FALSIFY, REPAIR, REPEAT: AUDITABLE LOCAL REPAIR OF TYPED DYNAMICAL EQUATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Equation discovery can fit observed trajectories while retaining the wrong term, regime, or delay. We introduce Falsify–Repair–Repeat (FRR), a bounded repair procedure for auditing and locally repairing executable dynamical laws. Statistical verifiers turn failed tests into typed witnesses, a constrained policy selects grammar-valid edits, and numerical tools retain authority over refitting, acceptance, replay, and abstention. The selector is replaceable; accepting a structural edit requires executable validation and replay of accumulated tests. This separation lets deterministic and learned policies share comparable evidence and outcome records without exposing unrestricted control over the model. We characterize the finite repair basin, conditional recovery, replay preservation, budgeted termination, and an optional error-spending construction under conditionally valid tests. Across one-edit, compound-error, held-out, and null studies, typed editing supplies the principal behavioral and structural gains, while replay and abstention expose preservation–performance and risk–coverage trade-offs. Deterministic FRR reaches 85.6% semantic success on the one-edit benchmark; paired intervals support behavioral improvement through AST distance three, and independent ordinary-ODE rollouts improve from unseen initial conditions. A natural-candidate stress test further separates global basin entry from local authorization. The proposal-agnostic interface therefore turns repair within a bounded typed neighborhood into an auditable repair process. Here, "auditable" means that each accepted edit is traceable to a typed witness and an executable acceptance/replay record; it does not claim formal correctness.

1 INTRODUCTION

Symbolic dynamical models serve as both predictors and scientific objects. Early systems reconstructed dynamics or free-form laws directly from observations (Bongard & Lipson, 2007; Schmidt & Lipson, 2009), and modern sparse, neural, and sequence methods have broadened this program (Brunton et al., 2016; Udrescu & Tegmark, 2020; Qian et al., 2022; d'Ascoli et al., 2024; Petersen et al., 2021; La Cava et al., 2021). Yet low interpolation error does not establish a defensible structure. A fitted equation may omit an interaction, introduce a noise-compensating term, merge regimes, or absorb a delay into biased coefficients. Discovery produces an expression but rarely identifies which component the data contradict, which local correction the evidence supports, or whether that correction invalidates an earlier finding (Brunton & Kutz, 2019).

This downstream problem resembles counterexample-guided and syntax-guided synthesis, where a learner proposes a program and a verifier guides the next proposal (Solar-Lezama et al., 2006; Jha & Seshia, 2014; Alur et al., 2013; Gulwani et al., 2017). Dynamical data alter that interaction: a noisy trajectory is not a logical counterexample, a calibrated statistic is not a proof obligation, and a repair that explains one residual pattern may revive an earlier failure. We study executable equations from a known typed grammar when the target is reachable by a short edit path; otherwise the system abstains and returns control to discovery.

Falsify–Repair–Repeat (FRR) formulates this setting as a bounded, tool-mediated repair procedure. Deterministic verifiers return a failure margin, localized numerical evidence, and grammar-valid one-edit actions. A replaceable policy—deterministic, cross-validated, or language-model based—selects an action, while numerical code performs coefficient fitting, validation, and acceptance

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.

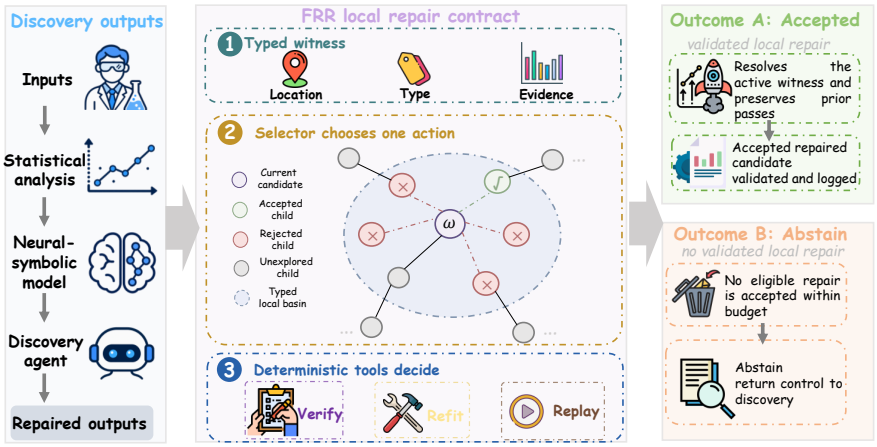

Figure 1: FRR audit contract from discovery outputs to accepted local repair or abstention. Typed witnesses localize evidence, the selector chooses a grammar-valid action, and deterministic verification, refitting, and replay determine the outcome.

and the controller rechecks accumulated tests after each edit. Selector choice can therefore vary without changing the executable repair and acceptance rules. Figure 1 summarizes the auditable contract: discovery outputs enter typed diagnosis, a selector chooses one grammar-valid action, and deterministic verification, refitting, and replay lead either to an accepted repair or to abstention and return to discovery.

Local repair complements rather than replaces discovery. Candidate equations may come from scientists, sparse regression, neural symbolic regression, or discovery agents, but FRR assumes that the proposal lies within a typed repair basin. It targets settings where the variables and broad equation family are known, while specific interactions, regimes, or delays remain uncertain. As scientific agents increasingly combine reasoning with external tools (Yao et al., 2023; Schick et al., 2023; Bran et al., 2024; Boiko et al., 2023), FRR provides a common procedure for evaluating candidate equations from different sources.

End-to-end accuracy cannot attribute gains to localization, constrained editing, search, or replay, and exact recovery need not imply behavioral improvement. We therefore compare matched policies on identical tasks and evaluate semantic, structural, computational, and preservation outcomes. The study combines a complete 2^4 component intervention with held-out systems, OOD shifts, compound errors, selective risk–coverage analysis, null inputs, and natural discovery candidates. This design yields four concrete contributions:

- Local equation repair is formulated as constrained statistical synthesis: a proposal-agnostic interface separates typed diagnosis, finite structural choice, numerical refitting, and replay, while deterministic and language-model policies share the same action space.
- Formal interface properties connect witness recall, grammar coverage, finite-sample separation, replay, and a run-level error-spending construction to conditional recoverability.
- Two evaluation protocols contribute 13,439 dynamical run records (5,168 confirmatory, 3,728 calibration, and 4,543 test runs): a complete 2^4 mechanism intervention and a compound-error recovery-radius study with replay and selective-deployment tests.
- Transfer and scope are evaluated on seven held-out ODEBench systems, distant SINDy candidates, four OOD shifts, two null controls, and unseen-initial-condition rollouts.

2 Related Work

Equation discovery. SINDy and PDE-FIND estimate sparse coefficients over supplied libraries (Brunton et al., 2016; Rudy et al., 2017); SINDy-PI, weak SINDy, and E-SINDy improve coverage or noise robustness through implicit, variational, or ensemble estimation (Kaheman et al., 2020; Messenger & Bortz, 2021; Fasel et al., 2022). D-CODE replaces observed derivatives with a

variational criterion (Qian et al., 2022), while PySINDy and PySR provide practical sparse and evolutionary search systems (de Silva et al., 2020; Cranmer, 2023). Neural equation learners and representation-based methods extract symbolic structure from learned models (Martius & Lampert, 2016; Cranmer et al., 2020; Champion et al., 2019); Neural ODEs and universal differential equations instead hybridize continuous dynamics with flexible function approximators (Chen et al., 2018; Rackauckas et al., 2020). ODEFormer maps trajectories to symbolic systems, and deep symbolic regressors learn distributions over expressions (d'Ascoli et al., 2024; Petersen et al., 2021; Holt et al., 2023). RAG-SR augments generation with retrieved expressions, while reactive SINDy and KinFormer impose reaction structure (Zhang et al., 2025; Hoffmann et al., 2019; Chen et al., 2025). These methods generate or re-estimate an expression; we instead ask which local change its residual evidence authorizes. Direct semantic modeling avoids a closed-form ODE (Kacprzyk & van der Schaar, 2025); FRR retains one because the expression itself is the object under audit.

Candidate refinement and agentic search. LLM-SR and RESTART refine symbolic candidates with language models or residual structure (Shojaee et al., 2025; Li & Pan, 2026). OPRO optimizes from candidate histories, while AgentRefine and Reflexion use environmental or verbal feedback to improve subsequent actions (Yang et al., 2024; Fu et al., 2025; Shinn et al., 2023). Such iterative refinement establishes that feedback can guide search, but a final expression alone does not reveal which test authorized each structural change. In FRR, controlled model arms return one identifier from a verifier-generated action list; numerical code performs fitting and acceptance. This separation makes feedback, edit constraints, selection, and replay experimentally identifiable and permits deterministic execution without an API.

Counterexample-guided synthesis. Classical CEGIS alternates synthesis with exact or decidable verification (Solar-Lezama et al., 2006; Jha & Seshia, 2014). Test-driven program repair likewise searches constrained patches using observed failures, semantic constraints, or learned ranking (Le Goues et al., 2012; Nguyen et al., 2013; Mechtaev et al., 2015; Long & Rinard, 2016). Passing a finite test suite, however, need not imply a correct patch (Qi et al., 2015). Equation repair faces the analogous statistical problem: a candidate can pass every observed test while remaining structurally wrong, and noise can reject a correct structure. We call a localized failed test a *statistical witness*, reserve *exact recovery* for AST equality, and measure behavioral agreement separately. Replay imports the CEGIS idea of accumulating evidence, but here preserves only observed test outcomes rather than logical validity. Unlike free-form refinement, FRR records each verifier result, admissible action, refit, rejection, and replay transition.

3 FALSIFY–REPAIR–REPEAT

At round k, FRR refits h_k and runs untested verifiers until one fails. The resulting witness localizes a term, regime, or lag and enumerates admissible one-edit successors, converting open-ended search into a finite comparison. A deterministic or learned policy ranks locally refitted candidates, but acceptance requires resolving the witness without a hard replay regression. The loop stops when all tests pass, no action remains, or the budget is exhausted: the policy orders admissible choices, while numerical tests and accumulated evidence decide.

3.1 VERIFIER THRESHOLDS AND TYPED WITNESSES

Each verifier $t \in \mathcal{T}$ returns a statistic $z_t(h; D)$, a status in $\{\text{PASS}, \text{INCONCLUSIVE}, \text{FAIL}, \text{INVALID}\}$, and a frozen threshold γ_t. The registry covers derivative fit, residual–feature correlation, coefficient relevance, residual whiteness, delay profiles, regime occupancy, and regime boundaries, drawing on residual model validation in system identification (Ljung, 1999; Ljung & Box, 1978). Thresholds are marginal 0.99 quantiles from correct-model development runs and remain fixed during evaluation; they do not control the error probability of a multi-round run.

Because the statistics have different scales and directions, define the oriented violation

$$v_t(h; D) = \left[\frac{s_t\{z_t(h; D) - \gamma_t\}}{q_t} \right]_+, \qquad [u]_+ = \max(u, 0), \tag{1}$$

where $s_t \in \{-1, +1\}$ orients failure toward positive values and $q_t > 0$ is a development-set scale. Zero denotes a passed finite test, not structural correctness.

A failed verifier emits

$$r_t(h) = \big(t,\ z_t - \gamma_t,\ L_t,\ W_t,\ A_t(h)\big), \tag{2}$$

where L_t localizes equations, regimes, or lags; W_t stores numerical evidence; and $A_t(h)$ enumerates grammar-valid actions. An action adds, removes, or replaces one term, delay, or threshold, tying every accepted change to a failed test.

3.2 Finite repair, selection, and replay

Let $h \oplus a$ apply action a. Every candidate belongs to the one-edit neighborhood

$$\mathcal{N}_1(h; r) = \{h \oplus a : a \in A(r),\ d_{\mathrm{AST}}(h, h \oplus a) = 1\}. \tag{3}$$

For target $h^\star$, define the admissible K-step repair basin

$$\mathcal{B}_K(h^\star) = \{h_0 : \exists\, h_0 \to \cdots \to h_m = h^\star,\ m \le K,\ h_{i+1} \in \mathcal{N}_1(h_i; r_i)\}. \tag{4}$$

Membership presumes a compatible variable set, equation family, grammar, and witness-supported path; the controlled benchmark samples $\mathcal{B}_1$. After local coefficient refitting, deterministic selection minimizes the frozen validation objective

$$a_k^* \in \arg\min_{a \in A(r_k)} \widehat{R}_{\mathrm{val}}(h_k \oplus a), \tag{5}$$

with ties resolved by a fixed beam order. CV-FRR uses a 70/30 fit/validation split and an eight-member beam. Language-model arms receive the same list and return one identifier; schema validation rejects responses outside $A(r_k)$, so the policy never emits executable code or leaves $\mathcal{N}_1(h_k; r_k)$.

Untested verifiers follow a frozen cost-aware priority based on detection, localization, repairability, novelty, and cost; the registry and ordering formula are in the supplement. Execution permits four repair rounds, eight verifier calls, 64 parameter fits, and one selector decision per round.

Let P_k denote tests that have previously passed. For a refitted proposal h', define the hard-regression count

$$G(h'; P_k) = \sum_{t \in P_k} \mathbf{1}[\sigma_t(h') \in \{\textsc{fail}, \textsc{invalid}\}]\,. \tag{6}$$

An edit is eligible only if it resolves the current witness, satisfies $G(h'; P_k) = 0$, and remains within budget. A PASS→INCONCLUSIVE transition is logged but is not a hard regression. This condition applies to the fitted proposal at acceptance. The implementation then refits coefficients and audits the returned state without rollback; any final regression is reported as a preservation failure, not a replay-stable return.

3.3 Formal properties

The interface separates four events at a required repair step: W, the diagnostic localizes the defect; A, the corrective action is enumerated; S, that action is selected; and Q, refitting and replay accept it. Defining the step-recovery event $E_{\mathrm{rec}} = W \cap A \cap S \cap Q$, the chain rule gives, without an independence assumption,

$$\Pr(E_{\mathrm{rec}}) = \Pr(W)\Pr(A \mid W)\Pr(S \mid A, W)\Pr(Q \mid S, A, W). \tag{7}$$

The factorial intervention targets these stages separately: feedback changes the information available for W, the edit grammar determines A, the policy governs S, and replay contributes to Q.

The decomposition also locates failure: W concerns diagnosis, A grammar coverage, S ranking, and Q numerical acceptance. A stronger selector cannot recover an action the grammar omitted, while a larger grammar can reduce the selection margin. Replay follows these search decisions and contributes preservation rather than discovery. This distinction underlies the factorial attribution.

The margin condition makes this interaction explicit. A larger grammar improves recall but gives validation more near-equivalent candidates to separate; extra search helps only when the correct successor remains distinguishable.

For a finite candidate set $\mathcal{C}_k$, suppose the correct successor h^* has population validation margin

$$\Delta_k = \min_{h \in \mathcal{C}_k \setminus \{h^*\}} \{R(h) - R(h^*)\} > 0 \tag{8}$$

and $\sup_{h \in \mathcal{C}_k} |\widehat{R}(h) - R(h)| \le \epsilon_k$.

Proposition 1 (Finite-step recovery)**.** If h^* is enumerated and $\Delta_k > 2\epsilon_k$, Eq. 5 uniquely selects h^*. Along an admissible path of length K, exact recovery occurs within K rounds if this enumeration and margin condition holds at every step and each selected successor passes refitting and replay.

The edit gate balances abstaining on an incorrect candidate against editing a correct one. A frozen calibration partition fixes this empirical risk–coverage trade-off, following the selective-prediction view of abstention (El-Yaniv & Wiener, 2010; Geifman & El-Yaniv, 2017); unlike conformal risk control (Angelopoulos & Bates, 2023; Bates et al., 2021), the evaluated gate does not provide a distribution-free guarantee. The proof and gate specification are given in the supplement.

An optional run-level error-spending construction allocates predictable levels $\alpha_{k,t}$ satisfying $\sum_{k,t} \alpha_{k,t} \leq \alpha$, related to sequential alpha-investing and online error control (Foster & Stine, 2008; Ramdas et al., 2018). If each verifier p-value is conditionally super-uniform while the current structure is correct, then

Proposition 2 (False-repair control)**.** If an edit requires $p_{k,t} \leq \alpha_{k,t}$, a run beginning from a correct structure satisfies

$$\Pr(\text{at least one false structural edit}) \leq \sum_{k,t} \alpha_{k,t} \leq \alpha. \tag{9}$$

The evaluated marginal thresholds do not instantiate this conditional guarantee; the stopping-time proof is given in the supplement. Independently, every run obeys $(N_R, N_T, N_F, N_A) \preceq (4, 8, 64, 4)$ for repair rounds, verifier calls, parameter fits, and policy decisions. Invalid evaluations remain failures, and the finite budget guarantees termination while preserving the complete evaluation history.

4 Evaluation

Tasks. Each task provides an executable equation h_0, repair data D_{d}, and a sealed audit split D_{a}. The controlled study has 180 paired one-edit tasks; the compound study has 768 tasks spanning three dynamical families, two noise levels, four corruption profiles, and AST distances $\{1, 2, 3, 5\}$. The language-model selector uses a prespecified 200-task subset. External tests cover seven held-out ODEBench systems, four OOD shifts, 112 SINDy candidates, and a participant-level MIMIC-III/VitalDB interface study (Johnson et al., 2016; Moody et al., 2020; Lee et al., 2022). This suite tests attribution, repair radius, transfer, unnecessary intervention, and basin entry.

Methods. Primary contrasts compare Agent-FRR, the deterministic FRR controller, a cross-validated variant (CV-FRR), and its matched no-replay arm. Agent-FRR tests whether a language model can rank audited actions; FRR isolates the verifier–grammar mechanism; CV-FRR tests additional local search; and the no-replay arm identifies the preservation cost of replay. All contrasts share tasks, grammar, thresholds, and budgets. Rewrite, random-selector, influence, refinement, and oracle controls complete the comparison.

Endpoints and statistics. Semantic success requires the final law to pass a derivative-NRMSE audit on the held-out final 30% of each trajectory; fitting, repair, and selection use only the first 70%. Exact recovery requires AST equality, rescue is evaluated on the 63 initially failed tasks, and replay stability excludes a final PASS→FAIL/INVALID transition. We treat semantic success and exact recovery as the primary behavioral and structural endpoints; replay stability and false-repair rate as safety endpoints; verifier calls and fits as efficiency endpoints; and localization measures as diagnostic endpoints. Comparisons are task-paired, with task-cluster bootstrap intervals (Efron & Tibshirani, 1993), Holm correction for prespecified confirmatory families (Holm, 1979), and selective thresholds fixed on calibration data before test evaluation. The supplement specifies the complete protocols and provenance.

5 Results

5.1 Local repair improves nearby equations

On the one-edit benchmark, bounded editing separates local correction from unrestricted rewriting. FRR and CV-FRR reach 85.6% semantic success across 180 matched tasks, compared with 66.1% for free rewriting. FRR also reconstructs 39.4% of the generating structures, and both typed policies rescue 37 of the 63 initially failed candidates. The factorial study below attributes this advantage to the typed repair interface.

Table 1: Performance on the one-edit benchmark.

Method	Semantic [95% CI]↑	Exact↑	Rescue† ↑	Exact† ↑	Stable↑	Loc. F1↑	Ver.↓	Fits↓	API↓
Initial	65.0 [57.8, 71.6]	0.0	–	–	180/180	–	0.0	0.0	0.00
Rewrite	66.1 [58.9, 72.6]	2.8	–	–	174/180	0.55	3.1	9.8	0.91
Typed/no replay	83.9 [77.8, 88.5]	33.3	34/63	17/63	174/180	0.67	4.2	9.6	0.76
Agent-FRR	81.1 [74.8, 86.2]	33.3	29/63	17/63	**180/180**	0.67	4.7	**7.7**	0.76
FRR	**85.6 [79.7, 89.9]**	**39.4**	**37/63**	**28/63**	**180/180**	**0.68**	4.8	8.2	**0.00**
Random-FRR	78.9 [72.4, 84.2]	27.2	25/63	11/63	**180/180**	0.53	5.4	18.3	0.84
Influence	82.2 [76.0, 87.1]	25.6	–	–	175/180	0.67	7.1	24.7	**0.00**
Refinement	70.6 [63.5, 76.7]	21.1	–	–	**180/180**	0.67	3.6	24.0	0.91
Oracle*	100.0 [97.9, 100.0]	100.0	–	–	180/180	–	0.0	1.0	0.00
CV-FRR	**85.6 [79.7, 89.9]**	38.3	**37/63**	**28/63**	179/180	**0.68**	4.7	9.1	**0.00**

Bold denotes the best non-oracle result; * is the oracle bound and † the 63 initially failed tasks.

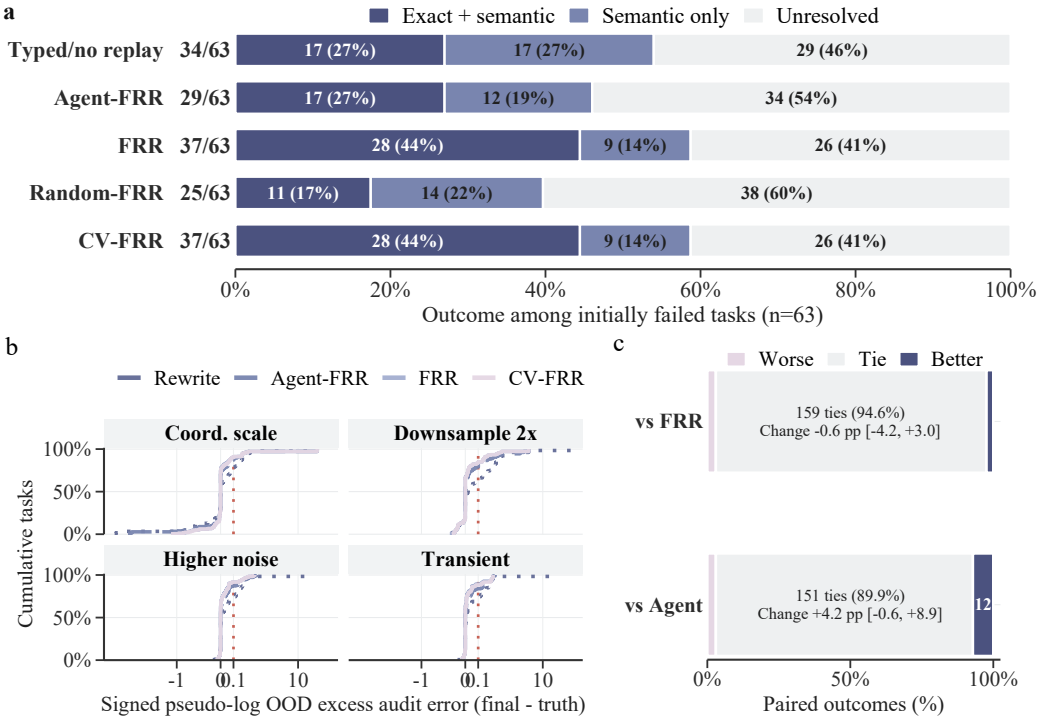

Figure 2: One-edit recovery, OOD robustness, and held-out transfer. Panel b uses a signed pseudo-log scale for the OOD excess audit error; the dotted line marks the primary $+0.10$ tolerance.

Figure 2 separates rescue, OOD robustness, and held-out transfer. Replacing deterministic ranking with Agent-FRR changes only action selection: semantic success shifts from 85.6% to 81.1%, while every run remains replay-stable. Because the witnesses, edits, and numerical checks are shared, the two selectors can be compared directly.

The additional local-search comparisons reveal a different advantage: structural recovery with few fits (supplement, Method identification and search cost). Across fit ceilings of 8–64, FRR retains 39.4% exact recovery. At the eight-fit ceiling it uses 3.80 fits on average, versus 7.87 and 7.76 for validation-ranked and residual-guided search, whose exact recovery is 7.2% and 11.7%. Residual-guided search has higher semantic success (90.0% versus 85.6%); FRR favors generating structure over further reductions in validation error.

5.2 Behavioral gains extend beyond exact recovery

Figure 3a–b shows where the typed neighborhood supports reconstruction. Typed policies recover about 60% of spurious-term cases, whereas wrong-delay cases retain the wrong structure even when

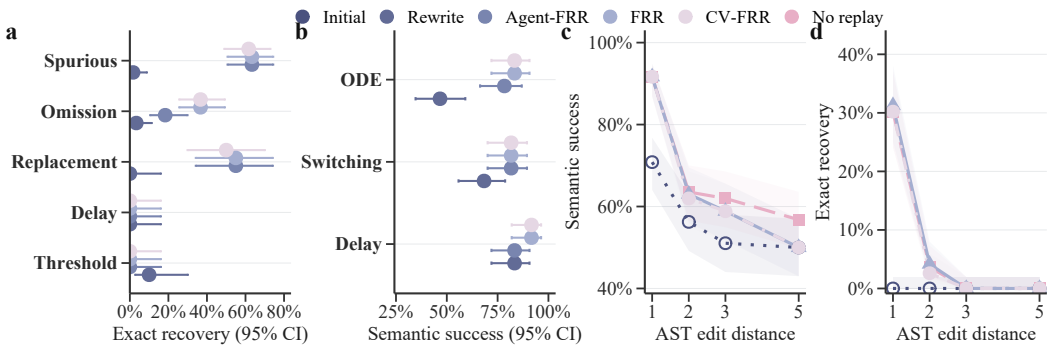

Figure 3: Recovery by defect type and initial AST distance.

Table 2: Compound-error recovery by initial AST distance.

	Initial	FRR			CV-FRR			CV-FRR w/o replay		
Depth	Sem.	Sem.	Δ	Exact	Sem.	Δ	Exact	Sem.	Δ	Exact
1	70.8	**91.7**	+20.8	**31.3**	**91.7**	+20.8	30.2	**91.7**	+20.8	30.2
2	56.3	63.0	+6.8†	**4.2**	62.0	+5.7†	2.6	**63.5**	+7.3†	3.6
3	51.0	58.9	+7.8†	**0.0**	58.9	+7.8†	**0.0**	**62.0**	+10.9†	**0.0**
5	50.0	50.0	+0.0	**0.0**	50.0	+0.0	**0.0**	**56.8**	+6.8	**0.0**
All	57.0	65.9	+8.9	**8.9**	65.6	+8.6	8.2	**68.5**	**+11.5**	8.5

$n = 192$ per depth and method (overall $n = 768$); Δ is gain over the initial candidate; † marks a paired 95% CI excluding zero.

behavioral fit improves. The two endpoints answer different questions: a trajectory can be explained by several equations, but exact recovery requires the diagnostic to expose the generating edit.

With increasing initial AST distance, the two endpoints separate further (Fig. 3c–d and Table 2). FRR achieves its largest behavioral and structural gains at distance one, but behavioral intervals remain above zero through distance three while exact recovery falls sharply beyond the first edit. At distance five, exact recovery is also beyond the four-round one-edit-per-round cap; the zero endpoint therefore reflects the finite controller budget as well as basin difficulty. Local repair therefore has a wider behavioral radius than syntactic recovery.

5.3 Behavioral transfer requires basin entry

Across seven unseen ODEBench systems, FRR has the highest descriptive semantic mean (56.5%) and exact recovery (26.8%), but CV-FRR is close (56.0% and 25.0%, respectively); all three prespecified contrasts remain inconclusive after Holm correction. On an independent endpoint, FRR reduces trajectory NRMSE among jointly integrable trajectories at every tested Van der Pol and Lorenz-63 horizon, with paired intervals excluding zero; Lorenz divergence rates are unchanged. The supplementary cross-family audit finds shorter-horizon improvement for the canonical delay system, smaller switching effects, and uncertainty at the longest delay horizon (Table 3).

Discovery-generated candidates probe basin entry rather than local ranking. In the original 112-candidate STLSQ/SR3 cohort (median AST distance 15), FRR reaches 74.0% semantic and 8.4% exact recovery, while CV-FRR reaches 76.9% and 13.9%; neither arm records a harmful edit. This cohort isolates basin entry: local repair improves a subset of discovery outputs, and its reach depends on the candidate-generation protocol. In an additional 80-candidate update study, coefficient refitting already achieves semantic success for every candidate; FRR makes no structural edits and retains all ten exact models (supplement, Discovery-generated model updates).

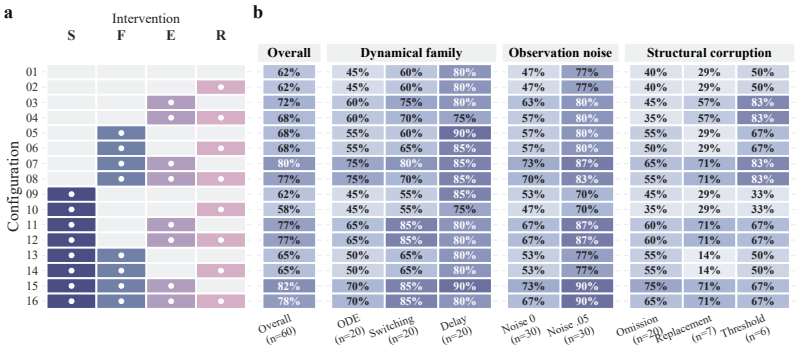

Figure 4: Complete 2^4 intervention and semantic-success strata.

Table 3: Held-out and OOD transfer.

Method	Semantic↑	Exact↑	OOD@.10↑	Stable↑	Ver.↓	Fits↓	API↓
Rewrite	42.3	0.6	72.4	93.5	2.9	**1.0**	0.98
Typed/no replay	53.0	18.5	86.7	94.0	3.8	7.7	0.98
Agent-FRR	51.8	20.8	84.6	**100.0**	4.3	7.7	0.98
FRR	**56.5**	**26.8**	**89.0**	**100.0**	5.3	7.9	**0.00**
Random-FRR	52.4	20.8	82.5	**100.0**	5.1	10.0	0.98
CV-FRR	56.0	25.0	88.8	94.6	5.4	9.6	**0.00**

Prespecified paired contrasts

Contrast	Estimate	95% CI	p_{Holm}	Decision
CV-FRR − FRR semantic	-0.006	[-0.042, +0.030]	0.394	Inconclusive
CV-FRR − FRR call reduction	-0.092	[-0.180, -0.015]	0.978	Inconclusive
CV-FRR − Agent-FRR semantic	+0.042	[-0.006, +0.089]	0.163	Inconclusive

Positive estimates favor CV-FRR; intervals are descriptive, whereas p_{Holm} tests the prespecified one-sided hypothesis.

Table 4: Factorial component effects.

Endpoint	Component	Effect (pp)	95% CI (pp)	$p_{\text{Holm},4}$
Semantic success	Active selector	+0.8	[-4.0, +5.8]	0.753
	Typed feedback	+5.8	[+1.5, +10.8]	0.044
	Typed edit	**+12.5**	**[+5.8, +20.2]**	**0.006**
	Replay	-1.7	[-4.0, +0.0]	0.211
Exact recovery	Active selector	+5.2	[+0.6, +10.6]	0.114
	Typed feedback	+3.1	[-0.6, +7.1]	0.235
	Typed edit	**+25.6**	**[+15.0, +36.9]**	< 0.001
	Replay	-0.6	[-1.7, +0.0]	0.235
AST closer	Active selector	+5.6	[+0.8, +11.0]	0.086
	Typed feedback	+4.4	[+0.6, +8.3]	0.086
	Typed edit	**+26.0**	**[+15.0, +37.1]**	< 0.001
	Replay	-0.6	[-1.7, +0.0]	0.128

5.4 Typed editing drives gains; replay preserves tests

Typed editing has the largest component effect in Table 4: +12.5 semantic points (95% CI [5.8,20.2]; Holm $p = .006$), alongside structural gains. Typed feedback adds 5.8 points ($p = .044$) under the four-component correction; only editing survives the broader ten-term sensitivity correction. Figure 4 shows how feedback and the surrounding policy modulate the benefit of the grammar.

Across 495 paired tasks, Agent-FRR and FRR agree on 96.8% of action sequences and 98.8% of act/abstain decisions, with 0.2% contract errors. Most choices therefore do not require a language model. The contribution lies in the diagnostic and executable decision rules shared by both selectors.

Replay addresses preservation rather than discovery. In the matched CV-FRR comparison, replay reduces runs with final regression from 39 to 11, at a cost of 2.86 semantic-success points; exact

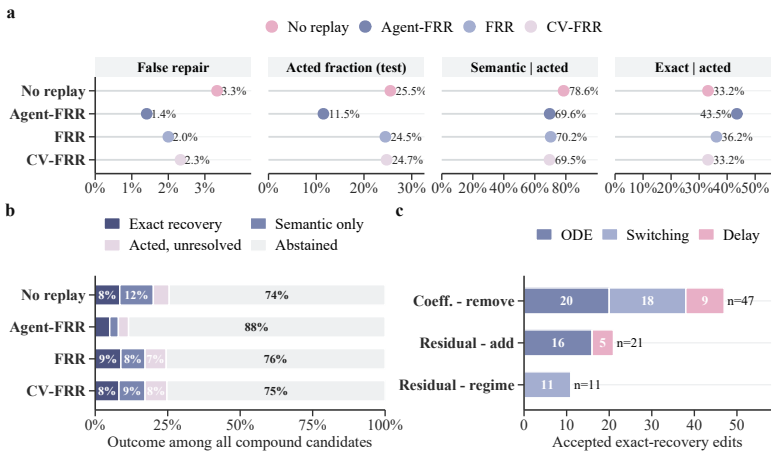

Figure 5: Selective operation at the 5% false-repair budget.

recovery changes little. Deterministic FRR separately records zero final regressions. Typed editing produces the repair gain; replay trades some acceptance for continuity with previously passed tests.

5.5 Selective deployment

Selective deployment restricts edits to candidates with sufficient measured improvement. A separate calibration split fixes the minimum NRMSE gain using an upper binomial bound on clean-law edits. At the 5% budget, FRR reports 84.5% calibrated coverage and 2.0% observed false repair; the untouched test partition contains 188 acted decisions (24.5% of 768 tasks), among which 70.2% achieve semantic success and 36.2% exact recovery. Coverage uses the frozen calibration denominator, whereas the acted fraction uses the test-task denominator. Figure 5 reports the acted-decision outcome mixture and witness–action routes of exact repairs. The risk estimate applies to the evaluated distribution.

6 Discussion and limitations

FRR uses residual evidence to restrict which structural changes deserve testing. The factorial study identifies typed editing as the largest source of gain, and the local-search comparisons show that high exact recovery can require few fits. Optimizing validation error alone instead favors behavioral agreement, sometimes at the expense of the generating structure. Replay and selective deployment add control over preservation and unnecessary intervention.

The method targets local structural debugging when the variables and equation family are known and the grammar contains a useful repair path. Compound errors reveal a behavioral radius wider than exact identification, while discovery-generated candidates expose basin entry as the complementary challenge. Cross-family rollouts broaden the independent behavioral audit, but long-horizon delayed dynamics remain uncertain and the selective risk estimate is tied to the evaluated distribution. Broader proposal mechanisms and sequential calibration can extend the same authorization contract without weakening replay or abstention.

7 Conclusion

FRR links local equation edits to explicit diagnostic evidence. Typed witnesses restrict the search, numerical refitting evaluates candidates, and replay checks consistency with earlier tests. The experiments identify typed editing as the main source of controlled gains and show that strong local structural recovery can be achieved with few fits, while behavioral improvement extends beyond exact reconstruction. Discovery-generated candidates delimit where this repair mechanism helps. Separating proposal from acceptance allows deterministic and learned selectors to share the same auditable procedure, including the decision to leave a model unchanged.

AI USE STATEMENT

Generative AI tools assisted with coding, data processing, and language polishing. All analyses, results, figures, citations, and final text were independently verified by the authors. A language-model selector was also evaluated as part of the reported method.

ETHICS STATEMENT

The physiological interface study uses de-identified waveform data from existing repositories. No re-identification was attempted, and the models are not intended for clinical use. All other experiments use synthetic systems or public benchmarks.

REPRODUCIBILITY STATEMENT

The main paper describes the repair interface, evaluation setup, and primary comparisons. The supplement provides implementation details, statistical procedures, additional results, and proofs. Fixed configurations and run records support reproducibility and link each reported result to its underlying experiment.

REFERENCES

Rajeev Alur, Rastislav Bodik, Garvit Juniwal, Milo M. K. Martin, Mukund Raghothaman, Sanjit A. Seshia, Rishabh Singh, Armando Solar-Lezama, Emina Torlak, and Abhishek Udupa. Syntax-guided synthesis. In *Formal Methods in Computer-Aided Design*, pp. 1–8, 2013.

Anastasios N. Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023. doi: 10.1561/2200000101.

Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael I. Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM*, 68(6):43:1–43:34, 2021. doi: 10.1145/3478535.

Daniil A. Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 624:570–578, 2023. doi: 10.1038/s41586-023-06792-0.

Josh Bongard and Hod Lipson. Automated reverse engineering of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 104(24):9943–9948, 2007. doi: 10.1073/pnas.0609476104.

Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D. White, and Philippe Schwaller. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, 6:525–535, 2024. doi: 10.1038/s42256-024-00832-8.

Steven L. Brunton and J. Nathan Kutz. *Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control*. Cambridge University Press, 2019. doi: 10.1017/9781108380690.

Steven L. Brunton, Joshua L. Proctor, and J. Nathan Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 113(15):3932–3937, 2016. doi: 10.1073/pnas.1517384113.

Kathleen Champion, Bethany Lusch, J. Nathan Kutz, and Steven L. Brunton. Data-driven discovery of coordinates and governing equations. *Proceedings of the National Academy of Sciences*, 116 (45):22445–22451, 2019. doi: 10.1073/pnas.1906995116.

Jindou Chen, Jidong Tian, Liang Wu, Xinwei Chen, Xiaokang Yang, Yaohui Jin, and Yanyan Xu. KinFormer: Generalizable dynamical symbolic regression for catalytic organic reaction kinetics. In *International Conference on Learning Representations*, 2025.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K. Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Miles Cranmer. Interpretable machine learning for science with PySR and SymbolicRegression.jl. *arXiv preprint arXiv:2305.01582*, 2023. doi: 10.48550/arXiv.2305.01582.

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. In *Advances in Neural Information Processing Systems*, volume 33, pp. 17429–17442, 2020.

Stéphane d'Ascoli, Sören Becker, Philippe Schwaller, Alexander Mathis, and Niki Kilbertus. ODEFormer: Symbolic regression of dynamical systems with transformers. In *International Conference on Learning Representations*, 2024.

Brian M. de Silva, Kathleen Champion, Markus Quade, Jean-Christophe Loiseau, J. Nathan Kutz, and Steven L. Brunton. PySINDy: A python package for the sparse identification of nonlinear dynamical systems from data. *Journal of Open Source Software*, 5(49):2104, 2020. doi: 10.21105/joss.02104.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall/CRC, 1993.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11:1605–1641, 2010.

Urban Fasel, J. Nathan Kutz, Bingni W. Brunton, and Steven L. Brunton. Ensemble-SINDy: Robust sparse model discovery in the low-data, high-noise limit, with active learning and control. *Proceedings of the Royal Society A*, 478(2260):20210904, 2022. doi: 10.1098/rspa.2021.0904.

Dean P. Foster and Robert A. Stine. Alpha-investing: A procedure for sequential control of expected false discoveries. *Journal of the Royal Statistical Society: Series B*, 70(2):429–444, 2008. doi: 10.1111/j.1467-9868.2007.00643.x.

Dayuan Fu, Keqing He, Yejie Wang, Wentao Hong, Zhuoma Gongque, Weihao Zeng, Wei Wang, Jingang Wang, Xunliang Cai, and Weiran Xu. AgentRefine: Enhancing agent generalization through refinement tuning. In *International Conference on Learning Representations*, 2025.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Sumit Gulwani, Oleksandr Polozov, and Rishabh Singh. Program synthesis. *Foundations and Trends in Programming Languages*, 4(1–2):1–119, 2017. doi: 10.1561/2500000010.

Moritz Hoffmann, Christoph Fröhner, and Frank Noé. Reactive SINDy: Discovering governing reactions from concentration data. *The Journal of Chemical Physics*, 150(2):025101, 2019. doi: 10.1063/1.5066099.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

Samuel Holt, Zhaozhi Qian, and Mihaela van der Schaar. Deep generative symbolic regression. In *International Conference on Learning Representations*, 2023.

Susmit Jha and Sanjit A. Seshia. Are there good mistakes? a theoretical analysis of CEGIS. *arXiv preprint arXiv:1407.5397*, 2014.

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, et al. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1), 2016. doi: 10.1038/sdata.2016.35.

Krzysztof Kacprzyk and Mihaela van der Schaar. No equations needed: Learning system dynamics without relying on closed-form ODEs. In *International Conference on Learning Representations*, 2025.

Kadierdan Kaheman, J. Nathan Kutz, and Steven L. Brunton. SINDy-PI: a robust algorithm for parallel implicit sparse identification of nonlinear dynamics. *Proceedings of the Royal Society A*, 476(2242):20200279, 2020. doi: 10.1098/rspa.2020.0279.

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*, volume 34, 2021.

Claire Le Goues, ThanhVu Nguyen, Stephanie Forrest, and Westley Weimer. GenProg: A generic method for automatic software repair. *IEEE Transactions on Software Engineering*, 38(1):54–72, 2012. doi: 10.1109/TSE.2011.104.

Hyung-Chul Lee, Yoonsang Park, Soo Bin Yoon, Seong Mi Yang, Dongnyeok Park, and Chul-Woo Jung. VitalDB, a high-fidelity multi-parameter vital signs database in surgical patients. *Scientific Data*, 9(1), 2022. doi: 10.1038/s41597-022-01411-5.

Yunlun Li and Sinno Jialin Pan. Robust equation structure learning with adaptive refinement. In *International Conference on Learning Representations*, 2026.

Greta M. Ljung and George E. P. Box. On a measure of lack of fit in time series models. *Biometrika*, 65(2):297–303, 1978. doi: 10.1093/biomet/65.2.297.

Lennart Ljung. *System Identification: Theory for the User*. Prentice Hall, 2 edition, 1999.

Fan Long and Martin Rinard. Automatic patch generation by learning correct code. In *ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages*, pp. 298–312, 2016. doi: 10.1145/2837614.2837617.

Georg Martius and Christoph H. Lampert. Extrapolation and learning equations. *arXiv preprint arXiv:1610.02995*, 2016.

Sergey Mechtaev, Jooyong Yi, and Abhik Roychoudhury. DirectFix: Looking for simple program repairs. In *International Conference on Software Engineering*, pp. 448–458, 2015. doi: 10.1109/ICSE.2015.63.

Daniel A. Messenger and David M. Bortz. Weak SINDy: Galerkin-based data-driven model selection. *Multiscale Modeling & Simulation*, 19(3):1474–1497, 2021. doi: 10.1137/20M1343166.

Benjamin Moody, George Moody, Mauricio Villarroel, Gari D. Clifford, and Ikaro Silva. MIMIC-III Waveform Database. *PhysioNet*, 2020. doi: 10.13026/c2607m. Version 1.0.

Hoang Duong Thien Nguyen, Dawei Qi, Abhik Roychoudhury, and Satish Chandra. SemFix: Program repair via semantic analysis. In *International Conference on Software Engineering*, pp. 772–781, 2013. doi: 10.1109/ICSE.2013.6606623.

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*, 2021.

Zichao Qi, Fan Long, Sara Achour, and Martin Rinard. An analysis of patch plausibility and correctness for generate-and-validate patch generation systems. In *International Symposium on Software Testing and Analysis*, pp. 24–36, 2015. doi: 10.1145/2771783.2771791.

Zhaozhi Qian, Krzysztof Kacprzyk, and Mihaela van der Schaar. D-CODE: Discovering closed-form ODEs from observed trajectories. In *International Conference on Learning Representations*, 2022.

Christopher Rackauckas, Yingbo Ma, Julius Martensen, Collin Warner, Kirill Zubov, Rohit Supekar, Dominic Skinner, Ali Ramadhan, and Alan Edelman. Universal differential equations for scientific machine learning. *arXiv preprint arXiv:2001.04385*, 2020. doi: 10.48550/arXiv.2001.04385.

Aaditya Ramdas, Tijana Zrnic, Martin J. Wainwright, and Michael I. Jordan. SAFFRON: An adaptive algorithm for online control of the false discovery rate. In *International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4286–4294, 2018.

Samuel H. Rudy, Steven L. Brunton, Joshua L. Proctor, and J. Nathan Kutz. Data-driven discovery of partial differential equations. *Science Advances*, 3(4), 2017. doi: 10.1126/sciadv.1602614.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems*, volume 36, pp. 68539–68551, 2023.

Michael Schmidt and Hod Lipson. Distilling free-form natural laws from experimental data. *Science*, 324(5923):81–85, 2009. doi: 10.1126/science.1165893.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 8634–8652, 2023.

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*, 2025.

Armando Solar-Lezama, Liviu Tancau, Rastislav Bodik, Sanjit Seshia, and Vijay Saraswat. Combinatorial sketching for finite programs. In *Proceedings of the 12th International Conference on Architectural Support for Programming Languages and Operating Systems*, pp. 404–415, 2006. doi: 10.1145/1168857.1168907.

Silviu-Marian Udrescu and Max Tegmark. AI Feynman: A physics-inspired method for symbolic regression. *Science Advances*, 6(16):eaay2631, 2020. doi: 10.1126/sciadv.aay2631.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V. Le, Denny Zhou, and Xinyun Chen. Large language models as optimizers. In *International Conference on Learning Representations*, 2024.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023.

Hengzhe Zhang, Qi Chen, Bing Xue, Wolfgang Banzhaf, and Mengjie Zhang. RAG-SR: Retrieval-augmented generation for neural symbolic regression. In *International Conference on Learning Representations*, 2025.

SUPPLEMENTARY MATERIAL:

FALSIFY, REPAIR, REPEAT:

AUDITABLE LOCAL REPAIR OF TYPED DYNAMICAL EQUATIONS

Anonymous authors

Paper under double-blind review

CONTENTS

List of Figures **15**

List of Tables **16**

A Experimental protocol and provenance **17**
A.1 Extension experiment summary . . . 17

B Executable hypothesis representation **18**
B.1 Verifier registry and calibrated thresholds . . . 18
B.2 Repair algorithm . . . 19
B.3 Conditional recoverability proof . . . 20
B.4 Conditional run-level error spending . . . 21

C Complete treatments and factorial design **22**

D Task generation and external registry **24**
D.1 Controlled tasks . . . 24
D.2 External ODEBench audit . . . 25
D.3 Natural-candidate stress test . . . 26

E Compound-error recovery radius and replay stress **27**
E.1 Paired uncertainty and independent rollout audit . . . 28
E.2 Exploratory basin-entry preconditioner . . . 30
E.3 Terminal outcomes and run accounting . . . 31
E.4 Cost–quality frontier . . . 32

F Selective deployment **33**

G Physiological-law interface study **34**

H Confirmatory statistics **35**

I Null and out-of-distribution controls **36**

J Independent cross-family behavior and agent attribution **37**
J.1 Switching and delay rollouts . . . 37
J.2 Constrained agent-selector attribution . . . 38
J.3 Calibration operating points . . . 39

K Matched residual-refinement baseline **40**

L Language-model interface and compute audit **41**

M Method identification and search cost **42**
M.1 Local-search comparison . . . 42
M.2 Paired effects and fit-budget sensitivity . . . 43
M.3 Compound budget dependence and common-pool decisions . . . 44
M.4 Discovery-generated model updates . . . 45

LIST OF FIGURES

S1 Family-resolved factorial outcomes. 23
S2 External repair by ODEBench system. 25
S3 Compound recovery across depth and corruption profile. 27
S4 Outcome composition across the local repair radius. 28
S5 Terminal outcomes by method and dynamical family. 31
S6 Negative controls and OOD stress tests. 36
S7 Independent rollouts and selector agreement. Panel b is shown over the 90–100% range for readability; labels above the bars report the exact percentages. 38
S8 Language-model outcomes by execution group. 41

LIST OF TABLES

S1 Confirmatory evaluation design. . . . 17
S2 Extension studies and outcome coverage. . . . 17
S3 Calibrated verifier thresholds. . . . 18
S4 Verifier sensitivity across calibration quantiles. . . . 18
S5 Matched treatment matrix. . . . 22
S6 Factorial main effects. . . . 23
S7 Raw natural-candidate stress test on 112 discovery outputs. . . . 26
S8 Natural-candidate extension: 224 candidates, 32 per system. . . . 26
S9 Compound-error outcomes under the compound-error protocol. . . . 27
S10 Paired semantic gains at AST distances two and three. . . . 28
S11 Free rollout from unseen initial conditions. . . . 28
S12 Full-denominator rollout audit and failure-aware sensitivity. . . . 29
S13 Exploratory basin-entry audit. . . . 30
S14 Terminal outcomes in the completed repair study. . . . 31
S15 Method-level cost and quality summary for the compound-error protocol. . . . 32
S16 Selective outcomes at the 5% false-repair budget. Coverage uses the frozen calibration denominator; Acted is the number of accepted test decisions. . . . 33
S17 Participant-level physiology repair. . . . 34
S18 Prespecified confirmatory contrasts. . . . 35
S19 Independent switching and delay rollouts. . . . 37
S20 M3–M4 selector attribution. . . . 38
S21 Calibration operating points on the untouched test partition. Coverage uses the frozen calibration denominator; Acted n is the number of accepted test decisions. . . . 39
S22 Matched residual-refinement comparison. . . . 40
S23 Compute and reliability summary. . . . 41
S24 Language-model selector decisions. . . . 41
S25 One-edit local-search comparison under the method-identification protocol (64-fit ceiling). . . . 42
S26 Compound local-search comparison under the method-identification protocol (192-fit ceiling; $n = 768$ per row). . . . 42
S27 Full-budget paired differences: FRR minus each search baseline. . . . 43
S28 Complete one-edit fit-budget ablation ($n = 180$ per row). . . . 43
S29 Complete compound fit-budget ablation under the method-identification protocol ($n = 768$ per row). . . . 44
S30 One-step decisions on identical fitted candidate pools. . . . 44
S31 All discovery-generated update candidates before structural repair. . . . 45
S32 Model-update outcomes after coefficient adaptation ($n = 80$). . . . 45

A EXPERIMENTAL PROTOCOL AND PROVENANCE

Verifier implementations, task generators, budgets, prompt schemas, and reporting rules were specified before confirmatory evaluation. Seven thresholds were calibrated at the 0.99 quantile using 1,500 correct-model development rows. The confirmatory study contains 5,168 runs over 548 tasks under common methods, margins, bootstrap procedures, and decision rules.

The primary compound-error, replay, and selective-deployment studies use a second protocol with disjoint calibration and test phases: 3,728 calibration runs, 4,543 test runs, and 10,425 selective-test rows. The later extension archive is separate: its calibration set contains 4,848 rows, of which 4,800 clean-law rows enter threshold fitting (see the extension summary below). Agent-FRR uses a predefined 200-task language-model subset; the remaining arms use all 768 compound tasks. A separate 240-run physiology study evaluates interface portability. Fixed configurations, model identities, source data, and analysis records link each reported value to its generating run.

Table S1: Confirmatory evaluation design.

Study	Evaluations	Purpose
Core	600	matched M0–M9 benchmark
Challenge	1,200	controlled corruption/noise stress
Mechanism	960	complete 2^4 factorial design
External	1,008	seven-system ODEBench repair audit
Null/OOD	1,400	false-repair, false-support, and shift controls
Total	5,168	repair and control runs

Three external contrasts were specified before observing external outcomes and tested with paired bootstrap procedures and Holm correction. Separately, 112 STLSQ/SR3 fits provide discovery candidates. Model settings, cache behavior, and failure accounting are held fixed.

A.1 EXTENSION EXPERIMENT SUMMARY

The supplementary studies examine four issues: basin entry, clean-law calibration, independent long-horizon behavior, and a matched residual-refinement selector. The natural-candidate study evaluates 224 discovery outputs with three policies (672 task-policy evaluations); all policies have 55.8% semantic success and 3.6% exact recovery. The calibration archive contains 1,200 clean-law rows plus 12 fixed compound-omission probe rows per policy (4,848 rows total); only the 4,800 clean-law rows enter threshold calibration, yielding 168/1,200 observed false-repair events for M4, M9, and M10, compared with 0/1,200 for the unedited control. The independent rollout audit contains 3,072 retained trajectories across switching and delay systems at $1\times$, $2\times$, and $4\times$ horizons; these outcomes are withheld from selection. The matched residual-refinement comparison includes 224 task pairs (448 runs): M4 and M7 obtain the same endpoint rates, while M7 uses 2.02 versus 2.36 verifier calls per run. Contract and repair failures remain in all denominators. The extension archive provides task-level results and analysis code.

Table S2: Extension studies and outcome coverage.

Study	Rows	Primary comparison	Outcome recorded
Natural candidates	672	M0/M4/M9	Semantic and exact recovery; basin-entry diagnosis
Clean-law calibration	4,848	M0/M4/M9/M10	False-repair events and one-sided upper bounds
Switching/delay rollout	3,072	4 policies, 3 horizons	Trajectory NRMSE and divergence, with all rows retained
Matched baseline	448	M4 vs M7	Endpoint rates, verifier calls, API and contract outcomes

B EXECUTABLE HYPOTHESIS REPRESENTATION

Each hypothesis is a typed record containing a family (ODE, regime, or delay), target-indexed terms, coefficient keys, an optional delay in integer samples, and optional regime metadata. Expressions are parsed into a restricted abstract syntax tree. Unknown free-form code is rejected.

For a state $\boldsymbol{x} \in \mathbb{R}^d$, an ODE candidate is

$$\dot{x}_j = \sum_{q \in \mathcal{S}_j(h)} \theta_{jq} \phi_q(\boldsymbol{x}), \qquad j = 1, \ldots, d. \tag{10}$$

Delay candidates may include $\phi_q(\boldsymbol{x}_t, \boldsymbol{x}_{t-\tau})$. Regime candidates use a threshold predicate $x_k \leq b$ and separate term sets for the two modes. Structural edits modify $\mathcal{S}$, τ, or b; coefficients θ are refitted by deterministic numerical code.

B.1 VERIFIER REGISTRY AND CALIBRATED THRESHOLDS

Table S3: Calibrated verifier thresholds.

Verifier	Threshold	Pass rule
Coefficient relevance	0.134596	statistic $\geq$ threshold
Delay profile	3.464496	statistic $\leq$ threshold
Normalized derivative fit	0.869695	statistic $\leq$ threshold
Regime boundary profile	0.009241	statistic $\leq$ threshold
Regime occupancy	0.467903	statistic $\geq$ threshold
Residual feature correlation	0.326376	statistic $\leq$ threshold
Residual whiteness	0.981351	statistic $\leq$ threshold

The verifier never observes the generating equation. It receives only the executable candidate and discovery trajectory. A failed result contains the statistic, threshold, effect, localized components, witness summaries, and allowed repair operator names. Invalid numerical evaluations are recorded as invalid, not converted to passes.

Threshold sensitivity. Re-thresholding the exhaustive initial-candidate diagnostic at the 0.975, 0.99, and 0.995 quantiles leaves detection unchanged for five of seven verifiers. Normalized derivative fit and regime boundary profile vary by at most 1.1 and 1.7 percentage points, respectively (Table S4). The least conservative quantile primarily increases clean-law regime-boundary triggers. This diagnostic changes no repair decision and measures verifier operating characteristics rather than end-to-end policy performance.

Table S4: Verifier sensitivity across calibration quantiles.

Verifier	Applicable n	Detection range	Clean-law FP range
Coefficient relevance	180	64.4%	1.7–3.3%
Delay profile	60	15.0%	0.0%
Normalized derivative fit	180	5.6–6.7%	0.0–1.7%
Regime boundary profile	60	60.0–61.7%	0.0–5.0%
Regime occupancy	60	33.3%	0.0%
Residual feature correlation	180	41.1%	0.0%
Residual whiteness	180	21.7%	0.0%

B.2 REPAIR ALGORITHM

The controller alternates four typed operations:

Falsify	Refit the current structure, rank unused family-compatible verifiers by the fixed cost-aware utility, and retain passing tests for later replay. A failure emits a localized typed witness.
Propose	Enumerate at most 32 unique one-edit children consistent with the witness, then choose one using the treatment-specific selector. Invalid or duplicate structures remain in the trace.
Validate	Refit the child and require both resolution of the active failure and preservation of every earlier pass. A PASS→INCONCLUSIVE transition is recorded but is not a hard regression.
Terminate	Refit an accepted structure once more, report final replay stability, and stop on acceptance or at four rounds, eight verifier calls, or 64 new fits.

The active utility is $U_k(t) = (\pi_t + 0.7\ell_t + 0.7\rho_t + 0.25\nu_{t,k})/c_t$, with fixed component values for each verifier. It defines an execution priority, not a calibrated probability, and is unchanged across evaluation tasks.

B.3 CONDITIONAL RECOVERABILITY PROOF

Let $\mathcal{A}_k = A(r_k)$, let $\mathcal{C}_k = \{h_k \oplus a : a \in \mathcal{A}_k\}$, and suppose $a^\star \in \mathcal{A}_k$ lies on an admissible path to $h^\star$. Assume

$$\min_{a \in \mathcal{A}_k \setminus \{a^\star\}} \{R(h_k \oplus a) - R(h_k \oplus a^\star)\} \geq \Delta_k \quad \text{and} \quad \sup_{h \in \mathcal{C}_k} |\hat{R}_{\text{val}}(h) - R(h)| \leq \epsilon_k. \tag{11}$$

Then

$$\begin{aligned} \hat{R}_{\text{val}}(h_k \oplus a) &\geq R(h_k \oplus a) - \epsilon_k && (12)\\ &\geq R(h_k \oplus a^\star) + \Delta_k - \epsilon_k && (13)\\ &> R(h_k \oplus a^\star) + \epsilon_k && (14)\\ &\geq \hat{R}_{\text{val}}(h_k \oplus a^\star) && (15) \end{aligned}$$

whenever $\Delta_k > 2\epsilon_k$. The strict inequality makes $a^\star$ the unique empirical minimizer. Apply the same argument inductively at rounds $1, \ldots, K$: if the witness localizes the next defect, $a_k^\star$ is enumerated, the uniform error bound and margin hold, and refitting plus replay accept the selected successor, the controller follows the admissible path and reaches $h^\star$ within K rounds.

B.4 CONDITIONAL RUN-LEVEL ERROR SPENDING

Let $E_{k,t} = \{p_{k,t} \leq \alpha_{k,t}\}$ be the first false rejection that can authorize a structural edit while the current equation is correct. Conditional super-uniformity gives

$$\Pr(E_{k,t} \mid \mathcal{F}_{k,t}^{-}) \leq \alpha_{k,t}, \tag{16}$$

where $\mathcal{F}_{k,t}^{-}$ is the full trace immediately before verifier (k, t) and each allocation is measurable with respect to that history. If a run starts from a correct structure, its first false edit must be preceded by at least one $E_{k,t}$. The tower property and union bound therefore yield

$$\Pr(\text{any false edit}) \leq \Pr\left(\bigcup_{k,t} E_{k,t}\right) \leq \sum_{k,t} \mathbb{E}[\Pr(E_{k,t} \mid \mathcal{F}_{k,t}^{-})] \leq \sum_{k,t} \alpha_{k,t} \leq \alpha. \tag{17}$$

Conditional validity permits adaptive verifier order and early stopping. The experimental marginal quantiles provide per-test thresholds; the run-level bound requires the stronger conditional premise above.

C COMPLETE TREATMENTS AND FACTORIAL DESIGN

The matched methods are listed below. M1 receives a 4,096-token ceiling because 2,400 tokens truncated some full-law rewrites; constrained language-model selectors use 2,400 tokens. All language-model arms share the same model, temperature 0, input limit, timeout, retry policy, and cache semantics. M8 receives true corruption localization and serves only as an upper bound.

Table S5: Matched treatment matrix.

ID	Method	Selector	Feedback	Repair	Replay	LM	Oracle	Role
M0	Initial candidate	none	none	none	no	no	no	lower bound
M1	LM full-law rewrite	random	scalar	full law	no	yes	no	rewrite control
M2	Typed/no replay	active	typed	one edit	no	yes	no	replay ablation
M3	Agent-FRR	active	typed	one edit	yes	yes	no	LM selector
M4	FRR	active	typed	beam	yes	no	no	primary implementation
M5	Random + replay	random	typed	one edit	yes	yes	no	selector ablation
M6	Influence repair	active	residual	influence	no	no	no	residual baseline
M7	RESTART-style	active	residual	refinement	yes	yes	no	literature style
M8	Oracle localization	oracle	oracle	one edit	yes	no	yes	upper bound
M9	CV-FRR	active	typed	CV beam	yes	no	no	CV selector alternative
M10	CV-FRR (no replay)	active	typed	CV beam	no	no	no	replay ablation

Rewrite is an end-to-end control that changes the selector, feedback, repair space, and replay relative to FRR; component attribution therefore comes from the complete factorial. Agent-FRR–FRR is the matched policy contrast because both arms share witnesses, admissible edits, refitting, replay, and budgets. Refinement is a residual-guided control inspired by RESTART, not a reproduction of its training and search system.

The mechanism experiment crosses four binary factors:

- selector: deterministic pseudo-random versus active cost-aware;
- feedback: scalar failure signal versus full typed record;
- repair: free rewrite versus finite typed edit;
- replay: disabled versus replay of previously passed tests.

All 16 cells are evaluated rather than using one-factor-at-a-time ablations.

For factor j in the balanced 2^4 design, the marginal effect averages uniformly over the remaining factors,

$$\tau_j = 2^{-3} \sum_{\boldsymbol{x}_{-j} \in \{0,1\}^3} \left\{ \bar{Y}(1, \boldsymbol{x}_{-j}) - \bar{Y}(0, \boldsymbol{x}_{-j}) \right\}, \tag{18}$$

with analogous difference-in-differences for interactions. Bootstrap resampling retains all 16 treatments within each task cluster; individual runs are not treated as independent replicates.

Main-paper Fig. 5c accounts for every accepted edit in the 69 exact-recovery runs. Its 79 actions comprise 47 coefficient-relevance witnesses followed by term removal, 21 residual–feature-correlation witnesses followed by an ordinary term addition, and 11 residual–feature-correlation witnesses followed by a regime-term addition. These are exhaustive counts, not selected examples.

The ten main-effect and pairwise-interaction summaries were not registered as confirmatory claims. We report task-cluster intervals descriptively and apply a Holm sensitivity correction across all ten semantic terms; only typed editing survives. The endpoint-specific analysis below applies Holm correction across the four components within each endpoint. Both analyses use the original factorial cells without changing methods or thresholds.

Table S6: Factorial main effects.

Endpoint	Component	Effect (pp)	95% CI (pp)	$p_{\mathrm{Holm},4}$
Semantic success	Active selector	+0.8	[-4.0, +5.8]	0.753
	Typed feedback	+5.8	[+1.5, +10.8]	0.044
	Typed edit	+12.5	[+5.8, +20.2]	0.006
	Replay	-1.7	[-4.0, +0.0]	0.211
Exact recovery	Active selector	+5.2	[+0.6, +10.6]	0.114
	Typed feedback	+3.1	[-0.6, +7.1]	0.235
	Typed edit	+25.6	[+15.0, +36.9]	< 0.001
	Replay	-0.6	[-1.7, +0.0]	0.235
AST closer	Active selector	+5.6	[+0.8, +11.0]	0.086
	Typed feedback	+4.4	[+0.6, +8.3]	0.086
	Typed edit	+26.0	[+15.0, +37.1]	< 0.001
	Replay	-0.6	[-1.7, +0.0]	0.128

Replay primarily protects previously passed tests rather than increasing semantic recovery. The one-edit study contains 693 pass→pass events, with estimated semantic, exact-recovery, and AST-closer effects of -1.7, -0.6, and -0.6 percentage points. Under compound errors, FRR intercepts 476 potential regressions in 124 runs and finishes with zero post-refit regressions. In the matched CV-FRR comparison, replay intercepts 457 potential regressions in 133 runs and reduces final-regression runs from 39 to 11. The no-replay arm nevertheless has higher aggregate semantic success (68.5% versus 65.9% for FRR), and FRR's semantic gain is zero at AST distance five; CV-FRR loses 2.86 semantic-success points relative to its no-replay arm (95% task-cluster bootstrap CI $[-4.04, -1.69]$). Replay therefore trades acceptance for continuity rather than delivering a uniform accuracy gain. The acceptance check precedes a coefficient refit; the final audit records any resulting regression without rollback, and these 11 returns count as preservation failures.

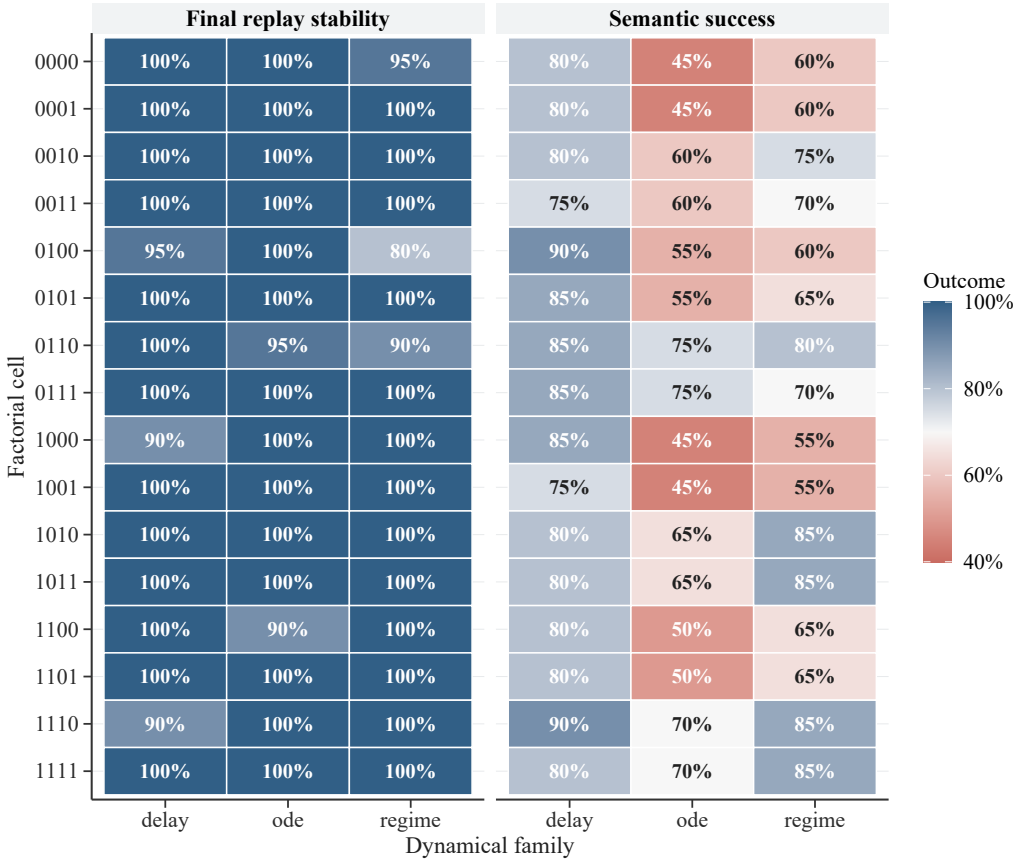

Figure S1: Family-resolved factorial outcomes.

D TASK GENERATION AND EXTERNAL REGISTRY

D.1 CONTROLLED TASKS

Synthetic trajectories have duration 20 and step size 0.02. Systems are Van der Pol, Lorenz-63, switching linear, switching Van der Pol, delayed Stuart–Landau, and a generic delay system. Noise levels are 0 and 0.05; seeds are 101, 113, 127, 139, and 151. Each dataset receives three controlled corruptions sampled from add-spurious-term, omit-term, replace-term, wrong-delay, and wrong-threshold operations when family-compatible. One-edit tasks make component attribution possible inside a grammar-covered local neighborhood; they do not represent arbitrary compound misspecification or estimate how often natural discovery errors fall inside that neighborhood.

Each trajectory is split chronologically: the first 70% supplies coefficient fitting, verifier execution, and repair selection, while the final 30% is reserved for outcome computation. Semantic success requires the candidate to remain in the correct equation family and to have derivative NRMSE on this audit block no larger than the refitted generating structure's NRMSE plus 0.10. Although the split is disjoint, derivative NRMSE remains related to evidence used by several verifiers. Exact recovery, AST-distance movement, four OOD views, the natural-candidate stress test, and unseen-initial-condition rollouts provide complementary structural and behavioral checks.

D.2 External ODEBench audit

The external registry is the Strogatz-expanded ODEBench file distributed with ODEFormer. Systems 54, 63, 49, 45, 56, 50, and 40 are crossed with four seeds, two noise levels, and three corruptions, yielding 168 matched tasks per repair method. Rewrite, Typed/no replay, Agent-FRR, FRR, Random-FRR, and CV-FRR are evaluated on every task. SINDy-STLSQ and SINDy-SR3 are fitted from scratch on 56 trajectories each using PySINDy; they contextualize proposal performance but are not paired repair treatments.

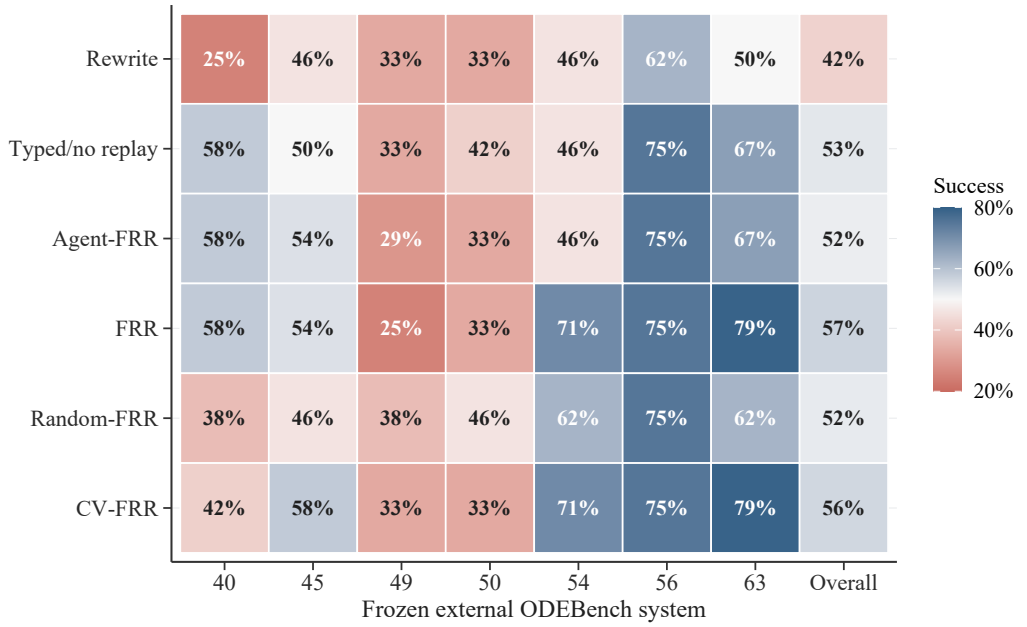

Figure S2: External repair by ODEBench system.

Classical baselines achieve mean semantic success 0.500 (STLSQ) and 0.429 (SR3), exact recovery 0 and 0.071, and mean structural F1 0.281 and 0.351. Their derivative NRMSE distribution is heavy-tailed; consequently, mean NRMSE is not used to assert rank.

D.3 NATURAL-CANDIDATE STRESS TEST

The controlled benchmark starts one admissible edit from the generating law. To test the boundary of that assumption without tuning on the main or external outcomes, we retain the 112 completed equations produced from scratch by STLSQ and SR3 on 56 ODEBench trajectories, then apply Initial, FRR, and CV-FRR with the same typed verifier and one-edit repair grammar. The candidate registry, methods, endpoints, cluster unit, multiplicity correction, and seed were fixed before execution. Candidate-to-truth edit distance is much larger than in the controlled benchmark (median 15; maximum 134). In the original 112-candidate STLSQ/SR3 cohort, FRR reaches 74.0% semantic and 8.4% exact recovery, while CV-FRR reaches 76.9% and 13.9%; neither controller records a harmful edit. The corresponding verifier-call counts are 10 and 22. This cohort quantifies basin-entry dependence: local repair improves a subset of discovery outputs, but remains conditional on the candidate-generation protocol.

Table S7: Raw natural-candidate stress test on 112 discovery outputs.

Method	Semantic (%)	Exact (%)	Harmful edits	Verifier calls
M0	56.2	3.6	0/112	0.0
M4 (FRR)	74.0	8.4	0/112	10
M9 (CV-FRR)	76.9	13.9	0/112	22

A separate extension evaluates 224 candidates (32 per system). Table S8 reports this larger cohort: all three policies attain 55.8% semantic success and 3.6% exact recovery. The coefficient-screening analysis below uses the original 112 candidates.

Table S8: Natural-candidate extension: 224 candidates, 32 per system.

		Shared outcomes: Initial, FRR, CV-FRR			Mean verifier calls		
System	n	Sem. (%)	Exact (%)	F1	Initial	FRR	CV-FRR
ODEBench-40	32	68.8	25.0	0.504	0.00	3.00	3.00
ODEBench-45	32	0.0	0.0	0.428	0.00	2.00	2.00
ODEBench-49	32	43.8	0.0	0.420	0.00	2.00	2.00
ODEBench-50	32	50.0	0.0	0.400	0.00	2.00	2.00
ODEBench-54	32	100.0	0.0	0.209	0.00	2.00	2.00
ODEBench-56	32	100.0	0.0	0.209	0.00	2.00	2.00
ODEBench-63	32	28.1	0.0	0.045	0.00	3.50	3.66

E COMPOUND-ERROR RECOVERY RADIUS AND REPLAY STRESS

The compound study crosses four initial AST distances $(1, 2, 3, 5)$, four corruption profiles (omission, spurious, mixed, and mechanism), three dynamical families, two observation-noise levels, and four seeds. FRR and CV-FRR retain replay; the matched CV-FRR ablation removes it. Agent-FRR replaces the numerical selector with the language-model policy on a predefined 200-task subset. Figure S3 reports the complete matrix, including candidates that are already behaviorally adequate.

Table S9: Compound-error outcomes under the compound-error protocol.

Method	Tasks	Semantic	95% CI	Exact	95% CI	Mean verifier calls
Initial	768	57.0%	[53.5, 60.5]	0.0%	[0.0, 0.5]	0.00
CV-FRR/no replay	768	**68.5%**	[65.1, 71.7]	8.5%	[6.7, 10.6]	3.96
Agent-FRR†	200	66.0%	[59.2, 72.2]	**9.5%**	[6.2, 14.4]	3.64
FRR	768	65.9%	[62.5, 69.2]	8.9%	[7.0, 11.1]	5.67
CV-FRR	768	65.6%	[62.2, 68.9]	8.2%	[6.5, 10.4]	5.72

†Agent-FRR uses the predefined language-model subset and is not paired against the 768-task arms.

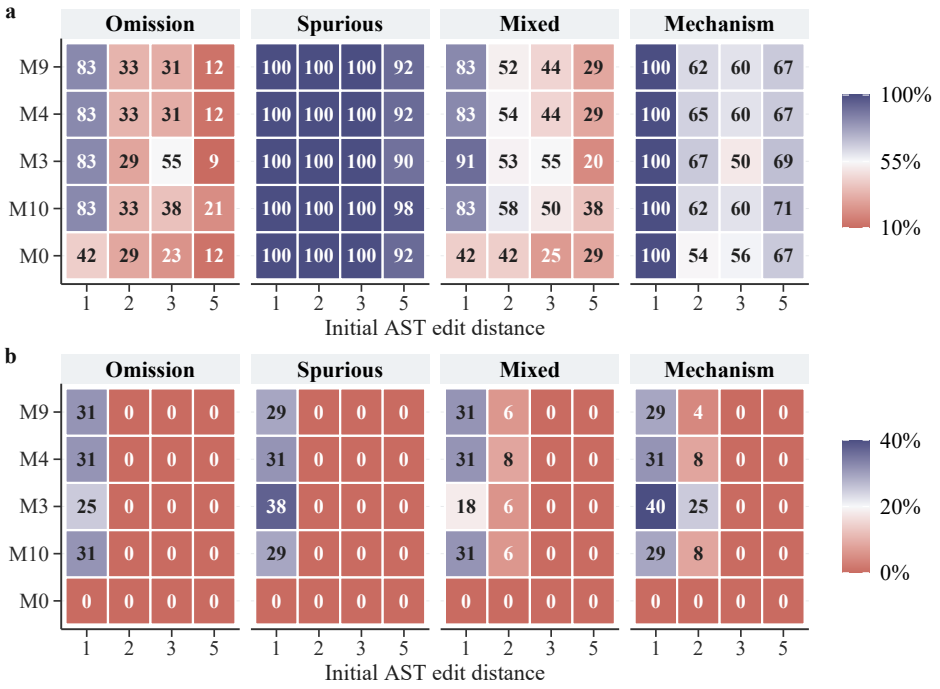

Figure S3: Compound recovery across depth and corruption profile.

At distance one, FRR raises semantic success from 70.8% to 91.7% and recovers the exact structure in 31.3% of tasks. Its semantic advantage narrows to 6.8 points at distance two and 7.8 at distance three, then vanishes at distance five. Across all depths, the no-replay arm reaches 68.5% semantic success versus 65.9% for FRR; its larger gains at distant candidates come with post-refit regressions. Figure S4 makes the endpoint separation explicit: semantic-only improvements persist after exact recovery has nearly disappeared.

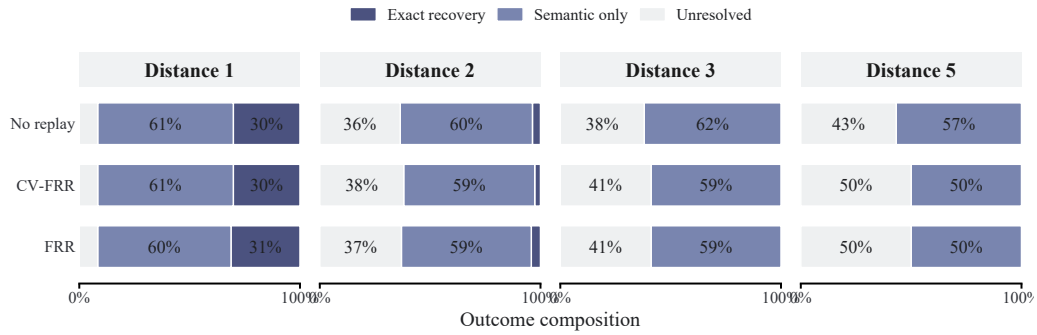

Figure S4: Outcome composition across the local repair radius.

E.1 Paired uncertainty and independent rollout audit

Task-paired bootstrap intervals confirm that the behavioral gains at AST distances two and three are not driven by unpaired task composition (Table S10). The intervals exclude zero for FRR, CV-FRR, and no replay. These estimates concern derivative-audit success; exact structural recovery remains concentrated at distances one and two.

Table S10: Paired semantic gains at AST distances two and three.

Distance	Method	Paired tasks	Semantic gain (pp)	95% CI
2	M4	192	6.77	[3.65, 10.42]
2	M9	192	5.73	[2.60, 9.38]
2	M10	192	7.29	[3.65, 10.94]
3	M4	192	7.81	[4.17, 11.98]
3	M9	192	7.81	[4.17, 11.98]
3	M10	192	10.94	[6.25, 15.63]

The free-rollout diagnostic integrates ordinary ODE candidates from two initial conditions unused by fitting, repair, or derivative auditing. A deterministic sampling rule selects eight tasks within each system and initial AST distance, giving 32 tasks and 64 trajectories per system and horizon. All integration records, including divergences, remain in the source data. Divergent trajectories have undefined NRMSE and are excluded from the finite-pair estimate in Table S11: Lorenz-63 contributes 64, 60, and 58 pairs across its three horizons, corresponding to 32, 31, and 29 task clusters. Van der Pol retains all 64 pairs at each horizon. Thus the negative intervals describe improvement among jointly integrable trajectories, not a reduction in divergence.

Table S11: Free rollout from unseen initial conditions.

System	Horizon	Task clusters	M4−M0 NRMSE	95% CI
Lorenz-63	0.5	32	-0.206	[-0.474, -0.028]
Lorenz-63	1.0	31	-0.261	[-0.592, -0.037]
Lorenz-63	2.0	29	-0.230	[-0.511, -0.046]
Van der Pol	2.0	32	-0.198	[-0.345, -0.067]
Van der Pol	5.0	32	-0.224	[-0.405, -0.075]
Van der Pol	10.0	32	-0.254	[-0.462, -0.081]

Table S12 restores the full trajectory denominators. Initial and repaired Lorenz models each diverge on 0, 4, and 6 trajectories; neither Van der Pol arm diverges. An additional failure-aware sensitivity analysis uses the bounded loss $e/(1+e)$ for finite NRMSE e and assigns loss one to divergence. It includes all 32 task clusters at every horizon, with 10,000 task-cluster bootstrap resamples. This descriptive sensitivity analysis uses a different estimand from the finite-pair NRMSE comparison.

Table S12: Full-denominator rollout audit and failure-aware sensitivity.

System	Horizon	Finite / total	Diverged (initial / FRR)	Δ bounded loss [95% CI]
Lorenz-63	0.5	64/64	0/0	-0.037 [-0.079, -0.003]
Lorenz-63	1	60/64	4/4	-0.039 [-0.075, -0.010]
Lorenz-63	2	58/64	6/6	-0.032 [-0.059, -0.009]
Van der Pol	2	64/64	0/0	-0.075 [-0.132, -0.026]
Van der Pol	5	64/64	0/0	-0.080 [-0.140, -0.028]
Van der Pol	10	64/64	0/0	-0.082 [-0.144, -0.030]

E.2 EXPLORATORY BASIN-ENTRY PRECONDITIONER

On the first 70% of each natural-candidate trajectory, terms below the calibrated coefficient-relevance threshold were removed, while retaining at least the largest-effect term per target. The rule never accessed the generating equation or final 30% audit block. Screening changes the initial candidate from 63/112 (56.25%) to 65/112 (58.04%) semantic success, from 4/112 (3.57%) to 11/112 (9.82%) exact recovery, and from 0.316 to 0.367 mean structural F1. Applying FRR to the screened candidates yields 70/112 (62.50%) semantic success, 33/112 (29.46%) exact recovery, and F1 0.378; M9 yields 78/112 (69.64%), 43/112 (38.39%), and F1 0.390, with no harmful edit in either arm (Table S13). Relative to the raw-candidate results above, these screened-arm endpoints are not uniformly higher, so the analysis is interpreted as descriptive evidence about basin entry rather than a guaranteed preconditioner or confirmatory repair result.

Table S13: Exploratory basin-entry audit.

Candidate / method	n	Semantic $(n/N, \%)$	Exact $(n/N, \%)$	Mean structural F1	Accepted edits	Harmful edits
Raw SINDy candidate	112	63/112 (56.25%)	4/112 (3.57%)	0.316	0	0
Coefficient-screened candidate	112	65/112 (58.04%)	11/112 (9.82%)	0.367	0	0
Screened + M4	112	70/112 (62.50%)	33/112 (29.46%)	0.378	22	0
Screened + M9	112	78/112 (69.64%)	43/112 (38.39%)	0.390	33	0

Replay interception subset. FRR intercepts 476 events in 124 of 768 runs (16.1%), with a median of three events per affected run (IQR 5). Within these 124 tasks, FRR minus CV-FRR/no replay is -12.9 percentage points in semantic success and $+0.209$ in audit NRMSE. This subset changes both the selector and replay status; the replay-only contrast instead pairs CV-FRR with its no-replay arm over all 768 tasks, as reported in the factorial analysis.

E.3 TERMINAL OUTCOMES AND RUN ACCOUNTING

Every method–family stratum receives one terminal category, including empty neighborhoods and contract failures. This separates an accepted edit from an explicit abstention or an unsuccessful repair path. Figure S5 and Table S14 provide the complete descriptive accounting. Agent-FRR contributes its 495-task paired comparison here, whereas the cost table uses its 200-task language-model subset; the two denominators are not combined.

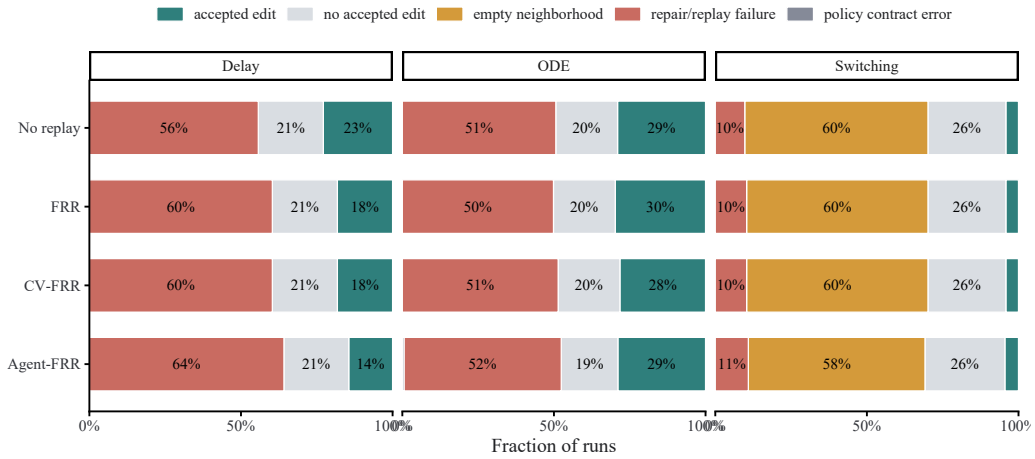

Figure S5: Terminal outcomes by method and dynamical family.

Table S14: Terminal outcomes in the completed repair study.

Method	Family	Accepted	No edit	Empty basin	Repair/replay failure	Contract error
No replay	ODE	188 (29.0%)	132 (20.4%)	0	328 (50.6%)	0
No replay	Switching	27 (4.2%)	166 (25.6%)	392 (60.5%)	63 (9.7%)	0
No replay	Delay	148 (22.8%)	139 (21.5%)	0	361 (55.7%)	0
Agent-FRR	ODE	48 (28.9%)	31 (18.7%)	0	86 (51.8%)	1 (0.6%)
Agent-FRR	Switching	7 (4.5%)	41 (26.3%)	91 (58.3%)	17 (10.9%)	0
Agent-FRR	Delay	25 (14.5%)	37 (21.4%)	0	111 (64.2%)	0
FRR	ODE	193 (29.8%)	132 (20.4%)	0	323 (49.8%)	0
FRR	Switching	27 (4.2%)	166 (25.6%)	388 (59.9%)	67 (10.3%)	0
FRR	Delay	118 (18.2%)	139 (21.5%)	0	391 (60.3%)	0
CV-FRR	ODE	183 (28.2%)	132 (20.4%)	0	333 (51.4%)	0
CV-FRR	Switching	27 (4.2%)	166 (25.6%)	388 (59.9%)	67 (10.3%)	0
CV-FRR	Delay	118 (18.2%)	139 (21.5%)	0	391 (60.3%)	0

For FRR, repair/replay failure accounts for 49.8% of ODE runs and 60.3% of delay runs, while 59.9% of switching runs terminate with an empty repair basin. These rates describe where the bounded controller stops without an accepted repair; they are not safety guarantees.

E.4 COST–QUALITY FRONTIER

The same analysis shows the quality–compute trade-off. Table S15 reports method-level semantic success, verifier calls, fits, and replay events. On the compound pool, FRR reaches 65.9% semantic success versus 68.5% for CV-FRR without replay, while generating 476 replay events across 124 runs. The replay arm therefore represents a preservation–accuracy trade-off, not a uniformly higher-quality operating point.

Table S15: Method-level cost and quality summary for the compound-error protocol.

Method	n	Semantic	Exact	Verifiers	Fits	Replay events	Intercepted runs
Initial	768	57.0%	0.0%	0.00	0.00	0	0
CV-FRR (no replay)	768	68.5%	8.5%	3.96	8.14	0	0
Agent-FRR	200	66.0%	9.5%	3.64	5.49	18	18
FRR	768	65.9%	8.9%	5.67	5.59	476	124
CV-FRR	768	65.6%	8.2%	5.72	7.32	457	133

F SELECTIVE DEPLOYMENT

Let C indicate that the supplied structure is correct under the target grammar and let $\delta = 1$ indicate an accepted structural edit. For cost ratio $\lambda > 0$, the edit-gate risk $\mathcal{R}_\lambda(\delta) = \Pr(\delta = 0, \neg C) + \lambda \Pr(\delta = 1, C)$ separates abstaining on an incorrect candidate from editing a correct one.

The selective score is the discovery-block improvement

$$s(h_0, h_f) = \mathrm{NRMSE}_{\mathrm{d}}(h_0) - \mathrm{NRMSE}_{\mathrm{d}}(h_f). \tag{19}$$

For each method and risk budget, calibration chooses the least restrictive threshold whose one-sided 95% binomial upper bound for editing a clean law is below that budget; the threshold is then applied once to untouched test tasks. At 5%, FRR has 84.5% calibrated coverage (188 accepted decisions on the 768-task test partition); coverage across the four policies ranges from 61.5% to 87.7%. At 10%, it ranges from 71.5% to 88.9%. Conditional semantic success spans 69.5–78.6%, conditional exact recovery 33.2–43.5% at 5% (33.2–44.2% at 10%), and observed false-repair rates 1.4–3.3%. Coverage follows the frozen calibration denominator, whereas Acted reports the number of accepted test decisions. The 1% point yields zero coverage because the calibration sample cannot resolve a nonzero operating point at that budget. The discrete score gives identical 5% and 10% thresholds for three arms.

Table S16: Selective outcomes at the 5% false-repair budget. Coverage uses the frozen calibration denominator; Acted is the number of accepted test decisions.

Method	Tasks	Acted	Coverage	Semantic|act	Exact|act	False repair
Agent-FRR†	200	23	61.5%	69.6%	**43.5%**	**1.4%**
FRR	768	188	84.5%	70.2%	36.2%	2.0%
CV-FRR	768	190	**87.7%**	69.5%	33.2%	2.3%
CV-FRR (no replay)	768	196	75.5%	**78.6%**	33.2%	3.3%

G PHYSIOLOGICAL-LAW INTERFACE STUDY

The physiology experiment evaluates whether the typed repair contract transfers to participant-level blood-pressure equations built from ECG/PPG-derived pulse-arrival features and clinical covariates. MIMIC-III uses 279 held-out participants and VitalDB uses 144; processed windows contain 10 seconds at 125 Hz with ECG Lead II, PPG/Pleth, and ABP channels. Participant counts, windows, and records remain distinct units. Twenty corruption–seed tasks are evaluated for each dataset and target. The study tests interface portability rather than clinical performance.

The executable family is a ridge-refitted additive law

$$\widehat{\mathrm{BP}} = \beta_0 + \sum_{q \in \mathcal{S}} \beta_q \phi_q, \quad \phi_q \in \{\mathrm{PAT}, \mathrm{PAT}^{-1}, \mathrm{PAT}^{-2}, \mathrm{PAT}{\times}\mathrm{HR}, \mathrm{HR}, \mathrm{age}, \mathrm{BMI}, \ldots\}. \tag{20}$$

Residual–feature correlation ranks absent terms, coefficient magnitude ranks removable terms, and participant-paired bootstrap improvement on a discovery-only validation split determines acceptance. For example, the seed-401 SBP task omits PAT^{-2}. The residual witness proposes adding that term; its validation MAE improvement is 0.0724 mmHg, and the accepted child exactly restores the reference term set. The subsequent audit is performed on held-out participants from both cohorts. This mapping preserves the same witness–typed action–refit–accept contract used by the dynamical benchmarks, although the endpoint is participant-level prediction rather than trajectory dynamics.

Table S17: Participant-level physiology repair.

Cohort	Target	Participants	Initial	Repaired	Reduction	Improved	Exact
MIMIC-III	SBP	279	15.912	15.894	0.0175	25%	20%
MIMIC-III	DBP	279	9.441	9.440	0.0014	15%	0%
VitalDB	SBP	144	14.598	14.574	0.0239	25%	20%
VitalDB	DBP	144	9.313	9.312	0.0012	10%	0%

H CONFIRMATORY STATISTICS

The cluster unit is a complete external task, preserving paired seeds, systems, noise levels, and corruptions. We use 10,000 recentered task-cluster bootstrap replicates with seed 2027. The non-inferiority margin is two percentage points. Holm correction is applied jointly to all three hypotheses at family-wise $\alpha = 0.05$.

Table S18: Prespecified confirmatory contrasts.

Contrast	Effect	CI low	CI high	p_H	Decision
M9−M4 semantic non-inferiority (margin -0.02)	-0.00595	-0.04167	0.02976	0.394	inconclusive
M9 versus M4 verifier-call reduction	-0.09167	-0.18045	-0.01471	0.978	inconclusive
M9−M3 semantic superiority	0.04167	-0.00595	0.08929	0.163	inconclusive

For H2, the negative effect indicates that M9 uses more verifier calls. Its interval excludes zero opposite to the prespecified one-sided reduction hypothesis, producing the large corrected p-value.

I NULL AND OUT-OF-DISTRIBUTION CONTROLS

Correct-model null tasks measure false repair and false support. Phase-randomized and fixed distribution shifts test whether an agent repairs nonspecific residual structure. The 200 null-control tasks retain all failures in their denominators. Figure S6 reports the complete controls.

Across the boundary analysis, 126 non-oracle null runs contain a false repair. The first failed verifier is coefficient relevance in 83 runs, residual–feature correlation in 31, normalized derivative fit in 6, and regime-boundary profile in 6.

The 0.99 calibration quantile applies to each verifier in isolation. A run may test several verifiers over several rounds, refit coefficients, and select among multiple admissible edits, so the run-level false-repair probability is not controlled at 1%. The correct-law null rates are 15% for FRR and 9% for the phase-null control, showing why the marginal thresholds should not be described as a general safe guarantee. The selective study fixes an abstention threshold on a separate calibration partition before test evaluation (Table S16). On this test distribution, observed clean-law false repair remains below the 5% budget; the result is not distribution-free sequential error control.

An exhaustive diagnostic applies every compatible verifier to all 180 initial misspecified equations and their corresponding correct equations. Coefficient relevance detects 91.7% of add-term and 85.0% of replace-term corruptions but 40.0% of omissions; residual–feature correlation detects 73.3% of omissions and 70.0% of replacements. Regime occupancy detects every threshold corruption in its applicable subset. On correct equations, coefficient relevance fires in 3/180 cases (1.67%); every other verifier has zero observed false positives in its applicable set. Because the boundary decomposition and exhaustive diagnostic were specified after the confirmatory study, their Wilson intervals are interpreted descriptively and did not determine thresholds or confirmatory decisions.

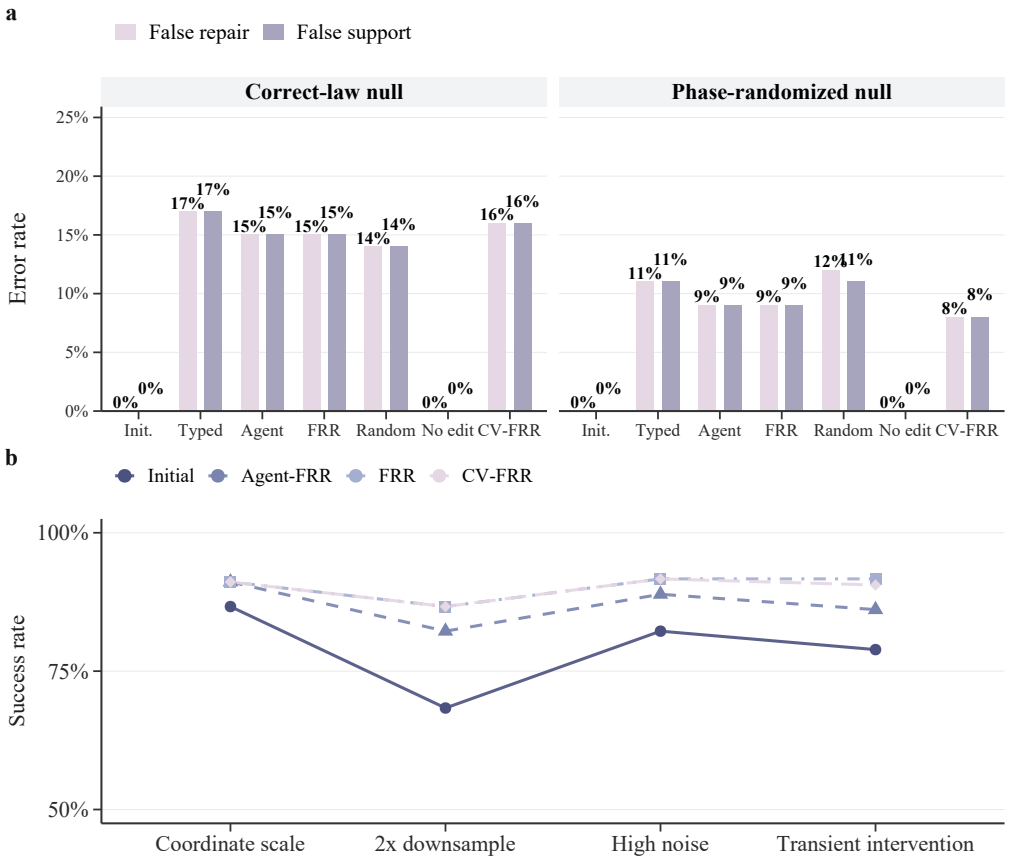

Figure S6: Negative controls and OOD stress tests.

J INDEPENDENT CROSS-FAMILY BEHAVIOR AND AGENT ATTRIBUTION

J.1 SWITCHING AND DELAY ROLLOUTS

The ordinary-ODE rollout analysis is extended to two switching and two delayed systems. Initial states or history functions are held out from fitting, witness construction, repair selection, and derivative auditing. Every eligible task is retained, including divergent trajectories, and the outcome is evaluated only after the frozen repair has terminated. Table S19 reports paired task-cluster bootstrap intervals for M4 minus the unedited candidate; negative values favor repair.

Short- and medium-horizon trajectory and terminal error decrease for the canonical delay system, without increased divergence, whereas its longest horizon is uncertain: at $4\times$, trajectory and terminal changes are $+1.254$ and $+1.824$ with intervals spanning zero. Delayed Stuart–Landau changes are effectively zero across horizons. The switching systems show smaller non-positive trajectory, terminal, and transition-timing estimates. These mechanism-specific endpoints extend the audit beyond derivative fit while locating the remaining difficulty in long-horizon delayed behavior.

Table S19: Independent switching and delay rollouts.

System	Horizon	n	Δ trajectory	95% CI	Δ terminal	Mechanism Δ
Delay	$1\times$	32	$\mathbf{-0.514}$	$[-0.970, -0.136]$	-0.026	0.0 pp div.
Delay	$2\times$	32	$\mathbf{-0.826}$	$[-1.823, -0.121]$	-0.122	0.0 pp div.
Delay	$4\times$	32	$+1.254$	$[-0.425, +4.421]$	$+1.824$	-3.1 pp div.
Delayed S–L	$1\times$	32	-0.0004	$[-0.0015, +0.0001]$	-0.0002	0.0 pp div.
Delayed S–L	$2\times$	32	-0.0006	$[-0.0021, +0.0004]$	-0.0005	0.0 pp div.
Delayed S–L	$4\times$	32	-0.0006	$[-0.0020, +0.0003]$	-0.0004	0.0 pp div.
Switching linear	$2\times$	32	-0.0067	$[-0.0202, 0.0000]$	-0.0051	-0.0009 timing
Switching linear	$5\times$	32	-0.0050	$[-0.0150, 0.0000]$	-0.0070	-0.0014 timing
Switching linear	$10\times$	32	-0.0078	$[-0.0233, 0.0000]$	-0.0071	-0.0017 timing
Switching VdP	$2\times$	32	-0.0387	$[-0.0969, 0.0000]$	-0.0626	-0.0450 timing
Switching VdP	$5\times$	32	-0.0447	$[-0.1119, 0.0000]$	-0.1583	-0.0569 timing
Switching VdP	$10\times$	32	-0.0806	$[-0.2017, 0.0000]$	-0.1456	-0.1535 timing

The displayed interval is for trajectory NRMSE. For Delay at $4\times$, the terminal-error change is $+1.824$ with 95% CI $[-0.082, +5.589]$; both intervals span zero.

J.2 CONSTRAINED AGENT-SELECTOR ATTRIBUTION

To isolate the role of the language-model selector, M3 and M4 are paired on all 495 common controlled and compound tasks. The comparison is made at the action-sequence and act/abstain levels before examining outcomes; no additional model calls are made. The selectors agree on 96.8% of action sequences and 98.8% of decisions, while schema or contract errors occur on 0.2% of tasks. Among the 16 action disagreements, twelve end in an outcome tie and four favor M4; none favor M3. The language-model selector usually chooses the same actions as the deterministic controller; both use identical numerical fitting and acceptance rules.

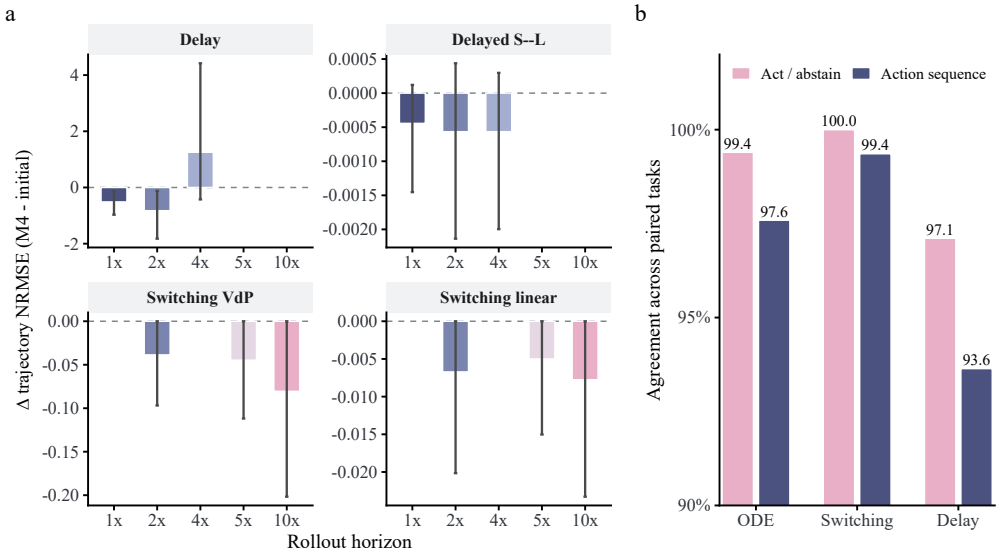

Figure S7: Independent rollouts and selector agreement. Panel b is shown over the 90–100% range for readability; labels above the bars report the exact percentages.

Table S20: M3–M4 selector attribution.

Stratum	n	Action agree	Decision agree	M3 semantic	M4 semantic	Contract error
All paired tasks	495	**96.8%**	**98.8%**	74.7%	75.6%	0.2%
ODE	166	97.6%	99.4%	66.3%	66.9%	0.6%
Switching	156	99.4%	100.0%	73.7%	73.7%	0.0%
Delay	173	93.6%	97.1%	83.8%	85.5%	0.0%

The sixteen disagreements comprise four ODE tasks, one switching task, and eleven delay tasks. FRR wins one ODE and three delay comparisons; the other twelve are ties. Thus the disagreements do not explain a performance advantage for the language-model selector.

J.3 CALIBRATION OPERATING POINTS

Calibration produces discrete operating points rather than a continuous risk curve. Table S21 reports coverage, conditional outcomes, and observed clean-law false repair at the prespecified 1%, 5%, and 10% budgets. A zero-coverage point is retained when calibration cannot resolve a nonzero threshold.

Table S21: Calibration operating points on the untouched test partition. Coverage uses the frozen calibration denominator; Acted n is the number of accepted test decisions.

Method	Budget	Coverage	Acted n	Semantic|act	Exact|act	False repair
Agent-FRR	1%	0.0%	0	–	–	0.0%
Agent-FRR	5%	61.5%	23	69.6%	43.5%	1.4%
Agent-FRR	10%	71.5%	43	72.1%	44.2%	2.8%
FRR	1%	0.0%	0	–	–	0.0%
FRR	5%	84.5%	188	70.2%	36.2%	2.0%
FRR	10%	87.7%	188	70.2%	36.2%	2.0%
CV-FRR	1%	0.0%	0	–	–	0.0%
CV-FRR	5%	87.7%	190	69.5%	33.2%	2.3%
CV-FRR	10%	88.9%	190	69.5%	33.2%	2.3%
CV-FRR (no replay)	1%	0.0%	0	–	–	0.0%
CV-FRR (no replay)	5%	75.5%	196	78.6%	33.2%	3.3%
CV-FRR (no replay)	10%	85.3%	196	78.6%	33.2%	3.3%

K MATCHED RESIDUAL-REFINEMENT BASELINE

We evaluated the residual-refinement selector (M7, a RESTART-style baseline) on the same 224 natural discovery candidates used by FRR (M4), with identical typed actions, verifier thresholds and four-round budget. The study contains 224 task pairs (448 runs); 24 M7 calls ended in schema or action-contract errors and remain in the denominator. Semantic success was 55.8% for both methods and exact recovery was 3.6% for both; M7 used 2.02 verifier calls per run versus 2.36 for M4. No accepted repair occurred in either arm. M7 reduces verifier calls on this cohort, but neither method changes the candidate structures; the result supports interface substitutability, not an FRR advantage over the matched baseline.

Table S22: Matched residual-refinement comparison.

Method	Runs	Semantic (%)	Exact (%)	Accepted edits	Verifier calls	API calls	Contract errors
FRR (M4)	224	55.8	3.6	0	2.36	0	0
RESTART-style (M7)	224	55.8	3.6	0	2.02	208	24

L LANGUAGE-MODEL INTERFACE AND COMPUTE AUDIT

Table S23: Compute and reliability summary.

Method	Verifiers	Fits	API	Anytime AUC	Stable	Correct-law FR	Phase-null FR
Rewrite (M1)	3.1	9.8	0.91	9.0	96.7	–	–
Typed/no replay (M2)	4.2	9.6	0.76	36.9	96.7	17.0	11.0
Agent-FRR (M3)	4.7	7.7	0.76	31.0	100.0	15.0	9.0
FRR (M4)	4.8	8.2	0.00	33.2	100.0	15.0	9.0
Random-FRR (M5)	5.4	18.3	0.84	27.8	100.0	14.0	12.0
CV-FRR (M9)	4.7	9.1	0.00	32.2	99.4	16.0	8.0

Table S24: Language-model selector decisions.

Status	Calls	Fraction
Live success	1,005	39.1%
Cache hit	1,392	54.2%
Live contract error	94	3.7%
Cached contract error	78	3.0%
Total	2,569	100%

The compound study records 295 M3 decisions: 260 successful live responses, 34 cache hits, and one schema-contract error. The error remains in the task denominator and produces no executable action. Token, latency, verifier-call, and parameter-fit measurements are recorded separately.

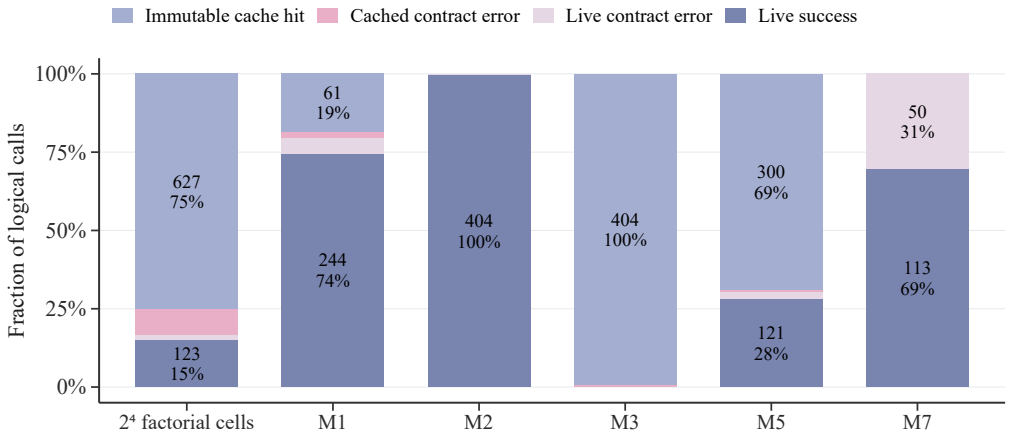

Figure S8: Language-model outcomes by execution group.

The language-model request contains the symbolic candidate, a compact falsification record, finite action identifiers, and scalar refit scores; it excludes time-series arrays and source identifiers. The constrained output schema accepts one known action identifier, while M1 must parse into the restricted grammar. Responses that fail schema or model-identity checks produce no action.

M METHOD IDENTIFICATION AND SEARCH COST

M.1 LOCAL-SEARCH COMPARISON

The factorial experiment establishes the value of typed editing; the comparisons here ask what diagnosis adds to a generic search over legal edits. We evaluate every original one-edit task ($n = 180$) and compound task ($n = 768$) under a separate, frozen method-identification protocol, keeping their generators, noise levels, seeds, thresholds, and task-specific inputs unchanged. These additional analyses are exploratory and use their own stated fit ceilings: 64 for one-edit tasks and 192 for compound tasks, with at most 32 candidates per step. The confirmatory compound result in the main paper uses a distinct 64-fit/8-verifier-call cap; it should not be conflated with this exploratory 192-fit sweep. Deterministic FRR uses the original implementation, but the protocol and resource ceiling determine the reported aggregate.

Validation-ranked local search enumerates legal additions, removals, replacements, and applicable lag or boundary changes without a witness. Its term library and edit operators match the FRR implementation. Within each operator, candidates are shuffled using the task seed and then interleaved across operators before truncation, so additions cannot consume the entire candidate allowance. Lag values follow the original data-length-dependent grid; boundary candidates use repair-data quantiles. Neither grid uses the generating equation. Residual-guided search changes this ordering using residual–feature projections for additions and replacements, and coefficient contributions for removals. It does not use typed localization or specialized lag/boundary witnesses.

Both searches fit coefficients on the first 70% of the repair block and rank candidates by derivative NRMSE on the remaining 30%. An edit is accepted only if it improves this validation score by more than 10^{-8}; the selected structure is then refitted on the full repair block. Candidate fits and accepted refits share the fit ceiling. FRR retains its witness-directed enumeration, original complexity-penalized ranking, and numerical acceptance tests. Thus the end-to-end comparison evaluates complete decision procedures under common resource ceilings; the shared-pool analysis below separates acceptance from candidate allocation. Verifier calls and realized fits are reported separately, with initialization and post-run evaluation excluded from adaptation cost under the original accounting convention.

Tables S25 and S26 reveal a structural–behavioral trade-off. On one-edit tasks, FRR reconstructs 39.4% of generating laws with 8.24 fits on average; generic and residual-guided search reconstruct 1.1% and 2.8% with approximately 64 fits. The searches attain higher semantic success in that one-edit setting, showing that validation improvement need not select the generating structure. Under the method-identification protocol, the compound comparison reports 95.9% semantic success and 8.9% exact recovery with 5.59 fits and zero tests, compared with 85.7%/0.0% and 90.0%/3.9% for local and residual-guided search, respectively. FRR records 99.8% acted tasks and no harmful edits in this table. Harm denotes loss of initial semantic success, not any increase in continuous audit error. Acted denotes at least one accepted structural edit.

Table S25: One-edit local-search comparison under the method-identification protocol (64-fit ceiling).

Method	n	Semantic (%)	Exact (%)	Acted (%)	Harm (%)	Fits	Tests
Initial (refitted)	180	65.0	0.0	0.0	0.0	0.00	0.00
Local search	180	92.8	1.1	100.0	0.6	63.98	0.00
Residual-guided	180	95.0	2.8	97.2	0.0	63.02	0.00
FRR	180	85.6	39.4	51.1	0.0	8.24	4.81

Table S26: Compound local-search comparison under the method-identification protocol (192-fit ceiling; $n = 768$ per row).

Method	n	Semantic (%)	Exact (%)	Acted (%)	Harm (%)	Fits	Tests
Initial (refitted)	768	57.0	0.0	0.0	0.0	0.00	0.00
Local search	768	85.7	0.0	99.7	0.3	174.98	0.00
Residual-guided	768	90.0	3.9	99.5	0.0	156.45	0.00
FRR	768	95.9	8.9	99.8	0.0	5.59	0.00

M.2 PAIRED EFFECTS AND FIT-BUDGET SENSITIVITY

Table S27 quantifies the full-budget differences using 10,000 paired bootstrap resamples. Resampling groups all noise conditions and corruptions sharing a system and seed, giving 30 clusters for one-edit tasks and 24 for compound tasks. Intervals describe variation across these task blocks within the fixed systems. Positive semantic and exact effects favor FRR; negative fit and harm effects favor FRR. These exploratory intervals do not change the original confirmatory hypotheses or multiplicity families.

Table S27: Full-budget paired differences: FRR minus each search baseline.

Setting	Comparator	Endpoint	Δ FRR	95% interval
One edit	Local search	Semantic (pp)	-7.22	[-13.33, -1.67]
One edit	Local search	Exact (pp)	+38.33	[+28.33, +48.89]
One edit	Local search	Harm (pp)	-0.56	[-1.67, +0.00]
One edit	Local search	Fits	-55.74	[-57.94, -53.28]
One edit	Residual-guided	Semantic (pp)	-9.44	[-15.56, -4.44]
One edit	Residual-guided	Exact (pp)	+36.67	[+26.67, +47.78]
One edit	Residual-guided	Harm (pp)	+0.00	[+0.00, +0.00]
One edit	Residual-guided	Fits	-54.78	[-56.99, -52.36]
Compound	Local search	Semantic (pp)	-19.79	[-24.22, -15.62]
Compound	Local search	Exact (pp)	+8.85	[+5.47, +12.50]
Compound	Local search	Harm (pp)	-0.13	[-0.78, +0.39]
Compound	Local search	Fits	-169.39	[-180.58, -156.20]
Compound	Residual-guided	Semantic (pp)	-24.09	[-28.78, -19.53]
Compound	Residual-guided	Exact (pp)	+4.95	[-0.52, +10.29]
Compound	Residual-guided	Harm (pp)	+0.13	[+0.00, +0.39]
Compound	Residual-guided	Fits	-150.86	[-162.13, -139.14]

The budget ablation repeats all tasks at fit ceilings 8, 16, 32, and 64, adding 128 and 192 for compound tasks. Only the fit ceiling changes; the original maximum-budget runs are reused. Table S28 shows that FRR's 39.4% exact recovery and 85.6% semantic success persist across the entire one-edit grid. At the eight-fit ceiling it uses 3.80 fits on average, while the two searches use about 7.8. This supports efficient local reconstruction rather than a need for extensive candidate fitting. Residual-guided search remains stronger on semantic success.

Table S28: Complete one-edit fit-budget ablation ($n = 180$ per row).

Ceiling	Method	Sem. (%)	Exact (%)	Harm (%)	Fits	Tests
8	Local search	76.7	7.2	0.6	7.87	0.00
8	Residual-guided	90.0	11.7	0.0	7.76	0.00
8	FRR	85.6	39.4	0.0	3.80	4.77
16	Local search	83.3	9.4	0.6	15.97	0.00
16	Residual-guided	95.0	21.1	0.0	15.92	0.00
16	FRR	85.6	39.4	0.0	6.20	4.81
32	Local search	87.8	11.7	0.6	32.00	0.00
32	Residual-guided	95.0	22.8	0.0	31.97	0.00
32	FRR	85.6	39.4	0.0	8.24	4.81
64	Local search	92.8	1.1	0.6	63.98	0.00
64	Residual-guided	95.0	2.8	0.0	63.02	0.00
64	FRR	85.6	39.4	0.0	8.24	4.81

M.3 COMPOUND BUDGET DEPENDENCE AND COMMON-POOL DECISIONS

With compound errors, FRR uses 2.7–5.6 fits and zero tests across the budget grid, while semantic success ranges from 78.8% to 95.9% and exact recovery from 0.0% to 9.1% (Table S29). The endpoint reaches 95.9% at the 192-fit ceiling; residual-guided search has higher semantic success at the lower compound budgets, whereas FRR uses substantially fewer fits. These figures should be read as a budget-sensitivity description, not as a monotone optimization law. Fits and tests are reported separately and neither should be interpreted as wall-clock cost.

Table S29: Complete compound fit-budget ablation under the method-identification protocol ($n = 768$ per row).

Ceiling	Method	Sem. (%)	Exact (%)	Harm (%)	Fits	Tests
8	Local search	66.1	1.6	0.4	7.90	0.00
8	Residual-guided	72.7	6.8	0.1	7.88	0.00
8	FRR	78.8	9.0	0.0	2.71	0.00
16	Local search	71.2	3.1	0.4	15.98	0.00
16	Residual-guided	76.4	9.2	0.0	15.96	0.00
16	FRR	65.9	9.1	0.0	4.19	0.00
32	Local search	74.9	6.5	0.4	31.99	0.00
32	Residual-guided	77.0	9.0	0.0	31.99	0.00
32	FRR	79.6	8.9	0.0	5.52	0.00
64	Local search	80.6	1.7	0.4	63.88	0.00
64	Residual-guided	84.9	6.0	0.0	63.75	0.00
64	FRR	85.6	8.9	0.0	5.59	0.00
128	Local search	85.0	0.0	0.3	123.34	0.00
128	Residual-guided	89.3	3.5	0.0	118.78	0.00
128	FRR	90.2	0.0	0.0	5.59	0.00
192	Local search	85.7	0.0	0.3	174.98	0.00
192	Residual-guided	90.0	3.9	0.0	156.45	0.00
192	FRR	95.9	8.9	0.0	5.59	0.00

To isolate acceptance conditional on the candidate pool, we also evaluate two one-step decisions on an identical generic, witness-blind pool of at most 32 fitted candidates. Both use the same validation ordering. The validation rule takes the best strict improvement; the FRR gate searches the first eight ranked candidates for resolution of the first failed verifier and preservation of earlier passes, subject to the original verifier ceiling. Both return coefficients refitted on the full repair block. This is an acceptance ablation, not an end-to-end replacement for FRR.

The common-pool comparison isolates the acceptance decision while holding the candidate pool and validation ordering fixed (Table S30). FRR improves semantic success from 88.3% to 90.3% on one-edit tasks and from 75.4% to 85.2% on compound tasks; exact recovery rises by 4.4 and 2.6 percentage points, respectively. Both settings reach 100% acted decisions under the revised gate and record zero harmful edits, compared with 0.6% and 0.4% under validation. The model-update arm is different: FRR abstains on all 80 tasks, so its 100% semantic and 12.5% exact rates reflect the initial candidates rather than accepted repairs. The comparison therefore supports a gate-level preservation and recovery advantage on repair tasks, but not a claim that FRR improves every update setting.

Table S30: One-step decisions on identical fitted candidate pools.

Setting	Decision	n	Semantic (%)	Exact (%)	Acted (%)	Harm (%)
One edit	Validation gate	180	88.3	12.8	100.0	0.6
One edit	FRR gate	180	90.3	17.2	100.0	0.0
Compound	Validation gate	768	75.4	7.0	99.7	0.4
Compound	FRR gate	768	85.2	9.6	100.0	0.0
Model update	Validation gate	80	100.0	0.0	100.0	0.0
Model update	FRR gate	80	100.0	12.5	0.0	0.0

M.4 DISCOVERY-GENERATED MODEL UPDATES

We next ask whether a data update naturally produces a candidate needing local structural repair. STLSQ and SR3 use the unchanged polynomial-degree-three discovery configurations to fit five-second trajectories from initial states scaled to 0.25 times those of the original Van der Pol and Lorenz-63 generators. New observations comprise a fourteen-second repair trajectory from the original initial state and an independent six-second audit trajectory from 1.5 times that state, sampled every 0.02 time units. Two noise levels (0 and 0.05), ten seeds (601, 607, 613, 617, 619, 631, 641, 643, 647, 653), two systems, and two optimizers give 80 candidates. Every candidate is retained, without filtering by truth distance or outcome. The generating dynamics remain fixed: this is a simulated discovery-and-update workflow.

Table S31 gives the initial structural distribution. Noise-free trajectories repeat across seeds because the initial states are fixed; these are not ten distinct dynamical realizations. Initial quality is measured after coefficient-only fitting on the new repair data, so structural repair receives no credit for parameter adaptation. This refit already brings every candidate within the semantic-success threshold, although only ten candidates have the generating structure.

Table S31: All discovery-generated update candidates before structural repair.

System	Noise	Discovery	n	AST median [range]	Semantic (%)	Exact (%)
Lorenz-63	0	SR3	10	53 [53, 53]	100.0	0.0
Lorenz-63	0	STLSQ	10	53 [53, 53]	100.0	0.0
Lorenz-63	0.05	SR3	10	53 [53, 53]	100.0	0.0
Lorenz-63	0.05	STLSQ	10	53 [53, 53]	100.0	0.0
Van der Pol	0	SR3	10	0 [0, 0]	100.0	100.0
Van der Pol	0	STLSQ	10	15 [15, 15]	100.0	0.0
Van der Pol	0.05	SR3	10	16 [15, 17]	100.0	0.0
Van der Pol	0.05	STLSQ	10	16 [15, 16]	100.0	0.0

FRR accepts no structural edits and retains all ten exact models (Table S32). Both local searches edit every candidate: semantic success remains 100%, but exact recovery falls to zero. The result illustrates restraint when coefficient adaptation suffices; it supplies no evidence of successful structural repair after a natural data update. In this workflow, the semantic endpoint has reached its ceiling before repair, and the distant candidates remain outside an effective repair path.

Table S32: Model-update outcomes after coefficient adaptation ($n = 80$).

Method	n	Semantic (%)	Exact (%)	Acted (%)	Harm (%)	Fits	Tests
Initial (refitted)	80	100.0	12.5	0.0	0.0	0.00	0.00
Local search	80	100.0	0.0	100.0	0.0	63.67	0.00
Residual-guided	80	100.0	0.0	100.0	0.0	63.70	0.00
FRR	80	100.0	12.5	0.0	0.0	0.88	2.25