S}.\text{probe}(\mathbf{x}^*))$ to $\mathcal{D}$; refit the frontier
7: **end if**
8: **end while**
9: Refit every $H \in \mathcal{P}$ on the final $\mathcal{D}$
10: **return** $\arg\min_{H \in \mathcal{P}}\ n \log(1 - R^2_{\text{val}}(H)) + \lambda\, \text{Comp}(H)$

G Component Ablations

Table 6: Component ablations, all at matched proposal budget $C_{\text{abduct}} = 32$. LLM calls are per instance.

Variant	Recovery (%)	Δ	Median $\mathcal{E}_I$	Probes	LLM calls
Proposer + iteration + probes	53.3 [42.8, 63.9]	–	0.01	6.8	7
w/o active probes	51.7 [41.1, 62.2]	-1.7	0.03	0.0	7
w/o Occam penalty ($\lambda = 0$)	52.8 [42.2, 63.3]	-0.6	0.02	6.8	7
Proposer, best-of-N	33.9 [23.3, 45.0]	-19.4	0.31	0.0	4
Random trees + identical search	6.1 [2.8, 10.0]	-47.2	1.23	7.3	0

H Coverage in Full

Table 7: Recovery split into proposal coverage and selection efficiency, over 180 instances, with the intervals Table 1 omits. Coverage is the fraction of instances whose recorded candidate pool contains a hypothesis meeting the recovery criterion, and is what an oracle selector would reach from the same proposals; the oracle gap is what any better selection rule has left to win. Intervals are 95% cluster bootstrap over environments.

Configuration	Coverage (%)	Recovery (%)	Selection eff. (%)	Oracle gap	Pool
Proposer, single shot	9.4 [3.3, 16.7]	9.4	100.0	0.0 [0.0, 0.0]	2
Proposer, best-of-N	33.9 [23.3, 45.0]	33.9	100.0	0.0 [0.0, 0.0]	32
Proposer + iteration	52.8 [41.7, 63.3]	51.7	97.9	1.1 [0.0, 2.8]	29
Proposer + iteration + probes	54.4 [43.9, 65.0]	53.3	98.0	1.1 [0.0, 2.8]	28
w/o Occam penalty ($\lambda = 0$)	54.4 [43.9, 65.0]	52.8	96.9	1.7 [0.0, 3.9]	28
Random trees + identical search	6.1 [2.8, 10.0]	6.1	100.0	0.0 [0.0, 0.0]	25

Table 8 gives proposal coverage and recovery for every structural family under the full loop. The gap between the two columns is a selection failure and their common level is a proposal failure, and the table shows that the second dominates: families sit near the diagonal almost everywhere, and the difficulty of the suite is concentrated in a few families where both columns are zero rather than spread across families where coverage exceeds recovery.

The identifiability that limits this is a property of the fitted family over this range rather than of our sample size, which a curve with no data in it shows most cleanly. Generating coverage from known parameters, sampling it without noise at the same budgets and refitting reproduces every fitted budget to within 0.22 points at $R^2 = 0.9999$, while returning a reachable fraction of 96.0% against a true 80.0% and missing $C_{\text{abduct}} = 128$ by +4.6 points. An asymptote is not determined by budgets that have not begun to approach it, however well they are fitted.

Table 8: Proposal coverage and recovery by structural family under the full loop. A family whose two columns agree is limited by what is proposed; a family whose coverage exceeds its recovery is limited by selection.

Category	Family	Vars	Coverage (%)	Recovery (%)
Screened Potential	dipole_field	2	11.1	11.1
	exp_screen	1	100.0	100.0
	gauss_screen	1	44.4	44.4
	yukawa	1	77.8	66.7
Latent Coordinate	dark_halo	1	0.0	0.0
	dark_pole	1	83.3	83.3
	image_sphere	1	0.0	0.0
Relativistic Bound	rel_mass	1	22.2	16.7
	vel_add	2	27.8	27.8
Saturating Kinetics	hill	1	77.8	77.8
	logistic_sat	1	88.9	88.9
	michaelis	1	100.0	100.0
	van_der_waals	3	0.0	0.0
In-Basis Control	cubic_drag	1	88.9	88.9
	quad_drag	1	100.0	100.0

The extension block re-runs $C_{\mathrm{abduct}} = 32$, which is the only reproduction check a language-model arm admits, since the arm is not bit-reproducible across hardware or library releases: on the 30 environments the two blocks share it returns 63.3% against the grid's 63.3% on different hardware and a later release. Of the instances the smaller budget missed, 16 of 30 were covered at 128 and 0 went the other way, so the aggregate movement is not an artefact of a few instances changing in both directions.

Table 9 gives the measured coverage curve along the proposal axis together with both fitted laws. The two rightmost columns are predictions in the rows whose source is given as held out, which were not used in either fit. Both fits are least squares on the measured proportions over a grid in $\log a$ and $\log b$, with the reachable fraction searched on a grid of its own; the objective is flat enough that a gradient method is sensitive to its starting point, and a grid at this size is exhaustive.

Comparing a held-out proportion against those curves treats the curve as known, and it is not. Resampling the 60 environments with replacement and refitting inside each resample places the saturating law's asymptote anywhere from 40.0 to 94.0%, so its parameters are barely determined by the range they are fitted over. Splitting the environments rather than resampling them makes the consequence concrete: the two halves return shape parameters differing by a factor of 18 and then contradict each other at the held-out budget, one measurement falling between its own two predictions and the other lying above both of its own. The synthetic curve above reaches the same conclusion with no data in it. A saturation law fitted over budgets that have not begun to saturate describes those budgets and predicts nothing beyond them, which is why Section 7 reports the measured response of coverage to budget and stops there.

A selection efficiency near one would be uninformative if a covered pool typically held a single hypothesis meeting the criterion, since any rule that noticed it would then score perfectly and the finding would be about the benchmark rather than about the rule. It does not: under the full loop a covered pool holds a median of 4 qualifying candidates and as many as 12, which is 17.9% of a median pool of 25, and 12.2% of covered instances hold exactly one. The rule is choosing, and on the 2 instances where it chose wrongly the pool held at most 2 qualifying candidates against that median.

Computing this for another system takes one line inside its evaluation loop. Whatever criterion decides success, record for each instance whether any candidate the search scored met it, alongside whether the candidate it returned did. The ratio of the two is selection efficiency and their difference is the headroom a better selection rule has; a system whose efficiency is near one cannot be improved by a better criterion, a better parsimony weight, or more careful stopping, and a system whose efficiency is low should not have effort spent on its proposal distribution. Two caveats travel with it. The oracle is over the pool the search realised and not over the hypothesis space, so a high efficiency says nothing about what a different proposer might have reached. And the verdict is regime-specific: the

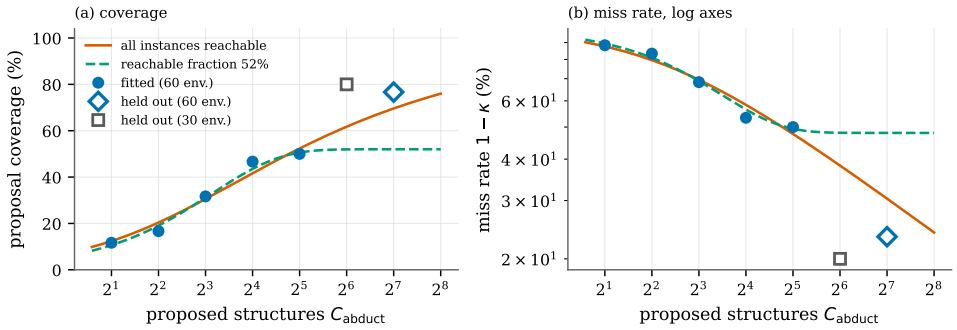


Figure 5: Coverage against proposal budget at fixed C_{deduct}, with the two candidate laws fitted on the budgets at or below 32 and the larger ones measured afterwards and held out of both fits. The intermediate budget was run on half the environments the others were and is drawn as its own series rather than pooled with them. Panel (b) redraws the same two laws as miss rates on logarithmic axes, where Proposition 5 says they differ in kind: a straight line of slope $-a$ against a curve flattening onto $1 - f$. Over the fitted range the two are barely separated, and both held-out measurements fall above both of them.

Table 9: Coverage along the proposal axis at fixed C_{deduct}, against the two candidate laws. The laws are fitted on the rows sourced from the grid; the held-out rows and the extension block's re-run of a fitted budget were not used in either fit.

C_{abduct}	Source	Coverage (%)	Recovery (%)	All reachable	Reachable fraction
2	grid / re-run	11.7	11.7	12.4	10.6
2	extension	20.0	20.0		
4	grid / re-run	16.7	16.7	20.4	19.1
4	extension	23.3	23.3		
8	grid / re-run	31.7	31.7	30.5	31.1
8	extension	33.3	33.3		
16	grid / re-run	46.7	46.7	41.6	43.5
16	extension	40.0	40.0		
32	grid / re-run	50.0	48.3	52.3	50.6
32	extension	50.0	46.7		
64	held out	80.0	80.0	61.7	52.0
128	held out	76.7	75.0	69.6	52.0

same loop is proposal-limited without observational noise and selection-limited with it, which is visible in Section 7 and would be invisible from either factor alone.

Both factors are defined against the recovery criterion, so Table 10 recomputes them at other thresholds from the same recorded pools, which needs no re-run because each pool carries every candidate's extrapolation score. Coverage moves as the threshold tightens, from a few points above its value at $\tau = 0.999$ to a few points below, and selection efficiency does not: it stays above ninety per cent for every configuration across four orders of magnitude in $1 - \tau$. The finding that selection is nearly saturated is therefore a property of the loop in this regime rather than of the particular criterion chosen.

I What the Complexity Charge Detects

Equation 3 is linear in λ, so for the recovering hypothesis H_+ and the incorrect H_- that the rule ranks highest at the operating point, the weight at which their order reverses is

$$\lambda^\star = \frac{n \log\big((1 - R^2_{\text{val}}(H_-))/(1 - R^2_{\text{val}}(H_+))\big)}{\big[\text{Comp}(H_+) - \text{Comp}(H_-)\big]_+}. \tag{11}$$

Table 10: Coverage, recovery and selection efficiency recomputed at four recovery thresholds τ from the same recorded pools. Selection is re-applied at each threshold rather than read off, since the hypothesis the objective returns does not depend on τ.

Configuration	$\tau = 0.99$			$\tau = 0.999$			$\tau = 0.9999$			$\tau = 0.99999$		
	Cov.	Rec.	Eff.	Cov.	Rec.	Eff.	Cov.	Rec.	Eff.	Cov.	Rec.	Eff.
Proposer, best-of-N	36.7	36.1	98.5	33.9	33.9	100.0	32.2	32.2	100.0	30.6	30.6	100.0
Proposer + iteration	60.6	57.2	94.5	52.8	51.7	97.9	48.9	48.3	98.9	47.8	47.2	98.8
Proposer + iteration + probes	59.4	57.8	97.2	54.4	53.3	98.0	51.7	50.6	97.8	50.0	48.9	97.8

Over the 98 covered instances the objective favours a recovering hypothesis on 98.0% of them, by a median of 1493 nats and 858 at the tenth percentile. On the 23 instances where the recovering hypothesis is also the larger, the median $\lambda^\star$ is 462 and the tenth percentile 64. The offline sweep of Figure 3c binds near that tenth percentile rather than at the median, which is the same statement arrived at without the algebra.

The offline penalty sweep of Section 7 is reported because the opposite result is easy to manufacture. Against the criterion we used earlier, linear in $1 - R^2_{\mathrm{val}}$ rather than logarithmic, the same ablation would have shown a large effect that was purely the penalty dominating a badly scaled fit term, and nothing about parsimony.

Removing the complexity ratio while keeping the anchor leaves the rule $\rho_E > 1$, whose verdict differs on 42.2% of 28156 scored hypotheses; among the 11853 that only the complexity charge condemns, 0.0% had recovered a law, so the charge never condemned a correct hypothesis here. The case error alone cannot resolve at any threshold is a pair of hypotheses that extrapolate identically and differ in size: in one relativistic environment a 10-node and a 158-node hypothesis extrapolate within one percent of each other and the index separates them at 0.31 against 5.02. Scoring both as detectors of the Ptolemaic failure mode, defined independently as fitting the observable regime above $R^2 = 0.99$ while extrapolating below $R^2 = 0.9$, the anchored index attains precision 45.2% and recall 69.5% against 35.7% and 20.2% for the error-only rule at the same anchor, so charging description length improves both.

The bound on that claim is the comparison against a tuned cut point. A threshold on extrapolation error selected with hindsight over these same hypotheses reaches $F_1 = 62.3$ against 54.8 for the untuned anchor, so the index is not the most accurate detector obtainable from these two quantities. Its claim is narrower: the threshold is fixed by the construction rather than fitted, and it means the same thing in every environment, which a cut point selected on extrapolation error does not, since that error is exactly the quantity the detector is trying to anticipate.

J How LLM-SR Was Run

The re-run that supplies the coverage decomposition for this baseline records every program its evaluator scored, which the original sweep did not. Recording changes nothing else in the loop, and the two runs agree on 53/60 of the instances they share at the same seed, so the decomposition is computed on a run that reproduces the one reported in Table 1 rather than on a different configuration.

LLM-SR is the closest published system to the loop instrumented here, so the comparison is only informative if the search being compared is theirs and the conditions are not tilted. We record exactly what was done.

The search is theirs, executed from the reference implementation at a pinned commit, with its experience buffer, island resets, prompt construction, and program-space representation unmodified. Only the two plug points its own configuration object declares were supplied: the language model and the execution sandbox. The proposer is the same Qwen3.5-35B-A3B used by every other language-model arm at the same temperature and top-p, so the comparison isolates the search loop rather than the serving stack or the model. The sandbox executes in process under an interval timer instead of forking a child per candidate, which is an execution mechanism rather than a search decision and was necessary because this driver holds a multi-billion-parameter model resident on the same device.

Each environment is given a specification synthesised in their format, seeded from the same incumbent theory H_0 that the instrumented loop starts from, restricted to the same primitive grammar, and allocated the same maximum number of free constants that `expr.MAX_PARAMS` imposes on every other proposer. An included test asserts that the governing law appears nowhere in the specification and that the seed program is H_0.

Their evolved artefact is a Python function, whereas every other hypothesis here is a closed form. Rather than compare across that gap, the evolved function is executed a second time with SymPy symbols substituted for the arrays, which reconstructs the expression the program computes exactly for straight-line arithmetic and refuses control flow or array reductions, both of which lie outside the grammar the other arms are held to. The recovered form is then scored by the same function that scores every hypothesis in this study.

Three choices favour the baseline and we state them rather than leave them to be discovered. Its parameters are refitted by the multi-restart least-squares fitter used everywhere else instead of the single BFGS pass from an all-ones initialisation that its own specification performs. Its final hypothesis is selected from every program it evaluated rather than from the last one, which is the treatment PySR's Pareto front also receives. And its instruction prompt carries an added output-format clause: their sampler targets a base-model completion server whose continuation is bare code, while the proposer used here is instruction-tuned, and without the clause it answers in prose and runs to the token cap without emitting a function. That mismatch was measured, not assumed: one generation in eight yielded an executable program without the clause against essentially all of them with it, and the wall time per environment fell by an order of magnitude at the same time. Scoring the baseline on a prompt-format incompatibility would have measured something other than the quality of its search.

The interface can be audited from the logs rather than taken on trust. Across every logged run, 97.8% of the programs the baseline evaluated yielded a closed form the shared scorer accepted, so what the recovery figure reflects is its search and not attrition between its representation and ours.

K Implementation Details

The study reports 13 paired tests. Bonferroni over all of them puts the threshold at $0.05/13$, which the proposer against the enumerative control, iteration over best-of-N, and the loop against polynomial regression and PySR all survive by many orders of magnitude. The 4B and 14B points of the scale curve do not, and are read as a trend rather than as individually significant. No correction turns a null result into a positive one, so the two components reported as non-contributing are unaffected; for those the relevant quantity is the equivalence bound of Section 7 rather than a p-value.

Table 11: Configuration used for all reported runs.

Setting	**Value**
Proposer (default)	Qwen3.5-35B-A3B, bfloat16, single GPU
Sampling	temperature 0.9, top-p 0.95, $\leq$ 200 new tokens
C_{abduct}	32 distinct candidate structures
C_{deduct}	6 least-squares restarts, $\leq$ 2000 evaluations each
Branching	4 candidates per proposal round
MDL weight λ	5.0 nats per AST node
Optimiser	`scipy.optimize.least_squares`, trust-region reflective
Active probes	$\leq$ 15, grid of 400 candidate points in the observable region
Train / validation / OOD	80 / 60 / 150 points
Recovery criterion	$R^2_{\text{ood}} > 0.999$
Seeds	42, 123, 456
PySR	`niterations` 40, 15 populations of 33, `maxsize` 30
PySR (high budget)	`niterations` 200, 25 populations of 50
Polynomial	degrees 1 to 6, order selected on validation

The proposer prompt supplies the incumbent law and its fit quality, a sample of the observations spread across the range of the first coordinate, the current Pareto frontier in structural form, and a description of the leading hypothesis's residual pattern. It states the permitted grammar in full and instructs that unknown constants be written symbolically, including constants standing for quantities

that were never measured. It never names a primitive associated with any particular ground-truth family. The enumerative control receives the same grammar and the same search loop and differs only in that its proposals are random expression trees drawn without reference to the data.

Two details of the frontier presentation proved to matter enough to record. Hypotheses are shown in structural form with constants still symbolic, because showing fitted numerical values contradicts the instruction to write constants symbolically and anchors the proposer on particular magnitudes. Fit quality is reported as unexplained variance rather than R^2, because once a fit is good R^2 rounds to 1.000 for both a near-miss and an exact recovery, concealing the very signal the proposer needs. A third detail belongs to verification rather than presentation: the multiple restarts in the fitting stage are necessary rather than cosmetic, since these objectives are non-convex and a single fixed initialisation systematically favours whichever functional form happens to be well conditioned at that point. In the selection criterion of Equation 3, driving the residual from 10^{-3} to 10^{-12} is worth about 21 nats per observation, so a correct but larger hypothesis can pay for its additional nodes.

Given a seed, environment construction, data sampling, and the enumerative arm reproduce bit-identically; we verified this by re-running instances and comparing the returned expression, its extrapolation score, and its probe count. The language-model arm seeds the sampler per call but is not bit-reproducible across hardware or library versions, which is why every reported figure aggregates over 3 seeds and 60 environments with intervals clustered by environment rather than resting on a single run.

L Representative Hypotheses

Table 12 lists, for each category, the hypothesis with the largest and the smallest Epicycle Index among those fitting validation data at $R^2 > 0.999$. The selection is mechanical rather than curated, so the examples are reproducible from the logs and are not chosen to flatter any method. The pattern they display is the one the index is designed to detect: expressions that fit the observable regime to five decimal places and then diverge by tens of orders of magnitude outside it. The paper's own loop appears among the failures, which is the expected consequence of scoring every method by the same criterion.

Table 12: Extreme hypotheses by Epicycle Index within each category, among those fitting validation data at $R^2 > 0.999$.

Category		Method	Hypothesis	R^2_{val}	R^2_{ood}
Screened Potential	worst	poly	`0.0075696923077173709*r**6 + 0.017235...`	0.99997	-1.56×10^{14}
	best	jump	`7.50034699067495*exp(-0.0159148356564...`	1.00000	1.00
Latent Coordinate	worst	jump	`10.6003986116116*exp(-0.1040186709543...`	1.00000	-5.38×10^{224}
	best	jump	`9.50866908572405/(r - 5.4361730099604...`	1.00000	1.00
Relativistic Bound	worst	pysr	`exp((v - sin(sin(v)))/(sin(v*sin(v)/(...`	0.99997	-1.63
	best	jump	`1.0*(u + v)/(0.0880010071340365*u*v +...`	1.00000	1.00
Saturating Kinetics	worst	poly	`-0.053229560971468276*x**6 + 0.515590...`	1.00000	-1.89×10^{15}
	best	llm_bestof	`4.17067723191746*tanh(0.5842863346361...`	1.00000	1.00
In-Basis Control	worst	llm_bestof	`-1.17373257835756*v/(1 - 0.2286164291...`	0.99980	-1.01×10^{3}
	best	llm_iter	`-2.2938669126812e-17*sqrt(v) - 0.2303...`	1.00000	1.00