DOES LLM DREAM OF DIFFERENTIAL EQUATION DISCOVERY?

Anonymous authors
Paper under double-blind review

ABSTRACT

Discovering a governing equation from data means choosing its terms and estimating their coefficients. We ask whether a general-purpose multimodal model, given no feedback, can do both itself, with no library of candidate terms and no optimiser fitting the coefficients. Four such models are shown ten dynamical systems in twelve ways, each run returning five candidate equations, and a run discovers the equation when a screened candidate has exactly the true terms and every coefficient within 10% of its value. Each model is also run with the data withheld, so that what it reads from the data can be told from what it recalls. In its strongest configuration GLM discovers ten laws, Qwen and DeepSeek eight and Gemma four, including laws none of the models writes without data, with coefficients they estimate themselves, and their reasoning shows how: they read values where competing terms vanish, and GLM tests what it reads. Under added noise the models seldom say that the data are noisy, yet on several systems they keep writing the true law. The hardest laws are missed because one of their less common terms is never written.

1 INTRODUCTION

A governing differential equation couples a discrete choice, which terms enter it, with a continuous one, the coefficient of each. Its terms are built from derivatives such as u_t or u_{xx}, which the data do not contain and which have to be computed from them. Classical discovery fixes a library of candidate terms and selects among them by sparse regression (Brunton et al., 2016; Rudy et al., 2017) or its weak form (Messenger & Bortz, 2021; Stephany & Earls, 2024), or searches expressions built from fixed operators by evolutionary or reinforcement-learning search (Maslyaev et al., 2021; Chen et al., 2022; Du et al., 2024b). In each of these, a fit to the data sets the coefficients; ODEFormer, trained for the task, writes an ordinary differential equation with its constants in one pass (d'Ascoli et al., 2023).

A language model can write a term that no prepared list contains, and it can be shown the data as numbers or as a plot. Where language models have been used to discover differential equations, they propose the terms and something else sets the coefficients: LLM-SR has the model write an equation skeleton as a program whose parameters an optimiser fits (Shojaee et al., 2025a), LLM4ED has the model generate and mutate equations that a numerical optimiser fits (Du et al., 2024a), a differentiable simulation fits the hypotheses a model proposes (Ma et al., 2024), and where the coefficients a model wrote for a partial differential equation were scored, it revised them against a solver (Luo et al., 2025).

We study the task with the model alone. A general-purpose multimodal model receives samples of a field, derivative fields estimated from them, or a plot, and writes the whole equation itself, which derivative terms enter it and the coefficient of each, with no library of candidate terms, no optimiser and no feedback on its answer. Such a model has read many of the laws it may be asked to find, and recall is a known problem in evaluating language models on equation discovery: benchmarks rewrite known laws into less common forms (Shojaee et al., 2025b) or alter them until no memorised form applies (Zheng et al., 2026; Chen et al., 2026), removing recall by construction. We measure it instead: each model is also run with the data withheld, and every answer is read against what it writes there. Four models are run on ten dynamical systems, varying what each is shown and how it is asked; each run returns five candidate equations, screened without the ground truth and scored against it. Further experiments fit a sparse regression to the same points, with a general library and with one that holds every true term, alter the prompt, and add observation noise.

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.

Contributions.

- **A first measurement.** To our knowledge, the first study of whether general-purpose models can discover differential equations on their own, terms and coefficients, read against a control for what each model recalls: the same model, asked again with the data withheld (§3).
- **What the models can and cannot do, and why.** We identify how discovery depends on the input a model is given, which models discover which laws, including laws they cannot recall, how their reasoning reaches the coefficients, and why the hardest laws are missed (§4).
- **Selection without the ground truth.** We screen candidate equations by two objectives computed from the data alone and show that together they keep nearly every discovery, where either alone loses some (§2, §4.2).

2 Problem Statement

An observation is a state variable u recorded on a grid of independent coordinates: time t for an ordinary differential equation, time and one spatial coordinate x for a partial one. Let $\mathcal{T}$ be a pool of candidate terms, each a product of the state, its partial derivatives such as u_t, u_{tt}, u_x or u_{xx}, their powers and functions of the coordinates, as in $u\,u_x$ or $u_t \sin 2t$. A governing equation is a relation among a subset of them,

$$\sum_{s\in S} p_s\, s \;=\; 0, \qquad S \subset \mathcal{T}, \tag{1}$$

in which one term of S, normally a time derivative, is written on the left with coefficient one, as in Viscous Burgers, $u_t = -u\,u_x + 0.1\,u_{xx}$. A solution is a pair, the active set S and the coefficients $P = \{p_s\}$, and three properties of the task shape the study.

The pool is open. Nothing fixes $\mathcal{T}$ in advance, as a library would: the model writes whatever terms it proposes from the operands of a grammar, and the pool of a run is the terms its candidates contain. A run can fail before any candidate is judged, by never writing a term the law needs, and no screen can keep a candidate that was never written. An open pool also admits memory: a model can write a law it knows whatever the data show, and whether the data or its memory produced the law shows only against the same model asked without the data (§4.3).

The derivatives are estimated. The data hold u, not its derivatives. Every derivative term is computed from the samples by some numerical scheme, and there are infinitely many such schemes, each with its own error, which grows with the order of the derivative and with noise in the data. A coefficient estimated from the computed fields equals the true one only up to a factor set by that error: no numerical estimator returns the true coefficients exactly, and the ideal equation differs from the best one the data allow.

The answer cannot be evaluated against the truth. Evaluating a candidate by solving it and comparing the solution with the data is itself a hard problem: it needs initial and boundary conditions and a numerical scheme that stays stable for whatever was written, and no single procedure does this for every equation a model may propose. Candidates are judged instead by two objectives taken from the literature and combined here, each computed on the data alone. With y and $\hat{y}$ the two sides of a candidate evaluated on the computed fields, and $\hat{a}_{k,w}$ the coefficient of its k-th term refitted on the w-th of W sliding windows along the d grid axes, with mean $\bar{a}_k$,

$$\text{Discrepancy} = \frac{\sum_i |y_i - \hat{y}_i|}{\sum_i |y_i|}, \qquad \text{Instability} = \frac{1}{d}\sum_k \frac{1}{W-1}\sum_{w=1}^{W}\left(\frac{\hat{a}_{k,w} - \bar{a}_k}{\bar{a}_k}\right)^2. \tag{2}$$

Discrepancy is scale-free (Hyndman & Koehler, 2006); Instability, adapted from Xu et al. (2023), asks whether the coefficients that fit one sub-domain fit the others. A candidate survives when no other candidate of the same answer is at least as good on both and better on one, and the survivors form the answer's *screened set*. The two objectives look at different things: Discrepancy judges the coefficients a candidate writes, while Instability refits them and asks whether its terms admit one set of coefficients across the domain. The screened set can hold more than one candidate; the screen narrows an answer without committing to a single equation, and a run is judged by the best candidate its screened set holds (§3.4).

3 MATERIALS AND METHODS

3.1 DATA

The suite holds solutions of ten known laws: two ordinary systems, the Van der Pol oscillator and the Forced damped oscillator, and eight partial systems covering wave propagation, shock formation, reaction and diffusion, nonlinear diffusion and chaotic spatiotemporal dynamics. A prompt can carry three kinds of evidence: numerical samples of the field, about two thousand points of an evenly spaced subgrid, or the whole series of an ordinary system; derivative fields estimated by finite differences; and a rendered plot of the field, or of the field and its derivatives, with nothing that names the system. More details are in Appendix A.

3.2 MODELS AND PROTOCOL

Four open-weight multimodal models are queried through OpenRouter: `glm-5.3-flash`, `qwen3.8-flash`, `deepseek-v4-flash-vision-exp` and `gemma-4-31b-it`, called GLM, Qwen, DeepSeek and Gemma below. The system prompt gives an equation grammar and a worked example and asks for reasoning in four steps, from the behaviour of the data to the coefficient estimates, followed by five candidate equations with numeric coefficients. No run receives feedback: the model is never shown a score of its answer, a solution of its equation or a request to improve it. The task is asked in one call or, in the two-call configurations, in two, the second asking for the coefficients of the structures from the first; every call is made at temperature 0.5 with no tools. A reply cut at the provider's length limit counts as a run with no answer. More details are in Appendix B.

3.3 EXPERIMENTS

The *main sweep* crosses what the model is shown, the samples, the plot or both, with whether derivative fields are supplied and whether the task is asked in one call or two, twelve configurations in all (Figure 1). A *cell* is 10 runs of one model on one system in one configuration. A model's *strongest cells* are those of the configuration whose ten cells hold the most discovering runs, and the other experiments run on them.

The arms that follow the sweep start with the *data-withheld arm*, a check on recall run in one call: it removes the samples and the plot and keeps the grid description and the type of equation, from which no law can be derived. A model *recalls a structure* when it recovers the structure there (§3.4), and *recalls a law* when it discovers the equation there. As a comparison, sparse regression (Brunton et al., 2016; Rudy et al., 2017) is fitted on the points the prompt carries, with a general library that, like the models, is not told the law, and, where no setting of it recovers the law, with a predefined library that holds every true term. The *prompt arms* each change one component of the prompt: the grammar is cut to an operand list or deleted, and the worked example is made abstract, replaced by another system's law, or replaced by the system's own law. The *noise arm* adds Gaussian noise to the stored field before anything is differentiated or sampled, at three levels per system: half, once and twice the system's *noise limit*, the level beyond which sparse regression on the same points no longer recovers the structure. The reasoning in every reply is read for the moves it makes. More details are in Appendix C.

3.4 EVALUATION

The objectives of §2 use the finite-difference fields over the cropped grid; each surviving candidate is then compared with the ground truth. With S and $\hat{S}$ the true and proposed active sets and a, $\hat{a}$ the coefficient vectors after each equation is divided by its own target coefficient,

$$J = \frac{|S \cap \hat{S}|}{|S \cup \hat{S}|}, \qquad \varepsilon_\infty = \max_{k \in S} \frac{|\hat{a}_k - a_k|}{|a_k|}. \tag{3}$$

A run *recovers the structure* when a surviving candidate has $J = 1$, and it *discovers the equation* when such a candidate also has every true coefficient within 10% of its value, $\varepsilon_\infty \leq 0.1$, the same bound on every system. Counts are out of the 10 planned runs of a cell, a run with no answer included. More details are in Appendix D.

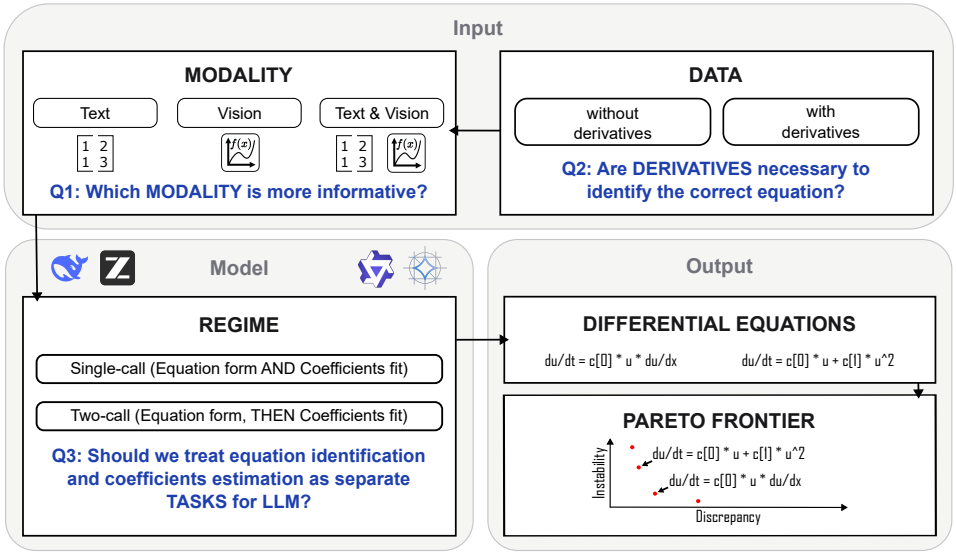


Figure 1: The main sweep. Text is the numerical samples and Vision the rendered plot; with derivative fields off or on and the task asked in one call or two, they give 12 configurations, and the candidates of each run are screened on two objectives.

4 Results

4.1 The main sweep

Across the twelve configurations of the main sweep, GLM discovers between five and ten laws, Qwen between four and nine, DeepSeek between three and eight and Gemma between none and five. Which input serves a model depends on the model: the numerical samples in one call serve GLM and Qwen best, the samples with the derivative fields DeepSeek, and the plot with the samples and the derivative fields Gemma. In its strongest configuration, GLM discovers ten laws, Qwen and DeepSeek eight, and Gemma four (Figure 2).

The system and the model decide discovery far more than the configuration does. Of the variation in discovery over the 4,800 runs, the system accounts for 20%, the model for 13% and their interaction for 8%, while 42% varies between runs of the same cell. Of the configuration factors on their own, what the model is shown accounts for 2%, one or two calls for 0.3%, and the derivative fields for nothing. The plot helps no model on average: in place of the samples it lowers the rate of discovery by 17 to 26 percentage points for GLM, Qwen and DeepSeek, and added to the samples by 9 points for GLM and Qwen. The derivative fields matter only through the system: they help on the Wave equation for every model, cost GLM, Qwen and DeepSeek discoveries on most other systems, and cost Gemma a discovery in a single pair. The fields are sampled at positions of their own, where neither u nor the coordinates are shown, so a model can check a relation pointwise from them only when every term of the law is a supplied derivative: on the Wave equation, and on no other system of the suite. The two-call configurations ask for the terms in one call and their coefficients in a second, to test whether coefficients estimated on their own are more accurate. They are not: of the runs that recover the structure, the share with every coefficient within 10% is six to nine percentage points lower with two calls, for every model.

The coefficients are the models' own, and mostly exact. GLM writes every coefficient exactly in every structure it recovers on eight systems, Qwen and DeepSeek on five and Gemma on none. An exact coefficient does not come from fitting: the derivatives are estimated (§2), and no fit to the computed fields returns one exactly (§4.4). It can be an estimate snapped to a round value, and every true coefficient of the suite is round: the reasoning states an estimate other than the one it writes in half of GLM's discoveries, a quarter of Qwen's, a sixth of DeepSeek's and none of Gemma's, and in 10 of GLM's and 6 of Qwen's it is over 10% off (Table 16). An inexact coefficient is usually far off: one more than 10% off marks 1% of the structures GLM recovers, 10% of Qwen's, 17% of DeepSeek's and 68% of Gemma's (Tables 1 and 15).

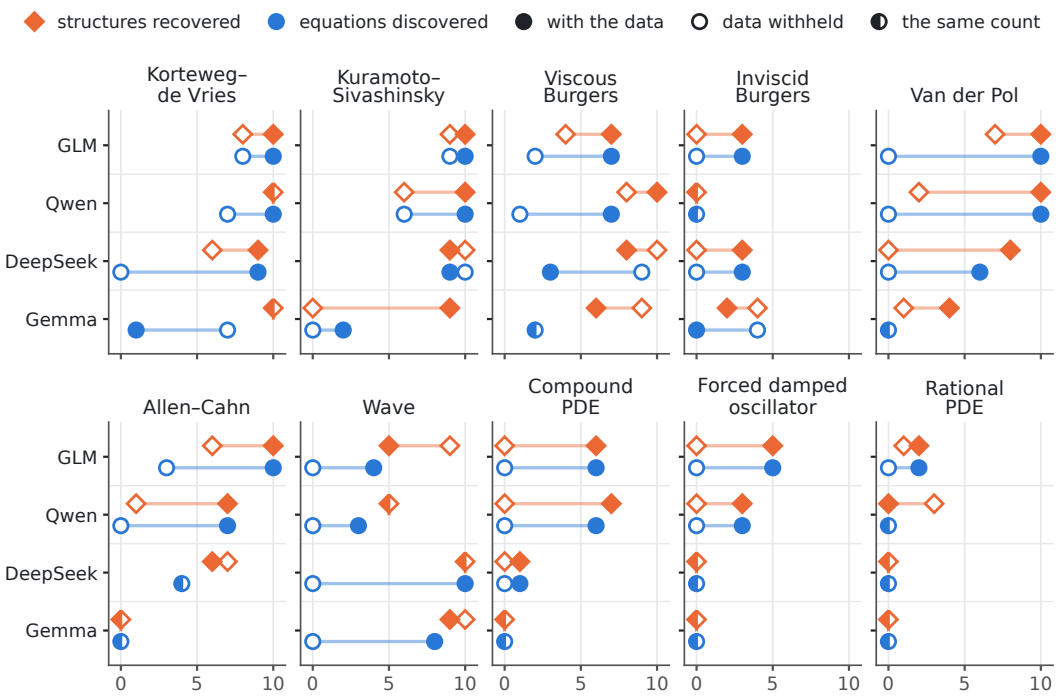


Figure 2: Structures recovered and equations discovered, out of 10 planned runs, per system and model, in each model's strongest configuration with the data and with the data withheld. The counts are in Table 50.

Table 1: ε_∞ of the recovered structures counted in Table 50, as mean (standard deviation), bold for the lowest mean on a system and a dash for none. The last column is the sparse regression of §4.4.

System	GLM	Qwen	DeepSeek	Gemma	Sparse regression
With the data (strongest configuration)					
Korteweg–de Vries	**0**	**0**	**0**	.68 (.28)	*.010*
Kuramoto–Sivashinsky	**0**	**0**	**0**	1.1 (.82)	*.0038*
Viscous Burgers	**0**	.13 (.22)	.39 (.37)	.53 (.44)	*.0079*
Inviscid Burgers	**0**	–	**0**	.75 (.35)	*.000059*
Van der Pol	**0**	**0**	.10 (.17)	1.0 (.74)	*.0037*
Allen–Cahn	**0**	**0**	1,800 (4,000)	–	*.089*
Wave	.11 (.25)	.14 (.24)	**0**	2.9 (8.7)	*.0045*
Compound PDE	.017 (.041)	.018 (.047)	**0**	–	*.00070*
Forced damped osc.	**0**	**0**	–	–	*.011*
Rational PDE	**0**	–	–	–	*.00022*
Data withheld					
Korteweg–de Vries	**0**	.25 (.40)	.89 (.083)	.29 (.47)	
Kuramoto–Sivashinsky	**0**	**0**	**0**	–	
Viscous Burgers	2.5 (2.9)	.79 (.32)	**.090 (.28)**	.70 (.40)	
Inviscid Burgers	–	–	–	**0**	
Van der Pol	**4.4 (1.7)**	5.3 (1.8)	–	6.0	
Allen–Cahn	500 (550)	**99**	1,700 (3,700)	–	
Wave	47 (40)	**24 (0)**	**24 (0)**	25 (.97)	
Compound PDE	–	–	–	–	
Forced damped osc.	–	–	–	–	
Rational PDE	1.5	**1.0 (.0043)**	–	–	

The coefficients are read from the data, at single points or from scales. A point where other terms vanish isolates a coefficient, as at the zero crossings of Van der Pol, where $u_{tt} + u$ reduces to $0.2\,u_t$. Values read at single points appear in 91% of GLM's discoveries, 75% of Qwen's, 64% of DeepSeek's

and 92% of Gemma's, and scales, such as a period, in 99%, 70%, 47% and 8%. Coefficients taken from a standard form appear in a fifth of GLM's Korteweg–de Vries and Kuramoto–Sivashinsky discoveries, two fifths of Qwen's and DeepSeek's and two thirds of Gemma's, and in no model's discovery of Van der Pol or the oscillator.

GLM tests what it reads. Given the same input, the samples in one call, it tests or checks the coefficients it found in nine tenths of its replies, Gemma and Qwen in about two fifths and DeepSeek in a third. GLM's tests are long, and no other model's are; twice as many of them end near the provider's reply limit, where a fifth of GLM's runs end with no answer. More details are in Appendices E.1 to E.4.

4.2 What limits the sweep

The pool is open (§2), and on the three hardest laws this is where runs fail: one of their less common terms is never written. Each failed run is traced to the earliest stage at which a true term stopped: never written or named, named in the reasoning but never written, or written and then screened out. On the Forced damped oscillator, the Rational PDE and the Compound PDE nearly every miss of Qwen, DeepSeek and Gemma is of the first kind, while most of GLM's have no answer. Elsewhere failures vary, among them a structure with a coefficient far off, which accounts for most of Gemma's misses on Korteweg–de Vries and Kuramoto–Sivashinsky.

Failing replies measure as much as successful ones but hold constant a coefficient that varies with t, x or u. None writes the damping modulated by $\sin 2t$ or the drift falling as $2/x$, and on the Compound PDE most keep constant the diffusivity that tracks u. Gemma, given the samples, names a law other than the one in the data in every reply on five systems.

The input decides which failure a run meets. Runs with no answer are frequent for GLM, whose tests are long (§4.1), nearly all in the configurations that carry the samples, where its replies are longest, and more input makes them more frequent: on the Forced damped oscillator a third of its runs from the samples end so, and every run once the derivative fields are added. DeepSeek has few such runs, and Qwen and Gemma almost none. Shown the plot in place of the samples, the runs in which a true term is never written or named become more frequent for GLM, Qwen and DeepSeek and slightly less frequent for Gemma. Under every input, nearly every miss of Qwen, DeepSeek and Gemma on the oscillator and the Rational PDE is a true term never written or named.

The models do not evaluate their own candidates. The prompt asks for five equations and no check of them, and a model that could evaluate its candidates against the data would return the true law first whenever it wrote it. Instead, it is not first in 4% of the replies holding it for GLM, 14% for Qwen, 11% for DeepSeek and 37% for Gemma. Selection has to happen outside the model. The answer cannot be evaluated against the truth (§2), and the candidates are screened on the data alone: outside Inviscid Burgers, this screen loses a discovery in one run, DeepSeek's, while Instability alone loses more for every model and Discrepancy alone for GLM, Qwen and DeepSeek. Instability alone would discard most Allen–Cahn discoveries, whose diffusion coefficient is more than four orders of magnitude below its reaction coefficients, and it cannot see a wrong coefficient on that term; Discrepancy alone prefers a law padded with a spurious term, as on the Wave equation (Figure 3). The Inviscid Burgers field does not identify its law: in the strongest cells GLM, Qwen and DeepSeek write the law in every run and Gemma in none, and in most of these runs the screen keeps in its place an equation that equals it on this field. More details are in Appendices E.5 and E.6.

4.3 Discovery or recall

The data raise every model's discoveries but Gemma's. Without the data a model can write a law only from memory (§3.3), and in the strongest cells GLM discovers the law in 67 runs with the data and 22 without, Qwen in 56 and 14, DeepSeek in 45 and 23, and Gemma in 13 and 13 (Figure 2). GLM discovers six laws it does not recall, Qwen and DeepSeek five and Gemma two; of these, GLM never writes even the structure of three without the data, Qwen of two, DeepSeek of three and Gemma of one, and for every model nearly every discovery of them carries the exact coefficients.

What the models recall is more often a structure than a law. Without the data, GLM and Qwen write the structure of seven laws and DeepSeek and Gemma of five, but recall the law itself for four, three, three and three; on every law a model recalls only in structure, its worst coefficient is off by 89% or

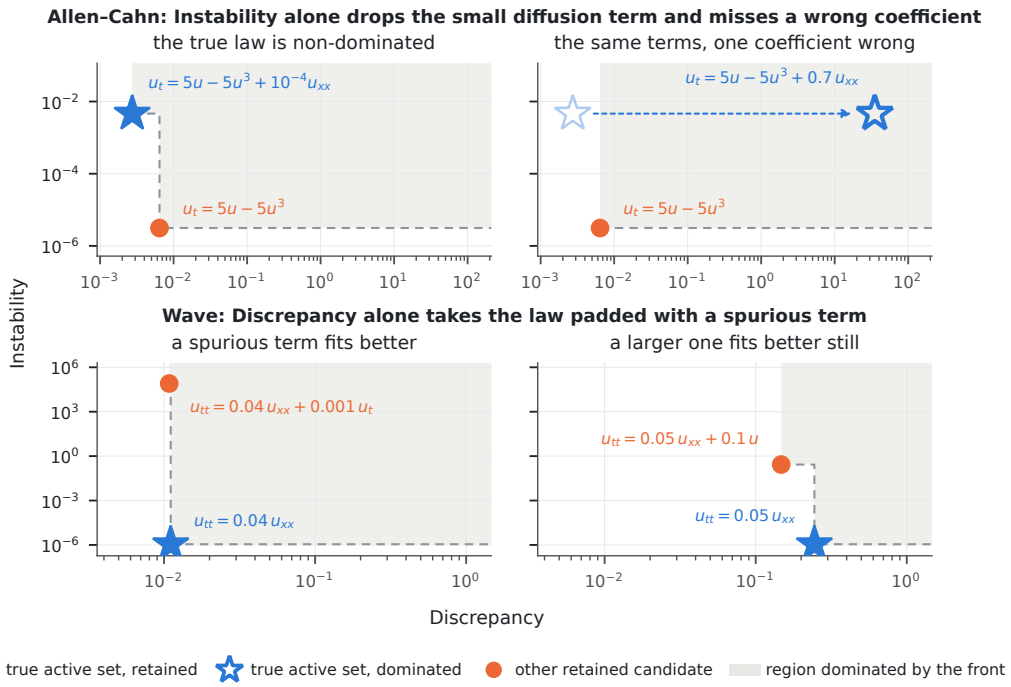


Figure 3: Four runs under the two screening objectives, one row for what goes wrong when a single objective screens alone. Axes are shared within a row, not across rows.

more on average (Table 1). Naming a law is not recalling it: every Van der Pol discovery names the law, yet no model recalls it. The data can also displace a recalled law: DeepSeek and Gemma each discover two laws in fewer runs with the data than without, and GLM and Qwen none. More details are in Appendix E.7.

4.4 Comparison with sparse regression

Sparse regression reaches every law only with help the models do not get. With a general library, every product of at most two factors, $\sin 2t$ and $1/x$ among them, and the left-hand side left open, as it is for the models, a setting chosen per system against the known law discovers eight laws, and none of the 2,200 searched discovers Korteweg–de Vries or Allen–Cahn. The settings that discover differ from system to system, in optimiser, threshold, ridge, column scaling and tolerance, and no one of them discovers all eight laws. The two laws it misses need the predefined library, which holds every true term, with a setting chosen against the known law. Assembled this way, the regression recovers all ten laws with an ε_∞ of at most 0.09, below 0.02 outside Allen–Cahn, and its coefficient error is smaller than every non-exact error of the models except in six of DeepSeek's runs and one of Qwen's, on Allen–Cahn and the Wave equation. Each model uses one configuration on all ten systems. More details are in Appendix C.3.

4.5 Prompt ablations

The grammar and the worked example of the prompt matter little. Cutting or deleting the grammar lowers Qwen's and Gemma's discoveries on at most three systems, and an example taken from another system's law raises them on more systems than it lowers them; for GLM and DeepSeek each arm mainly adds runs with no answer. Given the system's own law as the example, Gemma discovers more on five systems and fewer on none, and the models still estimate the coefficients themselves: no structure recovered under this arm that misses the 10% bound carries the example's coefficients. GLM writes Van der Pol's damping between 0.05 and 0.5, read from the growth of the amplitude, and Qwen writes its own estimates of the Viscous Burgers viscosity. More details are in Appendix E.8.



Figure 4: Equations discovered out of 10 planned runs in each model's strongest cells, from clean data to twice the noise limit, runs with no answer, and replies that say the data are noisy.

4.6 Discovery under noise

The models seldom say that the data are noisy (Figure 4). Most replies say so in four of GLM's cells under noise, one of Gemma's and none of Qwen's or DeepSeek's. DeepSeek and Gemma say it of the supplied derivative fields, which differentiation makes noisier. On the Compound PDE the noise changes under a tenth of the printed samples, on the Rational PDE a third at most, and on Inviscid Burgers mainly their print format.

At twice the noise limit GLM discovers seven laws against ten on clean data, Qwen six against eight, DeepSeek seven against eight and Gemma four against four. On the seven systems whose printed samples noise visibly changes, GLM keeps 86% of its laws at twice the limit, Qwen 57%, DeepSeek 83% and Gemma 75%. For every model nearly every such discovery carries the exact coefficients, and the replies say that the data are noisy in a third of GLM's such discoveries, a few of DeepSeek's and Gemma's and none of Qwen's. At this level the sparse regression recovers no structure on any system, which the definition of the limit anticipates; retuned against the known law, it recovers most laws.

Where a model notices the noise, it can act on it. GLM's replies on Van der Pol estimate the noise at a few thousandths (the true standard deviation at twice the limit is 0.0032), declare pointwise second differences unusable, and read the coefficients from the period, the envelope and the extrema instead. As noise grows, GLM, Qwen and DeepSeek read values at single points less often on most systems.

On the Wave and Allen–Cahn fields noise takes the coefficient before the structure. At twice the limit the Wave structures of Qwen, DeepSeek and Gemma all miss the 10% bound, and DeepSeek's Allen–Cahn structures mostly carry unit reaction coefficients. In all, noise removes most Allen–Cahn discoveries and nearly all Wave, oscillator and Rational PDE discoveries. For GLM it often removes the answer itself: at twice the limit nearly half of its runs end with no answer, against a sixth on clean data, while DeepSeek loses a few runs this way and Qwen and Gemma none. More details are in Appendices C.5 and E.9.

5 Discussion

The coefficients a model writes for a law it does not recall come from the data, and for GLM and Qwen on this suite of round coefficients, leaving the coefficients to an external optimiser is a design choice rather than a necessity. Which laws a model recalls differs between models, and between a structure and its coefficients; a benchmark that removes recall by construction cannot show this, while running each model with the data withheld, ten runs per system, measures it. Further work is

the iterative scenario, in which the model sees the two objectives of its candidates and revises them over several calls, supplying the evaluation it lacks here (§4.2).

Limitations. The study covers four open-weight models queried through one API, ten systems with one state variable and at most one spatial dimension, and ten runs per cell at one temperature, each run without feedback; a model that calls tools may behave differently. Each system is one solution shown through one fixed sample, and how discovery changes with the number of points is not measured. Another encoding of the derivative fields may change what they are worth. Each model's strongest configuration is chosen on the runs it is reported on, and Appendix C.1 sets it beside the runner-up. The 10% bound is a convention, and every true coefficient is round, which rewards snapping. The sparse regression uses a predefined library that holds every true term and settings chosen against the known law, and it serves as a comparison, not a competing method. The noise arm runs on each model's strongest cells with one draw per level. The analysis of the reasoning reads the text written in the reply, which need not be how the answer was reached (Turpin et al., 2023; Lanham et al., 2023), and not hidden traces. Comparisons across cells carry no correction for multiplicity.

6 Conclusion

Read against what each model writes with no data at all, the equations that GLM, Qwen and DeepSeek write without feedback, terms and coefficients, are on most laws of this suite read from the data rather than recalled, with no optimiser fitting the coefficients. The same models keep several of these laws under added noise while rarely saying that the data are noisy, and on the hardest laws what they miss is a less common term they never write.

Statement on the Use of AI

Language models are the object of this study: the four models of §3.2, queried through OpenRouter, produce every reply that is scored, and their replies are the data. Language models were also used as research tools: they read a sample of replies in checking the analysis of the reasoning, and a general-purpose language-model assistant helped write the analysis scripts and checked statements in the text against the scored runs. The same assistant was used to aid and polish the writing. Every number in the paper is computed by scripts from the archived runs, and the authors reviewed all text, code and results and take full responsibility for them.

Ethics Statement

The study involves no human subjects and no personal data. The ten fields are simulated or published solutions of known equations, credited in Appendix A, and the models are used through a public API under its terms. Appendix G reports the compute spent. The risk this work addresses is an overstated discovery: a model that returns a law from memory can look as if it read the law from the data, and every capability claim here is set against what the same model writes without the data.

Reproducibility Statement

The code and data are available at `https://anonymous.4open.science/r/DreamED-C186/` (anonymised for review). The repository holds the ten stored fields, with the generator of the Forced damped oscillator (the sources of the other nine are in Appendix A), the prompt builder with every variant, the runners that issue the calls, and the screening and scoring code. Each run stores the prompt of its final call, the raw reply of each call and the scored candidates. New runs need only an OpenRouter key; the model versions and decoding settings are in Appendix B, and the observation noise is drawn from a generator seeded by the system and the level. The record does not keep which provider served a call, and the DeepSeek model is an experimental endpoint that may be withdrawn.

REFERENCES

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.

Yimeng Chen, Piotr Piękos, Mateusz Ostaszewski, Firas Laakom, and Jürgen Schmidhuber. Physgym: Benchmarking llms in interactive physics discovery with controlled priors. *Advances in Neural Information Processing Systems*, 38, 2026.

Yuntian Chen, Yingtao Luo, Qiang Liu, Hao Xu, and Dongxiao Zhang. Symbolic genetic algorithm for discovering open-form partial differential equations (sga-pde). *Physical Review Research*, 4(2): 023174, 2022.

Mengge Du, Yuntian Chen, Zhongzheng Wang, Longfeng Nie, and Dongxiao Zhang. Large language models for automatic equation discovery of nonlinear dynamics. *Physics of Fluids*, 36(9), 2024a.

Mengge Du, Yuntian Chen, and Dongxiao Zhang. DISCOVER: Deep identification of symbolically concise open-form partial differential equations via enhanced reinforcement learning. *Physical Review Research*, 6(1):013182, 2024b.

Stéphane d'Ascoli, Sören Becker, Alexander Mathis, Philippe Schwaller, and Niki Kilbertus. Odeformer: Symbolic regression of dynamical systems with transformers. *URL https://arxiv. org/abs/2310.05573*, 2023.

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. Mathematical, Physical, and Engineering Sciences*, 478(2260):20210904, 2022.

Yuchen He, Hongkai Zhao, and Yimin Zhong. How much can one learn a partial differential equation from its solution? *Foundations of Computational Mathematics*, 24(5):1595–1641, 2024.

Rob J Hyndman and Anne B Koehler. Another look at measures of forecast accuracy. *International journal of forecasting*, 22(4):679–688, 2006.

Alan A Kaptanoglu, Brian M de Silva, Urban Fasel, Kadierdan Kaheman, Andy J Goldschmidt, Jared L Callaham, Charles B Delahunt, Zachary G Nicolaou, Kathleen Champion, Jean-Christophe Loiseau, et al. Pysindy: A comprehensive python package for robust sparse system identification. *arXiv preprint arXiv:2111.08481*, 2021a.

Alan A Kaptanoglu, Brian M de Silva, Urban Fasel, Kadierdan Kaheman, Andy J Goldschmidt, Jared L Callaham, Charles B Delahunt, Zachary G Nicolaou, Kathleen Champion, Jean-Christophe Loiseau, et al. Pysindy: A comprehensive python package for robust sparse system identification. *arXiv preprint arXiv:2111.08481*, 2021b.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, et al. Measuring faithfulness in chain-of-thought reasoning, july 2023. *URL http://arxiv. org/abs/2307.13702*, 2023.

Xiao Luo, Changhu Wang, Yizhou Sun, and Wei Wang. How do large language models perform on pde discovery: A coarse-to-fine perspective. In *EMNLP (Findings)*, pp. 2684–2697, 2025.

Pingchuan Ma, Tsun-Hsuan Wang, Minghao Guo, Zhiqing Sun, Joshua B Tenenbaum, Daniela Rus, Chuang Gan, and Wojciech Matusik. Llm and simulation as bilevel optimizers: A new paradigm to advance physical scientific discovery. *arXiv preprint arXiv:2405.09783*, 2024.

Mikhail Maslyaev and Alexander Hvatov. Comparison of single-and multi-objective optimization quality for evolutionary equation discovery. In *Proceedings of the Companion Conference on Genetic and Evolutionary Computation*, pp. 603–606, 2023.

Mikhail Maslyaev and Alexander Hvatov. Equation discovery framework epde: Towards a better equation discovery. *arXiv preprint arXiv:2501.14768*, 2024.

Mikhail Maslyaev, Alexander Hvatov, and Anna V Kalyuzhnaya. Partial differential equations discovery with epde framework: Application for real and synthetic data. *Journal of Computational Science*, 53:101345, 2021.

Daniel A Messenger and David M Bortz. Weak sindy for partial differential equations. *Journal of computational physics*, 443:110525, 2021.

Maziar Raissi, Paris Perdikaris, and George E Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational physics*, 378:686–707, 2019.

Samuel H Rudy, Steven L Brunton, Joshua L Proctor, and J Nathan Kutz. Data-driven discovery of partial differential equations. *Science advances*, 3(4):e1602614, 2017.

Philipp Scholl, Aras Bacho, Holger Boche, and Gitta Kutyniok. Symbolic recovery of differential equations: The identifiability problem. *URL https://arxiv. org/abs/2210.08342*, 2022.

Parshin Shojaee, Kazem Meidani, Shashank Gupta, Amir Barati Farimani, and Chandan Reddy. Llm-sr: Scientific equation discovery via programming with large language models. In *International Conference on Learning Representations*, volume 2025, pp. 16054–16085, 2025a.

Parshin Shojaee, Ngoc-Hieu Nguyen, Kazem Meidani, Amir Barati Farimani, Khoa D Doan, and Chandan K Reddy. Llm-srbench: A new benchmark for scientific equation discovery with large language models. *arXiv preprint arXiv:2504.10415*, 2025b.

Robert Stephany and Christopher Earls. Weak-pde-learn: A weak form based approach to discovering pdes from noisy, limited data. *Journal of Computational Physics*, 506:112950, 2024.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965, 2023.

Hao Xu, Junsheng Zeng, and Dongxiao Zhang. Discovery of partial differential equations from highly noisy and sparse data with physics-informed information criterion. *Research*, 6:0147, 2023.

Tianshi Zheng, Kiu Wai Tam, Kim Hue Nam Nguyen, Baixuan Xu, Zhaowei Wang, Cheng Jiayang, Hong Ting Tsang, Weiqi Wang, Jiaxin Bai, Tianqing Fang, et al. Newtonbench: Benchmarking generalizable scientific law discovery in llm agents. In *International Conference on Learning Representations*, volume 2026, pp. 99018–99088, 2026.

A DATA

The ten systems are solutions of the laws in Table 2 on fixed grids, two ordinary and eight partial, all numerical except the closed-form Inviscid Burgers.

Table 2: The ten systems: grid, domain, points shown in the prompt, grid points cropped from each end of the t and x axes, and prompt points inside the crop. The Rational PDE is scored after multiplying through by x.

System	Law	Grid	t	x	n shown	Crop	n in crop	Source
Forced damped osc.	$u_{tt} = -u_t \sin 2t - 4u + 1.5t$	320	$[0, 15.95]$	–	320	30	260	this work
Van der Pol	$u_{tt} = 0.2u_t - 0.2u^2 u_t - u$	320	$[0, 15.95]$	–	320	30	260	EPDE
Wave	$u_{tt} = 0.04\, u_{xx}$	101×101	$[0, 1]$	$[0, 1]$	1,936	10, 10	1,225	EPDE
Inviscid Burgers	$u_t = -u\, u_x$	101×101	$[0, 1]$	$[-1000, 0]$	1,936	10, 10	1,225	EPDE
Viscous Burgers	$u_t = -u\, u_x + 0.1\, u_{xx}$	101×256	$[0, 10]$	$[-8, 8]$	1,988	10, 25	1,254	PDE-FIND
Korteweg–de Vries	$u_t = -6u\, u_x - u_{xxx}$	201×512	$[0, 20]$	$[-30, 30]$	1,988	20, 50	1,254	PDE-FIND
Allen–Cahn	$u_t = 5u - 5u^3 + 10^{-4} u_{xx}$	51×128	$[0, 1]$	$[-1, 1]$	1,960	5, 12	1,254	PINN
Kuramoto–Sivashinsky	$u_t = -u\, u_x - u_{xx} - u_{xxxx}$	251×1024	$[0, 100]$	$[0, 32\pi]$	1,980	5, 5	1,760	PDE-FIND
Compound PDE	$u_t = u_x^2 + u\, u_{xx}$	251×100	$[0, 0.5]$	$[1, 2]$	1,960	10, 25	896	SGA-PDE
Rational PDE	$u_t = -2\, u_x/x + 0.5\, u_{xx}$	251×100	$[0, 0.5]$	$[1, 2]$	1,960	10, 25	896	SGA-PDE

Sources. Nine fields are taken from earlier studies (Rudy et al., 2017; Raissi et al., 2019; Chen et al., 2022; Maslyaev & Hvatov, 2023; 2024), and the Forced damped oscillator is integrated here. Viscous Burgers, Korteweg–de Vries and Kuramoto–Sivashinsky are the PDE-FIND files of Rudy et al. (2017), and the Compound and Rational PDEs are the open-form datasets of Chen et al. (2022), all identical to the published files. The Wave field is the solution of Maslyaev & Hvatov (2023) on its 101×101 grid, and Inviscid Burgers is the closed-form solution of the same study, $u = x/(t-2)$, on a smaller window. Van der Pol is the fourth-order Runge–Kutta series of Maslyaev & Hvatov (2024), and the Forced damped oscillator is integrated by the same scheme, with step 0.05 and initial state $u(0) = 0.8$, $u_t(0) = 2$. Allen–Cahn keeps every fourth point in each axis of the solution of Raissi et al. (2019). The Rational PDE places the array of Chen et al. (2022) on $t \in [0, 0.5]$ where their code uses $[0, 1]$, which doubles both coefficients of their $u_t = -u_x/x + 0.25\, u_{xx}$; no candidate written with the data withheld carries the published coefficients. On the periodic systems the stored grid omits the right end of the domain.

What a prompt carries. The samples are an evenly spaced subgrid of the stored array, capped at roughly $2{,}000$ points and allocated between the axes in proportion to their length; on the two ordinary systems they are the whole series of 320 points. Values are printed to four decimals, and the selection is integer `linspace` on each axis, the same in every configuration. Every derivative field is computed from the stored array by central differences, iterated for higher orders, on the full grid, and then cropped and sampled at $2{,}000$ evenly spaced positions of the cropped grid, shown without coordinates; these are mostly not the points of the field sample.

Boundary crop. A second-order difference rule extrapolates at the edges of the grid, so a margin of grid points is cropped from each end of each axis after differentiation (Table 2): 30 points of each ordinary series, and between 5 and 50 points per end on the partial systems. The candidates are screened and scored on every point of the cropped grid, and the sparse regression is fitted on the points of the field sample that fall inside it. A point the prompt showed outside the crop has no derivative and enters no fit; the crop leaves a little under two thirds of the shown points on most partial systems and under half on the Compound and Rational PDEs, where it removes half of the spatial axis.

Plot. The plot shows the field over the same grid, or the field and each supplied derivative as a composite panel, with axes labelled by coordinate and no title, legend or annotation naming the system.

B MODELS, PROMPTS AND PROTOCOL

The four models are called through OpenRouter (Table 3). Each identifier resolves to one dated version released before the first run of the main sweep, which ran from 3 to 18 September 2026; the

data-withheld arm ran on 17 and 21 September, the prompt arms from 19 to 21 September and the noise arm from 22 to 24 September. All four accept image input natively and write their reasoning into the reply; GLM, Qwen and DeepSeek also spend tokens on a hidden trace before it, and Gemma does not. Over the 1,200 runs of its main sweep, runs with no answer included, a run takes about 22 minutes on average for GLM, 8 for Qwen, 6 for DeepSeek and just over one for Gemma; the wall time and cost of every cell are in Appendix G.

Table 3: The four models as OpenRouter lists them, with their published weights. Limits are the top provider's, in tokens, prices in US dollars per million input / output tokens; Trace marks a hidden trace billed apart from the reply.

Model	Version / weights	Context	Max reply	$ in / out	Trace
GLM	`z-ai/glm-5.3-flash-20260826` `zai-org/GLM-5.3-Flash`	1,048,576	943,718	0.15 / 0.50	yes
Qwen	`qwen/qwen3.8-flash-20260826` `Qwen/Qwen3.8-Flash-Next`	1,000,000	131,072	0.15 / 0.47	yes
DeepSeek	`deepseek/deepseek-v4-flash-vision-exp-20260821` `deepseek-ai/DeepSeek-V4-Flash-Vision-Exp`	1,048,576	943,718	0.22 / 0.66	yes
Gemma	`google/gemma-4-31b-it-20260402` `google/gemma-4-31B-it`	262,144	16,384	0.09 / 0.34	no

Decoding. Temperature is 0.5 in every call. The system prompt is frozen, one text for the eight partial systems and one for the two ordinary ones in each task split, with the system itself described in the user prompt; the two arms whose example is another system's law or the system's own carry one text per system. A cell's 10 runs are 10 draws from one fixed request, except in three GLM two-call cells, where one run whose structure call returned nothing parseable was sent the one-call request. The completion length is left uncapped: a cap low enough to bound the spend is low enough to cut a model off before it answers. The only limit is the maximum reply of the provider that serves the call, and every run with no answer was cut there, at 131,072 completion tokens or fewer. The serving provider is not recorded, so GLM's runs with no answer, and with them its counts, depend on routing. Where a model exposes a reasoning-effort control, the provider default is kept, and the schema-constrained output mode is off.

B.1 The prompt

Every run of a cell carries one frozen system prompt, given below with the grammar block and the request for five candidates elided; two prompt arms change the grammar block.

```
You are an expert in differential equations and dynamical systems.
Given information about a dataset, identify the governing differential equation
with exact numerical coefficients.

Your task is to find the COMPLETE equation: both the correct terms AND their
numerical coefficient values. For example, if the equation is
u_tt = 0.75*u_x*u_xx - 0.02*u*u_xxx, you must identify the terms
(u_x*u_xx and u*u_xxx) AND estimate the coefficients (0.75 and -0.02).

{grammar_block}

Reason through the problem in structured steps:
1. Dynamical Behavior - describe what the data is doing (wave propagation,
   diffusion, shocks, oscillations, decay, blowup, reaction-diffusion, etc.).
2. Spatial and Temporal Patterns - call out separate mechanisms you see
   (e.g. Growth / Advection / Diffusion / Dispersion) with the magnitudes
   or rates you read off the data.
3. Candidate Equation - state the most plausible governing equation in
   standard notation, briefly justifying each term from the patterns above.
4. Coefficient Estimation - derive specific numerical coefficients per term
   from the observed magnitudes, rates, and balances.

Output format - JSON with target and terms (each term mapped to its
coefficient):
  {"target": "u_tt", "terms": {"u_x*u_xx": 0.75, "u*u_xxx": -0.02}}

This represents the equation: u_tt = 0.75*u_x*u_xx - 0.02*u*u_xxx

Important rules:
- Show your structured reasoning BEFORE the JSON answer (aim for ~300-500
  words).
```

```
- Do NOT echo data arrays, derivative values, or grid coordinates back in
  your reply.
- The JSON must be valid and parseable.
- Coefficients must be specific numerical values (not placeholders like "a"
  or "c1").
  Estimate them from the data magnitudes and derivative relationships.
- Use only the operands and spellings listed in the equation grammar above.

{five_candidate_request}
```

The worked example in the output-format line is the same for the eight partial systems and is written in time derivatives for the two ordinary ones; neither names a system of the suite. Where the task is split across two calls, the first asks for structures only and the second is given the structures and asked for coefficients; neither call sees the other's reasoning. The user prompt gives the grid description, then the data of the configuration, then analysis questions and a request for coefficients estimated from the data. The data-withheld arm sends the grid description alone, here for the Wave equation:

```
Propose the governing differential equation for a dataset with the
following description. No sampled values are provided.

Data shape: (101, 101)
Time range: [0.0, 1.0], dt = 0.010000, n_time = 101
Space range: [0.0, 1.0], dx = 0.010000, n_space = 101
Equation type: PDE
Variable(s): ['u']

Address:
1. Given only this description, which differential equation(s) would you
   consider most likely to have produced such a dataset?
2. Explain your reasoning from the equation type, dimensionality and grid.

State exact numerical coefficients for each term, as your best estimate.
```

A data prompt, here for Korteweg–de Vries with samples and derivative fields, each array cut after its first values and the analysis questions elided:

```
Discover the governing differential equation from the following dataset.

Data shape: (201, 512)
Time range: [0.0, 20.0], dt = 0.100000, n_time = 201
Space range: [-30.0, 29.8828125], dx = 0.117188, n_space = 512
Equation type: PDE
Variable(s): ['u']

Sampled data values:
t (sampled 28 of 201 points):
  [ 0. ,  0.7,  1.4,  2.2,  2.9,  3.7,  4.4,  5.1,  5.9,  6.6,  7.4,  8.1, ...
x (sampled 71 of 512 points):
  [-30.    , -29.1797, -28.3594, -27.5391, -26.6016, -25.7812, -24.9609, - ...
u(t, x) - rows=time, cols=space (28x71 sample):
[[ 9.0792e-05,  2.0618e-04,  4.6816e-04,  1.0627e-03,  2.7091e-03,  6.1318 ...

Pre-computed derivative fields (for analysis - these are raw data, NOT necessarily
terms in the equation; the equation may use different combinations of operands):
u_t (sampled 2000 of 66332 points):
  [-4.3008e-03, -1.3766e-01,  1.8566e-01,  8.3703e-03,  1.2601e-04, -1.100 ...
u_tt (sampled 2000 of 66332 points):
  [ 4.2846e-03,  8.2330e-02,  4.3230e-02,  8.8988e-03,  1.6274e-04, -5.842 ...
u_x (sampled 2000 of 66332 points):
  [ 4.2862e-03,  1.3765e-01, -1.8558e-01, -8.3860e-03, -1.6333e-04,  3.475 ...
u_xx (sampled 2000 of 66332 points):
  [ 4.1848e-03,  8.2679e-02,  4.3304e-02,  8.4212e-03,  1.4004e-04,  1.625 ...
u_xxx (sampled 2000 of 66332 points):
  [ 3.5254e-03, -3.8329e-03,  1.2632e-01, -6.4904e-03, -4.4932e-04,  7.017 ...

{analysis questions}

Estimate exact numerical coefficients for each term from the data magnitudes
and derivative relationships.
```

C EXPERIMENTS

C.1 THE MAIN SWEEP AND THE STRONGEST CELLS

A model's strongest configuration (§4.1) holds the most discoveries over its ten cells, with no tie at the top; Qwen's leads its runner-up by one run, 56 to 55. It is selected on the same runs it is reported on. Table 4 sets it beside the runner-up and splits it into its first and last five runs by time. The runner-up discovers nine laws for GLM and Qwen, eight for DeepSeek and four for Gemma, where the strongest configuration discovers ten, eight, eight and four; for GLM and Qwen it gains most on the Wave equation, where it supplies the derivative fields. Chosen on the first five runs of every cell, the strongest configuration is the same for GLM, Qwen and Gemma, and chosen on the last five, for Qwen and DeepSeek.

Table 4: Equations discovered in each model's strongest and runner-up configurations, out of 10 planned runs, and in the strongest one split by time into runs 1–5 and 6–10, out of 5.

Model	Configuration	KdV	K–S	$\mathrm{Bu_v}$	$\mathrm{Bu_i}$	VdP	A–C	Wave	Comp	Osc	Rat
GLM	samples (strongest)	10	10	7	3	10	10	4	6	5	2
	samples+derivs (runner-up)	10	10	8	3	9	9	8	5	0	1
	strongest, runs 1–5	5	5	5	2	5	5	3	4	4	0
	strongest, runs 6–10	5	5	2	1	5	5	1	2	1	2
Qwen	samples (strongest)	10	10	7	0	10	7	3	6	3	0
	samples+derivs (runner-up)	10	10	5	2	10	3	10	2	3	0
	strongest, runs 1–5	5	5	3	0	5	3	1	3	2	0
	strongest, runs 6–10	5	5	4	0	5	4	2	3	1	0
DeepSeek	samples+derivs (strongest)	9	9	3	3	6	4	10	1	0	0
	plot+samples (runner-up)	8	9	5	2	10	2	3	4	0	0
	strongest, runs 1–5	4	4	0	1	4	1	5	0	0	0
	strongest, runs 6–10	5	5	3	2	2	3	5	1	0	0
Gemma	plot+samples+derivs (strongest)	1	2	2	0	0	0	8	0	0	0
	samples+derivs (runner-up)	0	0	1	1	0	0	8	1	0	0
	strongest, runs 1–5	1	1	0	0	0	0	4	0	0	0
	strongest, runs 6–10	0	1	2	0	0	0	4	0	0	0

C.2 THE DATA-WITHHELD ARM

The first arm after the sweep checks recall. It runs every model in the one-call configuration with the data removed, over all ten systems, 10 runs per cell (Table 9). Its prompt keeps the grid description and the type of equation, whose domains and resolution may cue a familiar system, but no law can be derived from it.

C.3 SPARSE REGRESSION ON THE SAME POINTS

The comparison is sparse regression through PySINDy (Kaptanoglu et al., 2021a;b), fitted on the points the prompt carried and scored by the same scorer as the models. It is run first with a general library and then with a predefined one (Table 6), over the same 2,200 settings (Table 7), and Table 8 gives the setting chosen for each system and how many of the 2,200 discover the law.

The general library. It is the regression's counterpart of the models: it is not told the law, although its coordinate factors, $\sin 2t$ and $1/x$ among them, were chosen with the suite in view. Its columns are every product of at most two of the factors the predefined library is built from, and the left-hand side is not given: each setting is fitted once with u_t and once with u_{tt} on the left, and the two fits are screened by Discrepancy and Instability as the models' candidates are. Chosen against the known law, a setting discovers eight laws; on Korteweg–de Vries and Allen–Cahn no setting does, and the closest fits miss the dispersion and the diffusion term (Table 5); Rudy et al. (2017) recover Korteweg–de Vries from the full grid, not from about 1,250 rows. The discovering settings differ from system to system (Table 8): `STLSQ` on five systems, with thresholds from $3{\cdot}10^{-3}$ to 2, and `FROLS` on three, with one or eight iterations; a ridge of 10^{-10} except on Viscous Burgers; column scaling off on two systems; and coefficient tolerances from 0 to 0.1. Only Inviscid Burgers and the Wave equation share a setting. Wherever the fit with the correct left-hand side has the true terms, the screen keeps it.

The predefined library. It is used only on Korteweg–de Vries and Allen–Cahn, where no setting of the general library discovers the law. It supplies what the regression needs there: every true term, among the hand-picked products of the suite such as u_x/x and $u_t \sin 2t$, the system's own left-hand side, and a setting chosen against the known law, and with these it discovers both laws. Across the main sweep, every structure a model recovers has an ε_∞ of exactly zero or at least the regression's, except six DeepSeek runs and one Qwen run on Allen–Cahn and the Wave equation.

Table 5: Where no setting of the general library discovers the law: the closest fit of the 2,200 settings, with the correct left-hand side.

System	Law	Closest fit
Korteweg–de Vries	$u_t = -6u\,u_x - u_{xxx}$	$u_t = -0.760\,u_x - 0.624\,u\,u_x$
Allen–Cahn	$u_t = 5u - 5u^3 + 10^{-4}u_{xx}$	$u_t = 4.96\,u - 4.96\,u^3$

Table 6: The two libraries of the sparse regression, for the partial and the ordinary systems.

	General library	Predefined library
Field	u, u^2, u^3	u, u^2, u^3
Derivatives	u_x to u_{xxxx} (partial systems), u_t and u_{tt}, without the target	u_x to u_{xxxx} and the time derivative that is not the target; on the ordinary systems u_t
Products	every product of at most two of the factors above and below, squares included	each power of u times one derivative
Coordinate terms	factors t, $1/t$, $\sin t$, $\cos t$, $\sin 2t$, $\cos 2t$, and on the partial systems x, $1/x$, $\sin x$, $\cos x$, $\sin 2x$, $\cos 2x$	t, x, $\sin x$, $\cos t$, $\sin x \cos t$, $\cos x \sin t$; on the ordinary systems t, $\sin t$, $\cos t$
Hand-picked terms	none	u/x, u_x/x, u_x^2; on the ordinary systems $u_t \sin 2t$
Constant	yes	yes
Columns	226 partial, 62 ordinary	33 partial, 12 ordinary
Left-hand side	u_t or u_{tt}: both are fitted and the two fits screened	the system's own
Used on	every system	Korteweg–de Vries and Allen–Cahn, where no general setting discovers the law

Table 7: The settings searched with both libraries, and how a setting is chosen for each system.

Implementation	PySINDy: SINDy on the ordinary systems, PDE-FIND on the partial ones
Rows fitted	the prompt's points inside the crop, with the finite-difference fields at them
`STLSQ` threshold λ	0, 10^{-8}, 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, $3{\cdot}10^{-3}$, 0.01, 0.03, 0.1, 0.3, 0.5, 1, 2, 3
`FROLS` iterations k	1, 2, 3, 4, 5, 6, 8, 10, 12, 15
Ridge α	10^{-10}, 10^{-5}, 10^{-3}, 0.05, for both optimisers
Column scaling	on, off
Coefficient tolerance	0, 10^{-9}, 10^{-7}, 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 0.01, 0.05, 0.1, 0.2, applied after the fit
Settings	$(15 + 10) \times 4 \times 2 \times 11 = 2{,}200$
General library	the discovering setting with the lowest ε_∞
Predefined library	on Korteweg–de Vries and Allen–Cahn, the setting that best recovers the known law

Table 8: Sparse regression per system: the general library's best setting, and, where no general setting discovers the law, the predefined library's setting chosen against the known law, each with its ε_∞ and the number of the 2,200 settings that discover the law (λ threshold, k iterations, α ridge, column scaling, coefficient tolerance).

	General library			Predefined library		
System	Best setting	ε_∞	Settings	Setting	ε_∞	Settings
Korteweg–de Vries	none	–	0	FROLS, k 3, α 10^{-10}, scaled, tol 0.05	.01	161
Kuramoto–Sivashinsky	STLSQ, λ 0.03, α 10^{-10}, scaled, tol 0.05	.0038	528	not used	–	–
Viscous Burgers	STLSQ, λ $3{\cdot}10^{-3}$, α 10^{-5}, raw, tol 0.1	.0079	65	not used	–	–
Inviscid Burgers	FROLS, k 1, α 10^{-10}, scaled, tol 0	$5.9{\cdot}10^{-5}$	88	not used	–	–
Van der Pol	FROLS, k 8, α 10^{-10}, scaled, tol 0.01	.0037	179	not used	–	–
Allen–Cahn	none	–	0	FROLS, k 4, α 10^{-10}, scaled, tol 10^{-7}	.089	12
Wave	FROLS, k 1, α 10^{-10}, scaled, tol 0	.0045	396	not used	–	–
Compound PDE	STLSQ, λ 0.5, α 10^{-10}, scaled, tol 0.05	$7.0{\cdot}10^{-4}$	92	not used	–	–
Forced damped osc.	STLSQ, λ 0.1, α 10^{-10}, raw, tol 0	.011	109	not used	–	–
Rational PDE	STLSQ, λ 2, α 10^{-10}, scaled, tol 0	$2.2{\cdot}10^{-4}$	22	not used	–	–

C.4 THE PROMPT ARMS

Every model is run on its strongest cell over all ten systems for each arm, 10 runs per cell (Table 9). The grammar arms separate needing the vocabulary from needing the syntax; the example arms bracket what the worked example is worth, from placeholders to a misleading law to the answer itself.

Table 9: The six arms, the data-withheld arm first. Control is the same cell of the main sweep.

Arm	What it changes
data withheld	the samples and the plot are removed, and only the grid description and the type of equation remain
grammar to operands	the grammar block keeps its operand list and loses its factor rules, spelling rules and left/right-hand-side contract
grammar deleted	the grammar block is removed, and with it the closing rule that refers to it
example abstract	the worked example becomes placeholders with the same term count and coefficient magnitudes and no physical content; the line restating it as an equation reads `T0 = -0.35*A1`
example poisoned	the worked example becomes another system's true law, from a donor map fixed in code
example is the answer	the worked example becomes this system's own true law

C.5 NOISE

Noise is added to the stored field as $u + \sigma\varepsilon$, with ε standard normal and σ a fraction of the field's standard deviation, on the full grid before any derivative is taken, so that the derivative fields inherit it as a measured field would. One draw is made per system and level, from a generator seeded by the two, and the samples, the derivative fields and the arrays the screening objectives use are all taken from it. The levels follow one rule per system: the sparse regression is refitted on 30 independent draws at each level of an increasing ladder, and the noise limit is the highest level at which 3 to 5 of the 30 draws still return exactly the true active set (Table 10, first column). Each strongest cell is run at half, once and twice its system's limit.

What the models see. The noise is small against the field on every system, and what it changes in the prompt depends on the print format (Table 10). At twice the limit it changes most printed samples on seven systems, but at most a third on the Compound PDE and the Rational PDE, where its standard deviation is below the last printed digit of a typical sample. On Inviscid Burgers it changes mainly the print format of the samples, and at equal print format it changes 4% of them. Differentiation amplifies it in the derivative fields shown to DeepSeek and Gemma, which change on every system.

Table 10: Noise per system: the sparse regression's noise limit in percent of the field's standard deviation, the noise standard deviation at twice the limit, and the share of printed values that differ from the clean prompt, for the samples and, at twice the limit, the derivative fields; on Inviscid Burgers, in parentheses, the share at equal print format.

			Samples changed (%)			Derivatives, 2×	
System	Limit (% of std)	σ at 2×	0.5×	1×	2×	Changed (%)	Median rel. change
Korteweg–de Vries	$6.0 \cdot 10^{-3}$	$1.3 \cdot 10^{-5}$	83	89	94	>99	0.085
Kuramoto–Sivashinsky	$7.5 \cdot 10^{-3}$	$1.6 \cdot 10^{-4}$	70	80	90	>99	0.040
Viscous Burgers	$1.3 \cdot 10^{-3}$	$4.6 \cdot 10^{-6}$	68	73	79	99	0.029
Inviscid Burgers	$1.2 \cdot 10^{-5}$	$5.2 \cdot 10^{-5}$	90 (3)	90 (3)	90 (4)	81	$2.4 \cdot 10^{-4}$
Van der Pol	0.15	$3.2 \cdot 10^{-3}$	97	99	99	100	0.14
Allen–Cahn	0.47	$6.4 \cdot 10^{-3}$	>99	>99	>99	>99	0.77
Wave	0.040	$1.5 \cdot 10^{-3}$	92	95	98	>99	0.21
Compound PDE	$2.2 \cdot 10^{-5}$	$7.9 \cdot 10^{-8}$	7	7	8	63	$1.5 \cdot 10^{-4}$
Forced damped osc.	0.87	0.036	>99	100	>99	100	0.55
Rational PDE	$1.6 \cdot 10^{-4}$	$6.8 \cdot 10^{-7}$	17	20	32	94	$1.3 \cdot 10^{-3}$

Strength of the reference. The limit measures one regression, not the strongest one. Table 11 refits the calibration draws with the regression retuned at each level against the known law, with central or Savitzky–Golay derivatives. Retuning alone recovers the structure at twice the limit on most systems, Savitzky–Golay derivatives on the full grid on nearly all, and Allen–Cahn and the Wave equation resist both; retuned on the draw the models were given, the regression also meets the 10% bound at twice the limit on seven systems. Bagging over subsamples and the weak form recover laws at still higher noise (Fasel et al., 2022; Messenger & Bortz, 2021).

Table 11: Sparse regression near its noise limit: draws out of 30 that return exactly the true active set for the reference (a), and retuned at each level with central differences on the prompt's points (a′), Savitzky–Golay derivatives on the prompt's points (b) and on the full cropped grid (c), with the best single setting for those 30 draws in parentheses. The first rows give the condition number of the library on clean data.

Noise	Fit	KdV	K–S	Bu_v	Bu_i	VdP	A–C	Wave	Comp	Osc	Rat
cond.	prompt points	$1.9{\cdot}10^4$	63	$1.9{\cdot}10^3$	$> 10^{15}$	63	279	$1.9{\cdot}10^4$	$5.2{\cdot}10^5$	135	$3.8{\cdot}10^5$
	full grid	$1.8{\cdot}10^4$	63	$1.7{\cdot}10^3$	$> 10^{15}$	63	280	$1.9{\cdot}10^4$	$5.3{\cdot}10^5$	135	$4.0{\cdot}10^5$
0.5×	(a)	30	0	26	16	8	14	30	23	9	12
	(a′)	29 (30)	26 (30)	26 (30)	30 (30)	22 (30)	14 (22)	30 (30)	20 (30)	20 (30)	12 (30)
	(b)	30 (30)	26 (30)	30 (30)	30 (30)	29 (30)	0 (0)	30 (30)	30 (30)	5 (30)	7 (30)
	(c)	30 (30)	30 (30)	30 (30)	30 (30)	29 (30)	0 (0)	30 (30)	30 (30)	5 (30)	30 (30)
1×	(a)	3	3	3	5	3	5	3	4	4	4
	(a′)	30 (30)	4 (23)	16 (30)	30 (30)	29 (30)	0 (13)	27 (27)	13 (30)	9 (30)	10 (30)
	(b)	30 (30)	28 (30)	30 (30)	20 (30)	23 (30)	0 (0)	30 (30)	19 (30)	7 (30)	18 (30)
	(c)	30 (30)	30 (30)	30 (30)	5 (30)	23 (30)	0 (3)	30 (30)	27 (30)	7 (30)	28 (30)
2×	(a)	0	0	0	0	0	1	0	0	0	0
	(a′)	23 (30)	0 (0)	0 (30)	30 (30)	29 (30)	0 (3)	1 (11)	29 (30)	30 (30)	30 (30)
	(b)	30 (30)	17 (30)	19 (30)	28 (30)	28 (30)	0 (2)	23 (30)	29 (30)	30 (30)	12 (30)
	(c)	30 (30)	30 (30)	30 (30)	30 (30)	28 (30)	0 (5)	1 (30)	26 (30)	30 (30)	22 (30)

D EVALUATION

Screening. A proposal is a target term equal to a weighted sum of other terms, and evaluating both sides on the stored fields gives y and $\hat{y}$. Discrepancy is 1 for the do-nothing proposal and comparable across systems whose magnitudes differ by orders of magnitude. Instability refits the proposal's own terms by least squares, with a ridge of 10^{-6}, on sliding windows of half the axis length along each grid axis, and sums the squared coefficient of variation of each term across the windows, divided by the number of grid dimensions; a term whose mean is at or below 10^{-12}, or that is refitted in a single window, is charged a large penalty. Summing over terms, dividing by d and minimising Instability as an objective are departures from Xu et al. (2023). The screen of §2 compares the objectives exactly, and a candidate that cannot be evaluated is never kept. Where the archived objectives of six Kuramoto–Sivashinsky candidates were not finite, both were recomputed by the same code.

Sensitivity of Instability. On each model's strongest cells, Instability was recomputed with an average over terms in place of the sum, with half and twice the windows per axis, and with windows of a third and two thirds of the axis (Table 12). Outside Inviscid Burgers no variant gains or loses a structural recovery or a discovery. On Inviscid Burgers several variants change which of the forms equal to the law on this field survives (Appendix D.1), and with it the discoveries, in both directions.

Table 12: Runs of the strongest cell, out of its 10 planned runs, whose screened set changes when Instability averages over terms, halves or doubles the windows per axis, or uses windows of a third or two thirds of the axis; where any change, the structural recoveries and the discoveries gained (+) and lost (−) follow in parentheses; · marks no change.

Model		KdV	K–S	$\mathrm{Bu_v}$	$\mathrm{Bu_i}$	VdP	A–C	Wave	Comp	Osc	Rat
GLM	average over terms	·	·	·	4	·	·	·	·	1	·
	$\frac{1}{2}\times$ windows	·	·	·	5 (+5 −0, +5 −0)	·	·	·	·	·	·
	$2\times$ windows	·	·	·	·	·	·	·	·	·	·
	window $\frac{1}{3}$	·	·	1	·	·	·	1	1	1	·
	window $\frac{2}{3}$	·	·	3	·	2	·	1	·	1	1
Qwen	average over terms	·	·	·	1 (+1 −0, +1 −0)	·	·	1	1	2	·
	$\frac{1}{2}\times$ windows	·	·	·	·	·	·	·	1	·	2
	$2\times$ windows	·	·	·	·	·	·	·	2	·	·
	window $\frac{1}{3}$	·	·	·	1	·	1	2	1	2	2
	window $\frac{2}{3}$	3	·	3	·	·	·	1	1	2	·
DeepSeek	average over terms	·	·	·	6 (+5 −1, +5 −1)	·	1	·	2	2	·
	$\frac{1}{2}\times$ windows	·	·	·	3 (+3 −0, +3 −0)	·	1	·	3	·	·
	$2\times$ windows	·	·	1	·	·	·	·	2	·	2
	window $\frac{1}{3}$	·	1	2	·	·	2	·	2	1	1
	window $\frac{2}{3}$	1	1	2	·	·	1	·	2	·	·
Gemma	average over terms	·	·	·	·	·	3	·	·	1	·
	$\frac{1}{2}\times$ windows	·	·	·	·	·	6	1	·	·	·
	$2\times$ windows	·	·	1	·	·	·	·	·	·	·
	window $\frac{1}{3}$	·	1	2	·	·	2	1	·	2	1
	window $\frac{2}{3}$	·	·	·	·	·	3	·	1	3	·

Canonical form. Two proposals that differ only in notation are one equation: terms with a coefficient of exactly zero are dropped, a monomial factor carried by every term is divided out, and a proposal whose left-hand side differs from the truth's by a monomial in the coordinates is multiplied into the truth's frame. None of these steps moves a coefficient or creates a term. Dividing out a common factor decides 2 of GLM's Van der Pol discoveries and 3 of Qwen's Compound PDE discoveries in the strongest cells; each of these runs also writes the law with the bare time derivative on the left, a candidate the screen ranks below the multiplied form.

D.1 The Inviscid Burgers field

The Inviscid Burgers data are the self-similar solution $u = x/(t-2)$, on which $u\,u_x$, $x\,u_x^2$ and u^2/x are the same column to machine precision and $x\,u_x = u$ holds exactly. The field satisfies $u_t = -x\,u_x^2$ and $u_t = -u^2/x$ as exactly as the law, a single solution need not identify its equation (He et al., 2024; Scholl et al., 2022), and the sparse regression's library is singular on it. The paper scores the true form $u_t = -u\,u_x$ only. Table 13 shows what that costs: in the strongest cells every run of GLM, Qwen and DeepSeek writes the law within 10%, but most runs also write an equivalent form, reached through the closed form of the solution, and the screen keeps it in place of the law by a relative margin of order 10^{-11}. Noise separates the forms: at half the limit every run that writes the law keeps it, and at higher levels u^2/x, which needs no derivative, often fits the noisy field better. Runs from the plot rarely reach the closed form and keep the law more often, which reverses the samples-versus-plot pairs on this system for GLM and Qwen (Appendix E.2).

Structure and coefficients. Support agreement is the Jaccard index J, with precision and support recall reported where needed. Both equations are divided by their own target coefficient, which pins the scale without a free rescaling that would report every coefficient doubled as perfect. ε_∞ takes the

Table 13: Inviscid Burgers, per cell of 10 planned runs: runs discovered when the forms equal to the law on this field are accepted, runs discovered on the true form $u_t = -u\,u_x$, and runs that wrote $u_t = -u\,u_x$ within 10% in any candidate; · marks none.

Cell	Discovered	GLM	Qwen	DeepSeek	Gemma
strongest cell	forms equal here	10	10	10	·
	true form	3	0	3	·
	law, any candidate	10	10	10	·
data withheld	forms equal here	·	·	·	4
	true form	·	·	·	4
	law, any candidate	·	·	·	4
grammar to operands	forms equal here	10	7	7	·
	true form	9	1	4	·
	law, any candidate	10	10	7	·
grammar deleted	forms equal here	10	7	8	·
	true form	9	7	8	·
	law, any candidate	10	10	8	·
example abstract	forms equal here	10	10	10	·
	true form	3	1	2	·
	law, any candidate	10	10	10	·
example poisoned	forms equal here	10	10	1	1
	true form	5	2	0	1
	law, any candidate	10	10	1	1
example is the answer	forms equal here	10	10	8	·
	true form	7	1	3	·
	law, any candidate	10	10	8	·
noise $0.5\times$	forms equal here	10	10	10	1
	true form	10	10	10	1
	law, any candidate	10	10	10	1
noise $1\times$	forms equal here	10	10	10	·
	true form	6	7	4	·
	law, any candidate	10	10	10	·
noise $2\times$	forms equal here	8	10	7	1
	true form	5	9	3	1
	law, any candidate	7	10	7	1

worst relative error over the true terms; it is relative per term, which holds the Allen–Cahn diffusion of 10^{-4} beside reaction coefficients of 5 to the same standard. A true coefficient with its sign reversed is an error of at least 100%. It is reported only for proposals that recovered the structure.

Denominators. Every planned run of the main sweep took place (§3.4); a prompt-arm cell lacks at most 4 runs, requests refused before any reply. A count out of 10 is a coarse estimate of a cell's rate (Table 14).

Table 14: 95% Wilson interval for the rate of a cell that discovers the equation in k of its 10 planned runs.

k	0	1	2	3	4	5	6	7	8	9	10
lower	0	.02	.06	.11	.17	.24	.31	.40	.49	.60	.72
upper	.28	.40	.51	.60	.69	.76	.83	.89	.94	.98	1

D.2 THE 10% BOUND

The 10% bound (§3.4) is uncritical: in the strongest cells a recovered structure is nearly always exact or far off (Table 15), and a bound anywhere between 5% and 10% reclassifies one structure of GLM's and one of DeepSeek's, each a round estimate exactly 10% off, and none of Qwen's or Gemma's.

An exact coefficient can be a reading snapped to a round value. Table 16 gives the estimates behind the shares of §4.1, such as a Van der Pol ratio found between 0.198 and 0.203 and written as 0.2.

The bound also separates reading from wrong recall. Table 17 gives, per system, the set of coefficients written with the data withheld on a surviving true structure that comes closest to the truth without being exact: the nearest is 83% off, and most are off by factors, such as a squared wave speed of 0.25 against 0.04 or a drift of the wrong sign.

Table 15: Recovered structures in each model's strongest cell by their worst true coefficient: exact, within 10% but not exact, and further off.

Model	Worst coefficient	KdV	K–S	Bu_v	Bu_i	VdP	A–C	Wave	Comp	Osc	Rat
GLM	exact	10	10	7	3	10	10	4	5	5	2
	within 10%	0	0	0	0	0	0	0	1	0	0
	further off	0	0	0	0	0	0	1	0	0	0
Qwen	exact	10	10	7	·	10	7	3	6	3	·
	within 10%	0	0	0	·	0	0	0	0	0	·
	further off	0	0	3	·	0	0	2	1	0	·
DeepSeek	exact	9	9	3	3	5	3	10	1	·	·
	within 10%	0	0	0	0	1	1	0	0	·	·
	further off	0	0	5	0	2	2	0	0	·	·
Gemma	exact	1	2	2	0	0	·	8	·	·	·
	within 10%	0	0	0	0	0	·	0	·	·	·
	further off	9	7	4	2	4	·	1	·	·	·

Table 16: Discoveries in each model's strongest cell whose reasoning states an estimate of a true coefficient that differs from the coefficient written (Stated), the mean and maximum relative error of those estimates in percent, and those with every such estimate within 10%.

Model	System	Disc.	Stated	Err. % mean	max	$\leq$10%
GLM	KdV	10	5	3.2	10.0	5
	K–S	10	0	·	·	·
	Bu_v	7	6	13.7	40.0	2
	Bu_i	3	0	·	·	·
	VdP	10	7	2.4	15.0	6
	A–C	10	4	26.5	100	1
	Wave	4	3	6.2	15.0	2
	Comp	6	5	1.7	7.0	5
	Osc	5	2	5.3	15.0	1
	Rat	2	2	0.3	0.6	2
Qwen	KdV	10	1	5.0	5.0	1
	K–S	10	1	15.0	15.0	0
	Bu_v	7	5	28.3	50.0	0
	VdP	10	0	·	·	·
	A–C	7	2	10.0	10.0	2
	Wave	3	2	7.1	10.0	2
	Comp	6	2	8.0	10.0	2
	Osc	3	1	4.0	4.0	1
DeepSeek	KdV	9	1	5.0	5.0	1
	K–S	9	0	·	·	·
	Bu_v	3	2	2.3	3.0	2
	Bu_i	3	0	·	·	·
	VdP	6	0	·	·	·
	A–C	4	1	0.6	0.6	1
	Wave	10	4	1.1	2.5	4
	Comp	1	0	·	·	·
Gemma	KdV	1	0	·	·	·
	K–S	2	0	·	·	·
	Bu_v	2	0	·	·	·
	Wave	8	0	·	·	·

Table 17: For each system, the non-exact coefficients written with the data withheld on a surviving true structure that come closest to the truth: their worst relative error and the models that wrote them. A dash marks a system where none were written.

System	Worst error	Model	Coefficients
Korteweg–de Vries	83%	Qwen, DeepSeek	$u\,u_x$: -1, u_{xxx}: -1
Kuramoto–Sivashinsky	–		
Viscous Burgers	90%	Qwen, DeepSeek, Gemma	$u\,u_x$: -1, u_{xx}: 0.01
Inviscid Burgers	–		
Van der Pol	150%	GLM	u: -1, u^2u_t: -0.5, u_t: 0.5
Allen–Cahn	9900%	Qwen	u: 1, u^3: -1, u_{xx}: 0.01
Wave	525%	GLM	u_{xx}: 0.25
Compound PDE	–		
Forced damped osc.	–		
Rational PDE	100%	Qwen	u_x/x: 0.01, u_{xx}: 0.01

E RESULTS: DETAILS

E.1 THE MAIN SWEEP

Tables 18 to 29 give, per model, the discoveries, the structural recoveries and the ε_∞ of every cell of the main sweep, configurations down and systems across, each run read at its smallest ε_∞ among its surviving true structures.

Table 18: GLM: equations discovered.

Configuration	KdV	K–S	Bu_v	Bu_i	VdP	A–C	Wave	Comp	Osc	Rat
samples	10	10	7	3	10	10	4	6	5	2
plot	8	10	2	4	9	7	1	0	0	0
plot+samples	10	10	8	2	10	9	3	6	0	1
samples, 2 calls	9	9	5	3	9	7	3	3	4	1
plot, 2 calls	9	10	0	7	10	7	0	0	0	0
plot+samples, 2 calls	10	9	6	5	7	7	1	4	0	1
samples+derivs	10	10	8	3	9	9	8	5	0	1
plot+derivs	8	10	2	4	10	1	4	2	0	0
plot+samples+derivs	10	10	4	2	8	4	10	7	0	0
samples+derivs, 2 calls	9	9	5	3	7	5	9	1	1	0
plot+derivs, 2 calls	8	10	1	5	9	0	4	0	0	0
plot+samples+derivs, 2 calls	8	10	3	3	4	0	3	0	0	0

Table 19: Qwen: equations discovered.

Configuration	KdV	K–S	Bu_v	Bu_i	VdP	A–C	Wave	Comp	Osc	Rat
samples	10	10	7	0	10	7	3	6	3	0
plot	10	6	1	4	5	0	0	0	0	0
plot+samples	10	8	3	2	9	5	2	7	2	0
samples, 2 calls	10	10	2	0	10	3	1	7	0	1
plot, 2 calls	8	3	0	4	7	0	0	0	0	0
plot+samples, 2 calls	10	7	4	2	9	5	2	6	0	0
samples+derivs	10	10	5	2	10	3	10	2	3	0
plot+derivs	8	9	0	4	4	0	5	0	0	0
plot+samples+derivs	7	7	2	2	8	1	9	4	0	0
samples+derivs, 2 calls	10	9	8	2	8	3	8	4	0	0
plot+derivs, 2 calls	9	4	2	3	7	0	2	0	0	0
plot+samples+derivs, 2 calls	8	9	2	1	7	0	8	1	0	1

Table 20: DeepSeek: equations discovered.

Configuration	KdV	K–S	Bu_v	Bu_i	VdP	A–C	Wave	Comp	Osc	Rat
samples	9	3	3	0	7	3	1	0	0	0
plot	2	7	0	0	1	0	0	0	0	0
plot+samples	8	9	5	2	10	2	3	4	0	0
samples, 2 calls	3	5	1	7	8	5	1	1	0	0
plot, 2 calls	7	4	0	1	0	0	0	0	0	0
plot+samples, 2 calls	5	7	0	5	7	2	3	1	0	0
samples+derivs	9	9	3	3	6	4	10	1	0	0
plot+derivs	3	2	2	0	1	0	1	0	0	0
plot+samples+derivs	6	5	4	5	7	0	10	0	0	0
samples+derivs, 2 calls	4	8	3	6	7	2	8	0	0	0
plot+derivs, 2 calls	3	3	1	1	2	0	6	0	0	0
plot+samples+derivs, 2 calls	4	5	1	2	1	0	7	0	0	0

Table 21: Gemma: equations discovered.

Configuration	KdV	K–S	Bu_v	Bu_i	VdP	A–C	Wave	Comp	Osc	Rat
samples	1	0	0	1	0	0	0	0	0	0
plot	0	2	0	0	0	0	0	0	0	0
plot+samples	0	1	0	0	0	0	0	0	0	0
samples, 2 calls	0	0	0	0	0	0	0	0	0	0
plot, 2 calls	1	3	0	0	0	0	0	0	0	0
plot+samples, 2 calls	0	0	0	0	0	0	0	0	0	0
samples+derivs	0	0	1	1	0	0	8	1	0	0
plot+derivs	1	2	3	2	0	0	1	0	0	0
plot+samples+derivs	1	2	2	0	0	0	8	0	0	0
samples+derivs, 2 calls	0	1	0	2	0	0	7	0	0	0
plot+derivs, 2 calls	1	3	0	1	0	0	0	0	0	0
plot+samples+derivs, 2 calls	1	3	0	1	0	0	5	0	0	0

Table 22: GLM: structural recoveries.

Configuration	KdV	K–S	Bu_v	Bu_i	VdP	A–C	Wave	Comp	Osc	Rat
samples	10	10	7	3	10	10	5	6	5	2
plot	8	10	5	4	10	7	8	3	0	1
plot+samples	10	10	8	2	10	9	3	6	0	1
samples, 2 calls	9	9	8	3	10	9	3	3	5	1
plot, 2 calls	9	10	7	7	10	9	2	0	0	0
plot+samples, 2 calls	10	9	8	5	7	7	2	5	0	1
samples+derivs	10	10	9	3	10	9	8	5	0	1
plot+derivs	8	10	4	6	10	2	5	3	0	0
plot+samples+derivs	10	10	4	2	9	4	10	7	0	0
samples+derivs, 2 calls	9	9	6	3	9	5	9	2	1	0
plot+derivs, 2 calls	10	10	1	8	10	8	8	4	0	0
plot+samples+derivs, 2 calls	8	10	4	3	5	0	4	0	0	0

Table 23: Qwen: structural recoveries.

Configuration	KdV	K–S	Bu_v	Bu_i	VdP	A–C	Wave	Comp	Osc	Rat
samples	10	10	10	0	10	7	5	7	3	0
plot	10	8	5	5	6	4	3	0	0	0
plot+samples	10	8	8	2	9	8	3	8	2	0
samples, 2 calls	10	10	9	0	10	9	2	7	0	1
plot, 2 calls	8	4	5	7	7	9	2	0	0	0
plot+samples, 2 calls	10	8	8	2	9	9	3	6	0	0
samples+derivs	10	10	9	2	10	3	10	2	3	0
plot+derivs	9	9	4	8	8	4	10	0	0	0
plot+samples+derivs	8	8	5	2	10	2	9	5	0	0
samples+derivs, 2 calls	10	9	8	2	9	8	8	4	0	0
plot+derivs, 2 calls	10	8	4	8	7	8	4	1	0	0
plot+samples+derivs, 2 calls	8	9	4	1	8	6	8	3	0	1

Table 24: DeepSeek: structural recoveries.

Configuration	KdV	K–S	Bu_v	Bu_i	VdP	A–C	Wave	Comp	Osc	Rat
samples	10	6	9	0	9	6	1	0	0	0
plot	2	9	5	0	5	3	0	0	0	0
plot+samples	8	9	9	2	10	8	5	4	0	0
samples, 2 calls	8	6	8	7	9	9	2	1	0	0
plot, 2 calls	8	9	7	2	2	3	0	0	0	0
plot+samples, 2 calls	9	10	9	7	7	7	3	3	0	0
samples+derivs	9	9	8	3	8	6	10	1	0	0
plot+derivs	5	3	7	4	7	2	4	0	0	0
plot+samples+derivs	6	5	8	5	10	5	10	1	0	0
samples+derivs, 2 calls	5	9	5	7	9	6	8	0	1	0
plot+derivs, 2 calls	6	4	5	4	4	3	7	0	0	0
plot+samples+derivs, 2 calls	4	6	5	2	3	5	8	0	0	0

Table 25: Gemma: structural recoveries.

Configuration	KdV	K–S	Bu_v	Bu_i	VdP	A–C	Wave	Comp	Osc	Rat
samples	10	0	0	3	2	2	6	0	0	0
plot	4	7	7	0	0	6	5	3	0	0
plot+samples	6	3	3	0	0	2	4	0	0	0
samples, 2 calls	10	0	0	0	3	7	3	0	0	0
plot, 2 calls	8	5	7	0	1	10	0	0	0	0
plot+samples, 2 calls	8	1	4	1	2	9	3	1	0	0
samples+derivs	9	0	1	1	0	0	9	1	0	0
plot+derivs	7	8	4	3	0	0	7	0	0	0
plot+samples+derivs	10	9	6	2	4	0	9	0	0	0
samples+derivs, 2 calls	10	1	4	2	3	1	7	0	0	0
plot+derivs, 2 calls	10	7	6	3	1	0	2	1	0	0
plot+samples+derivs, 2 calls	10	10	6	3	1	0	5	0	0	0

Table 26: GLM: ε_∞ of the recovered structures, mean (standard deviation).

Configuration	KdV	K–S	Bu_v	Bu_i	VdP	A–C	Wave	Comp	Osc	Rat
samples	0	0	0	0	0	0	.11 (.25)	.017 (.041)	0	0
plot	0	0	.30 (.27)	0	.10 (.32)	0	13 (35)	.78 (.076)	–	1.1
plot+samples	0	0	0	0	0	0	0	0	–	0
samples, 2 calls	0	0	.49 (.80)	0	.26 (.82)	23 (66)	0	0	.22 (.49)	0
plot, 2 calls	0	0	.69 (.20)	.014 (.038)	.015 (.034)	22 (44)	.76 (.37)	–	–	–
plot+samples, 2 calls	0	0	.21 (.40)	0	0	0	12 (17)	.21 (.47)	–	0
samples+derivs	0	0	.039 (.12)	0	.075 (.24)	0	0	0	–	0
plot+derivs	0	0	.30 (.36)	.12 (.12)	.0050 (.016)	7.0 (9.9)	.20 (.45)	.083 (.10)	–	–
plot+samples+derivs	0	0	0	0	.028 (.083)	0	0	0	–	–
samples+derivs, 2 calls	0	0	.083 (.20)	0	.47 (1.3)	0	0	.35 (.49)	0	–
plot+derivs, 2 calls	.25 (.63)	0	0	.14 (.16)	.038 (.091)	120 (280)	3.1 (8.4)	1.5 (.71)	–	–
plot+samples+derivs, 2 calls	0	0	.63 (1.3)	0	.25 (.56)	–	.063 (.13)	–	–	–

Table 27: Qwen: ε_∞ of the recovered structures, mean (standard deviation).

Configuration	KdV	K–S	Bu_v	Bu_i	VdP	A–C	Wave	Comp	Osc	Rat
samples	0	0	.13 (.22)	–	0	0	.14 (.24)	.018 (.047)	0	–
plot	0	.23 (.43)	.47 (.27)	.10 (.22)	.083 (.20)	760 (1,200)	2.2 (2.7)	–	–	–
plot+samples	0	0	.36 (.31)	0	0	.30 (.42)	.083 (.14)	.041 (.057)	0	–
samples, 2 calls	0	0	.38 (.36)	–	0	1.2 (1.5)	.27 (.32)	.029 (.049)	–	0
plot, 2 calls	0	.50 (1.0)	.62 (.027)	.31 (.43)	0	89 (30)	13 (16)	–	–	–
plot+samples, 2 calls	0	.063 (.18)	.17 (.20)	0	0	2.5 (3.9)	.15 (.26)	.017 (.041)	–	–
samples+derivs	0	0	.18 (.23)	0	0	0	0	0	0	–
plot+derivs	.093 (.28)	0	.53 (.050)	.11 (.12)	.60 (.68)	130 (240)	1.6 (3.6)	–	–	–
plot+samples+derivs	.063 (.18)	.10 (.28)	.33 (.30)	0	.050 (.11)	10 (15)	0	.24 (.48)	–	–
samples+derivs, 2 calls	0	0	0	0	.017 (.050)	15 (34)	0	0	–	–
plot+derivs, 2 calls	.067 (.21)	.51 (.63)	.38 (.43)	.12 (.090)	0	34 (40)	.13 (.14)	1.5	–	–
plot+samples+derivs, 2 calls	0	0	.25 (.38)	0	.031 (.088)	27 (35)	0	.17 (.18)	–	0

Table 28: DeepSeek: ε_∞ of the recovered structures, mean (standard deviation).

Configuration	KdV	K–S	Bu_v	Bu_i	VdP	A–C	Wave	Comp	Osc	Rat
samples	.21 (.63)	9.0 (21)	.52 (.77)	–	.20 (.49)	120 (290)	0	–	–	–
plot	0	.13 (.26)	.53 (.084)	–	.40 (.28)	430 (490)	–	–	–	–
plot+samples	0	0	.18 (.23)	0	0	1.3 (2.3)	.26 (.34)	0	–	–
samples, 2 calls	.29 (.30)	.18 (.43)	.45 (.39)	0	.11 (.34)	1.3 (2.9)	.13 (.18)	0	–	–
plot, 2 calls	.13 (.35)	2.0 (4.9)	.85 (.43)	.17 (.24)	.33 (.25)	3,500 (5,600)	–	–	–	–
plot+samples, 2 calls	.21 (.34)	2.7 (6.5)	1.7 (2.8)	.32 (.75)	.014 (.038)	.60 (.73)	.017 (.029)	.64 (.98)	–	–
samples+derivs	0	0	.39 (.37)	0	.10 (.17)	1,800 (4,000)	0	0	–	–
plot+derivs	.13 (.22)	.30 (.52)	.77 (.75)	.24 (.079)	.55 (.28)	14 (0)	.69 (.90)	–	–	–
plot+samples+derivs	0	0	.23 (.32)	0	.20 (.47)	24 (42)	0	.50	–	–
samples+derivs, 2 calls	.040 (.089)	.22 (.67)	.15 (.21)	.062 (.16)	.46 (1.3)	3.4 (7.6)	0	–	.73	–
plot+derivs, 2 calls	.21 (.32)	.28 (.55)	.68 (.54)	.48 (.57)	.31 (.33)	15 (1.2)	.036 (.094)	–	–	–
plot+samples+derivs, 2 calls	0	.33 (.82)	1.9 (2.3)	0	6.8 (10)	270 (370)	.066 (.19)	–	–	–

Table 29: Gemma: ε_∞ of the recovered structures, mean (standard deviation).

Configuration	KdV	K–S	Bu_v	Bu_i	VdP	A–C	Wave	Comp	Osc	Rat
samples	.79 (.31)	–	–	.50 (.50)	1.0 (.35)	3,000 (2,800)	6.3 (9.7)	–	–	–
plot	.94 (.044)	1.1 (.90)	.69 (.17)	–	–	1,500 (1,700)	200 (100)	.83 (.058)	–	–
plot+samples	.91 (.020)	1.2 (1.0)	.63 (.058)	–	–	5,000 (0)	1.8 (1.1)	–	–	–
samples, 2 calls	.90 (.052)	–	–	–	.67 (.14)	140 (53)	9.2 (15)	–	–	–
plot, 2 calls	.81 (.43)	.62 (.91)	1.1 (.37)	–	.75	1,100 (1,500)	–	–	–	–
plot+samples, 2 calls	.89 (.32)	1.0	.70 (0)	1.0	.75 (0)	630 (660)	.75 (0)	.95	–	–
samples+derivs	.96 (.12)	–	0	0	–	–	.028 (.083)	0	–	–
plot+derivs	.72 (.32)	.81 (.57)	.13 (.25)	.067 (.12)	–	–	2.8 (3.9)	–	–	–
plot+samples+derivs	.68 (.28)	1.1 (.82)	.53 (.44)	.75 (.35)	1.0 (.74)	–	2.9 (8.7)	–	–	–
samples+derivs, 2 calls	.96 (.13)	0	1.3 (.49)	0	2.3 (1.4)	1,000	0	–	–	–
plot+derivs, 2 calls	.77 (.28)	1.1 (1.1)	1.8 (.61)	.14 (.15)	1.1	–	1.3 (1.1)	1.1	–	–
plot+samples+derivs, 2 calls	.81 (.29)	1.2 (.92)	1.2 (.39)	.67 (.58)	.50	–	0	–	–	–

E.2 MATCHED PAIRS

A matched pair is two cells of the main sweep that share a model and a system and differ in one factor (Table 30). The samples win more pairs than they lose against the plot on most systems for GLM, Qwen and DeepSeek; the exceptions include Inviscid Burgers for GLM and Qwen, where the plot keeps the true form more often (Appendix D.1). One call wins on more systems than it loses on, for every model. Of the runs that recover the structure, the share that places every coefficient within 10% is 93% with one call and 84% with two for GLM, 79% and 73% for Qwen, 66% and 58% for DeepSeek, and 23% and 16% for Gemma. Table 31 averages the pairs into differences in the rate of discovery, with 95% intervals from a two-stage cluster bootstrap that resamples the systems and then the runs within each cell, and checks them against a logistic GEE per model with an exchangeable correlation within a system; the two agree in sign for every model and factor.

Table 30: Matched pairs per model and system: cells sharing everything but one factor, counted by whether the first discovers more equations, as many, or fewer.

Model	Pairs	KdV	K–S	Bu_v	Bu_i	VdP	A–C	Wave	Comp	Osc	Rat
samples vs plot											
GLM	first better	3	0	4	0	1	3	4	4	3	3
	level	1	2	0	0	0	1	0	0	1	1
	second better	0	2	0	4	3	0	0	0	0	0
Qwen	first better	3	4	4	0	4	4	4	4	2	1
	level	1	0	0	0	0	0	0	0	2	3
	second better	0	0	0	4	0	0	0	0	0	0
DeepSeek	first better	3	3	4	3	4	4	4	2	0	0
	level	0	0	0	1	0	0	0	2	4	4
	second better	1	1	0	0	0	0	0	0	0	0
Gemma	first better	1	0	0	2	0	0	2	1	0	0
	level	0	0	3	1	4	4	2	3	4	4
	second better	3	4	1	1	0	0	0	0	0	0
one call vs two											
GLM	first better	3	3	6	0	5	5	4	5	1	2
	level	2	3	0	2	0	1	1	1	4	4
	second better	1	0	0	4	1	0	1	0	1	0
Qwen	first better	1	4	2	2	2	2	4	2	3	0
	level	3	1	1	4	2	4	2	2	3	4
	second better	2	1	3	0	2	0	0	2	0	2
DeepSeek	first better	4	3	4	1	3	1	2	2	0	0
	level	1	1	2	0	0	4	3	3	6	6
	second better	1	2	0	5	3	1	1	1	0	0
Gemma	first better	1	1	3	2	0	0	3	1	0	0
	level	4	1	3	2	6	6	3	5	6	6
	second better	1	4	0	2	0	0	0	0	0	0
derivs vs none											
GLM	first better	0	1	2	0	1	0	6	2	0	0
	level	4	5	2	4	0	0	0	1	4	2
	second better	2	0	2	2	5	6	0	3	2	4
Qwen	first better	1	3	2	2	0	0	6	0	0	1
	level	2	1	0	2	2	3	0	2	5	4
	second better	3	2	4	2	4	3	0	4	1	1
DeepSeek	first better	2	2	4	2	1	1	6	1	0	0
	level	1	0	1	2	1	2	0	2	6	6
	second better	3	4	1	2	4	3	0	3	0	0
Gemma	first better	3	3	3	4	0	0	5	1	0	0
	level	2	3	3	2	6	6	1	5	6	6
	second better	1	0	0	0	0	0	0	0	0	0

Table 31: Difference in discovery rate between the two levels of each factor: the mean over a model's matched pairs with a 95% two-stage cluster bootstrap interval, the systems where it is positive, zero or negative, and the average marginal difference of a logistic GEE clustered on the system with a 95% robust interval.

Model	Factor, first vs second	Matched pairs	Systems +	0	−	GEE
GLM	samples vs plot	$+.17$ [$+.02$, $+.32$]	7	0	3	$+.17$ [$+.04$, $+.31$]
	derivatives supplied vs not	$-.06$ [$-.20$, $+.08$]	2	0	8	$-.06$ [$-.19$, $+.07$]
	one call vs two	$+.11$ [$+.02$, $+.19$]	8	1	1	$+.10$ [$+.03$, $+.18$]
Qwen	samples vs plot	$+.26$ [$+.09$, $+.39$]	9	0	1	$+.25$ [$+.11$, $+.40$]
	derivatives supplied vs not	0 [$-.12$, $+.15$]	4	1	5	0 [$-.13$, $+.13$]
	one call vs two	$+.05$ [$-.008$, $+.10$]	6	2	2	$+.05$ [$+.006$, $+.09$]
DeepSeek	samples vs plot	$+.23$ [$+.12$, $+.36$]	8	2	0	$+.23$ [$+.12$, $+.34$]
	derivatives supplied vs not	$+.02$ [$-.08$, $+.16$]	3	2	5	$+.02$ [$-.10$, $+.14$]
	one call vs two	$+.04$ [$-.04$, $+.11$]	5	4	1	$+.04$ [$-.03$, $+.10$]
Gemma	samples vs plot	$+.007$ [$-.07$, $+.10$]	3	4	3	$+.007$ [$-.08$, $+.09$]
	derivatives supplied vs not	$+.08$ [$+.01$, $+.18$]	6	4	0	$+.08$ [$-.007$, $+.16$]
	one call vs two	$+.01$ [$-.02$, $+.05$]	3	6	1	$+.01$ [$-.01$, $+.04$]

E.3 Factors of the sweep

The main sweep is a balanced full factorial: four models, ten systems, three inputs, derivative fields off or on, and one call or two, with 10 runs in every cell. The sum of squares of the discovery indicator over its 4,800 runs therefore splits exactly into one term per factor and per interaction, computed from the marginal means, and a term between runs of the same cell (Table 32). A term's share is its η^2, and its F test uses the mean square between runs. A binomial logistic model gives the main effects in the same order, system, model, input shown, calls and derivative fields, and again makes the interaction of the system and the derivative fields the largest term that involves the configuration. Figure 5a repeats the decomposition within each model. Figure 5b gives the effect of each factor per system: the difference in the rate of discovery between its two levels, averaged over the cells that share the model, the system and the other factors; for the three factors of Table 31, its average over the systems is the matched-pair mean there. The plot added to the samples changes the rate of discovery by $-.09$ [$-.18$, $-.01$] for GLM, $-.09$ [$-.17$, $-.01$] for Qwen, $-.03$ [$-.10$, $+.05$] for DeepSeek and $+.00$ [$-.04$, $+.04$] for Gemma, with 95% intervals from the same two-stage cluster bootstrap.

Table 32: The variation in discovery over the 4,800 runs of the main sweep, by term: degrees of freedom, share of the sum of squares (η^2), F test against the variation between runs of a cell, and the share of the null deviance a logistic model loses without the term.

Term	df	η^2 (%)	F	p	Deviance (%)
system	9	20.1	230.1	$<.001$	39.5
model	3	12.7	438.3	$<.001$	28.0
input shown	2	2.3	119.8	$<.001$	5.5
derivative fields	1	0.0	1.3	.250	0.0
calls	1	0.3	32.1	$<.001$	0.8
model × system	27	7.6	29.1	$<.001$	5.4
system × input shown	18	1.4	8.1	$<.001$	4.0
system × derivative fields	9	3.9	44.2	$<.001$	7.1
system × calls	9	0.3	3.8	$<.001$	0.7
model × input shown	6	0.9	15.3	$<.001$	1.3
model × derivative fields	3	0.3	10.7	$<.001$	0.7
model × calls	3	0.1	4.5	.004	0.2
input shown × derivative fields	2	0.2	9.6	$<.001$	0.7
input shown × calls	2	0.2	8.3	$<.001$	0.3
derivative fields × calls	1	0.0	1.8	.182	0.0
three-way and higher	383	7.8	2.1	$<.001$	–
between runs of a cell	4320	41.8	–	–	–

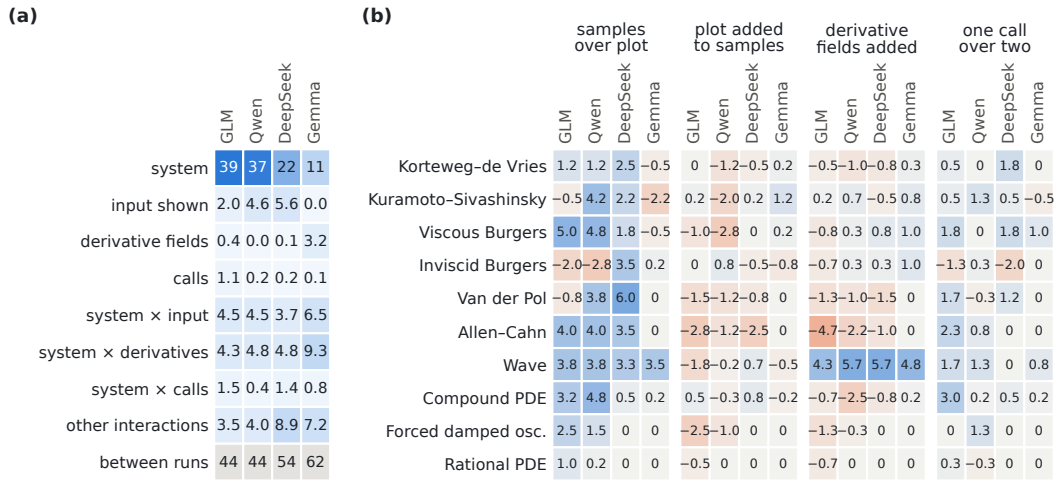


Figure 5: (a) Share of the variation in discovery, in %, that each factor of the main sweep and its interaction with the system explain within each model, and the share between runs of the same cell. (b) Change in discoveries, out of 10 runs, from one level of a factor to the other, per system, averaged over the matched cells.

E.4 How discoveries are made

Moves. The prompt asks for reasoning in four steps before the answer (Appendix B.1), and the tables below count the moves that written reasoning makes, in eight families: naming a law; taking coefficients from a standard form; reading values at points, including a balance where the other terms vanish; testing that a coefficient stays constant or verifying it on data not used to set it; reading a scale or a closed form; ruling out a mechanism by a signature of the data; balancing magnitudes only; and acknowledging or suppressing noise. The hidden traces of three models are not in the archive.

Reading the tables. Tables 33 to 36 set the replies of discovering runs against the other replies of each strongest cell. Every model names the law in at least four fifths of its discoveries. Coefficients from a standard form appear in at most a fifth of any model's discoveries, mainly on Korteweg–de Vries and Kuramoto–Sivashinsky. GLM tests or checks in nearly every discovery and the other models in under half. Among the runs that miss the law, Gemma's replies balance magnitudes in a third and the other models' in at most a seventh, and Gemma's take the coefficients from a standard form more often than any other model's.

Failing replies. The families appear in failing replies as well. On the oscillator and the Rational PDE the failing replies of Qwen and DeepSeek read points and scales as often as the successful ones and never write the missing term, and on the Compound PDE most failures keep the diffusivity constant. Where the structure is recovered and a coefficient misses the bound, GLM, Qwen and DeepSeek test it no more often than the discoveries of the same cell.

Models on the same input. Table 37 compares the four models on the one input they all received, the samples in one call. Appeals to a standard form concentrate on Korteweg–de Vries and Kuramoto–Sivashinsky, and on the Compound and Rational PDEs, two of the five systems where Gemma returns an equation other than the one in the data in every reply; there it names the heat equation.

Inputs within a model. Table 38 compares inputs within a model and a system as matched pairs. Supplying the derivative fields raises the share of replies that read values at points on eight systems for GLM, seven for Qwen and DeepSeek and nine for Gemma, and lowers the share that reads a scale on five systems for GLM, seven for Qwen and all ten for DeepSeek and Gemma; showing the samples rather than the plot raises the share that tests or checks for GLM, DeepSeek and Gemma and lowers it for Qwen.

Table 33: GLM, strongest cell: replies using each family of moves, among the replies of the runs that discover the equation and among the other replies; the first row counts the replies of each kind.

Family	Replies	KdV	K–S	Bu_v	Bu_i	VdP	A–C	Wave	Comp	Osc	Rat
replies	discovering	10	10	7	3	10	10	4	6	5	2
	other	0	0	1	7	0	0	2	2	2	2
names the law	discovering	10	10	7	3	10	10	4	6	2	0
	other	0	0	1	7	0	0	2	2	0	1
takes coefficients from a standard form	discovering	·	4	·	·	·	·	·	0	·	·
	other	·	0	·	·	·	·	·	1	·	·
reads values at points	discovering	6	8	7	3	10	10	4	6	5	2
	other	0	0	1	7	0	0	2	2	1	1
tests constancy or verifies	discovering	10	8	7	3	10	9	3	6	4	2
	other	0	0	1	6	0	0	1	1	2	2
reads a scale or closed form	discovering	10	10	6	3	10	10	4	6	5	2
	other	0	0	1	7	0	0	1	2	2	2
rules out by a signature	discovering	9	9	7	3	7	8	4	5	4	2
	other	0	0	0	7	0	0	1	2	2	2
balances magnitudes	discovering	·	1	1	·	·	·	0	1	·	·
	other	·	0	0	·	·	·	1	0	·	·
acknowledges or suppresses noise	discovering	3	·	3	·	4	·	·	0	2	1
	other	0	·	0	·	0	·	·	1	0	0

Table 34: Qwen, strongest cell: replies using each family of moves, among the replies of the runs that discover the equation and among the other replies; the first row counts the replies of each kind.

Family	Replies	KdV	K–S	Bu_v	Bu_i	VdP	A–C	Wave	Comp	Osc	Rat
replies	discovering	10	10	7	0	10	7	3	6	3	0
	other	0	0	3	10	0	3	7	4	7	10
names the law	discovering	10	10	7	0	10	3	3	2	·	0
	other	0	0	3	10	0	1	5	3	·	8
takes coefficients from a standard form	discovering	4	4	·	·	·	0	1	0	·	0
	other	0	0	·	·	·	1	0	2	·	1
reads values at points	discovering	6	4	3	0	10	7	3	6	3	0
	other	0	0	2	10	0	3	7	4	6	9
tests constancy or verifies	discovering	4	3	3	0	6	2	0	5	2	0
	other	0	0	1	2	0	1	3	2	2	5
reads a scale or closed form	discovering	9	6	5	0	7	4	2	3	3	0
	other	0	0	2	10	0	1	4	4	7	8
rules out by a signature	discovering	9	8	6	0	9	7	2	6	1	0
	other	0	0	2	8	0	2	7	3	5	8
balances magnitudes	discovering	·	7	3	·	2	1	1	·	·	0
	other	·	0	0	·	0	1	2	·	·	2
acknowledges or suppresses noise	discovering	2	·	·	0	1	·	·	·	0	·
	other	0	·	·	1	0	·	·	·	1	·

Tests and the reply limit. Every run with no answer spent the provider's whole reply budget, nearly all of it in the hidden trace, which is not archived; what those runs did cannot be read. The replies that answered show what testing costs. Within a cell, GLM's replies that test or check are longer than its other replies in 34 of the 50 cells that hold both (sign test $p = .015$), with medians of $47{,}000$ and $29{,}000$ completion tokens, and 17% of them end above three quarters of the $131{,}072$-token limit, against 8% of the others. For Qwen the replies that test are shorter, and for DeepSeek and Gemma about as long. By input, GLM tests in 95% of the replies shown the samples, 93% of those shown the samples with the plot and 86% of those shown the plot alone; their median lengths are $61{,}000$, $53{,}000$ and $35{,}000$ tokens, and the runs with no answer 25%, 37% and 2%.

Best guess and alternatives. Each reply states a best guess first and four alternatives (Table 39). In the strongest cells an alternative alone makes none of GLM's discoveries, about a tenth of Qwen's and DeepSeek's and nearly half of Gemma's, half or more of them on the Wave equation; across the three noise levels it makes a seventh of Qwen's, over two fifths of Gemma's and almost none of GLM's or DeepSeek's.

Table 35: DeepSeek, strongest cell: replies using each family of moves, among the replies of the runs that discover the equation and among the other replies; the first row counts the replies of each kind.

Family	Replies	KdV	K–S	Bu_v	Bu_i	VdP	A–C	Wave	Comp	Osc	Rat
replies	discovering	9	9	3	3	6	4	10	1	0	0
	other	1	1	7	7	4	6	0	8	10	10
names the law	discovering	9	9	3	3	6	4	9	1	0	0
	other	1	1	6	7	4	5	0	5	2	4
takes coefficients from a standard form	discovering	3	4	·	·	·	1	1	·	·	0
	other	0	0	·	·	·	2	0	·	·	1
reads values at points	discovering	5	4	2	2	6	2	7	1	0	0
	other	1	1	1	7	4	4	0	8	5	8
tests constancy or verifies	discovering	8	1	3	2	3	·	3	1	0	0
	other	0	1	0	5	1	·	0	4	3	1
reads a scale or closed form	discovering	6	4	1	3	3	2	2	0	0	0
	other	0	1	2	7	3	1	0	3	9	6
rules out by a signature	discovering	7	6	2	1	5	2	9	1	0	0
	other	0	1	4	6	2	4	0	7	7	8
balances magnitudes	discovering	1	2	0	·	·	0	2	·	0	0
	other	0	0	2	·	·	1	0	·	2	3
acknowledges or suppresses noise	discovering	0	·	1	1	1	·	3	·	·	·
	other	1	·	0	2	0	·	0	·	·	·

Table 36: Gemma, strongest cell: replies using each family of moves, among the replies of the runs that discover the equation and among the other replies; the first row counts the replies of each kind.

Family	Replies	KdV	K–S	Bu_v	Bu_i	VdP	A–C	Wave	Comp	Osc	Rat
replies	discovering	1	2	2	0	0	0	8	0	0	0
	other	9	8	8	10	10	10	2	10	10	10
names the law	discovering	1	2	2	·	0	0	8	0	·	0
	other	9	8	8	·	9	10	1	10	·	10
takes coefficients from a standard form	discovering	0	2	0	·	·	·	·	0	0	·
	other	3	5	1	·	·	·	·	3	1	·
reads values at points	discovering	1	1	2	0	0	0	8	0	0	0
	other	6	4	6	10	10	9	2	9	8	10
tests constancy or verifies	discovering	0	1	1	0	0	0	3	0	0	0
	other	4	3	5	10	4	1	1	9	9	4
reads a scale or closed form	discovering	1	·	0	0	0	·	·	·	0	0
	other	7	·	2	6	10	·	·	·	10	1
rules out by a signature	discovering	1	2	2	0	0	0	6	0	0	0
	other	9	6	6	7	8	5	2	8	3	9
balances magnitudes	discovering	1	1	1	·	0	0	6	0	0	0
	other	5	5	3	·	1	9	1	1	2	2
acknowledges or suppresses noise	discovering	·	1	·	0	·	0	·	0	·	0
	other	·	2	·	3	·	1	·	1	·	1

Table 37: Every model on the same input, the samples in one call: replies using each family of moves, and all replies.

Model	Replies	KdV	K–S	Bu_v	Bu_i	VdP	A–C	Wave	Comp	Osc	Rat
names the law											
GLM	using it	10	10	8	10	10	10	6	8	2	1
	all	10	10	8	10	10	10	6	8	7	4
Qwen	using it	10	10	10	10	10	4	8	5	0	8
	all	10	10	10	10	10	10	10	10	10	10
DeepSeek	using it	10	9	9	10	10	8	5	4	1	5
	all	10	9	10	10	10	10	10	10	10	10
Gemma	using it	10	10	10	0	5	10	10	10	4	10
	all	10	10	10	10	10	10	10	10	10	10
standard-form coefficients											
GLM	using it	0	4	0	0	0	0	0	1	0	0
	all	10	10	8	10	10	10	6	8	7	4
Qwen	using it	4	4	0	0	0	1	1	2	0	1
	all	10	10	10	10	10	10	10	10	10	10
DeepSeek	using it	4	1	2	0	0	0	0	0	0	0
	all	10	9	10	10	10	10	10	10	10	10
Gemma	using it	4	7	2	0	0	1	0	2	0	1
	all	10	10	10	10	10	10	10	10	10	10
values at points											
GLM	using it	6	8	8	10	10	10	6	8	6	3
	all	10	10	8	10	10	10	6	8	7	4
Qwen	using it	6	4	5	10	10	10	10	10	9	9
	all	10	10	10	10	10	10	10	10	10	10
DeepSeek	using it	3	4	2	9	10	8	8	9	6	3
	all	10	9	10	10	10	10	10	10	10	10
Gemma	using it	2	6	7	9	5	7	8	9	3	7
	all	10	10	10	10	10	10	10	10	10	10

Table 37 (continued).

Model	Replies	KdV	K–S	Bu_v	Bu_i	VdP	A–C	Wave	Comp	Osc	Rat
tests or checks											
GLM	using it	10	8	8	9	10	9	4	7	6	4
	all	10	10	8	10	10	10	6	8	7	4
Qwen	using it	4	3	4	2	6	3	3	7	4	5
	all	10	10	10	10	10	10	10	10	10	10
DeepSeek	using it	2	3	2	5	3	4	4	5	2	4
	all	10	9	10	10	10	10	10	10	10	10
Gemma	using it	2	2	7	7	4	4	3	7	1	6
	all	10	10	10	10	10	10	10	10	10	10
scale or closed form											
GLM	using it	10	10	7	10	10	10	5	8	7	4
	all	10	10	8	10	10	10	6	8	7	4
Qwen	using it	9	6	7	10	7	5	6	7	10	8
	all	10	10	10	10	10	10	10	10	10	10
DeepSeek	using it	7	7	6	10	9	4	3	6	10	8
	all	10	9	10	10	10	10	10	10	10	10
Gemma	using it	8	5	3	8	10	6	2	7	10	9
	all	10	10	10	10	10	10	10	10	10	10
magnitudes only											
GLM	using it	0	1	1	0	0	0	1	1	0	0
	all	10	10	8	10	10	10	6	8	7	4
Qwen	using it	0	7	3	0	2	2	3	0	0	2
	all	10	10	10	10	10	10	10	10	10	10
DeepSeek	using it	0	1	0	0	0	2	0	0	0	0
	all	10	9	10	10	10	10	10	10	10	10
Gemma	using it	0	3	1	0	0	2	2	0	0	1
	all	10	10	10	10	10	10	10	10	10	10

Table 38: Matched pairs within a model and a system: pairs in which the first input has the higher, the same or the lower share of replies using the family.

Model	First input	KdV	K–S	Bu_v	Bu_i	VdP	A–C	Wave	Comp	Osc	Rat
samples vs plot: values at points											
GLM	higher	2	2	3	0	2	1	1	1	1	2
	same	1	1	1	4	2	1	2	3	2	1
	lower	1	1	0	0	0	2	1	0	0	1
Qwen	higher	1	1	1	1	3	1	3	4	3	1
	same	0	0	0	0	1	0	0	0	0	0
	lower	3	3	3	3	0	3	1	0	1	3
DeepSeek	higher	1	3	1	3	3	1	0	3	1	1
	same	0	0	1	1	0	2	0	0	1	0
	lower	3	1	2	0	1	1	4	1	2	3
Gemma	higher	2	3	4	2	3	3	1	2	2	3
	same	0	0	0	2	0	1	3	2	2	1
	lower	2	1	0	0	1	0	0	0	0	0
samples vs plot: tests or checks											
GLM	higher	2	3	3	2	0	3	3	3	3	1
	same	2	1	0	1	4	0	0	0	0	2
	lower	0	0	1	1	0	1	1	1	0	1
Qwen	higher	0	0	0	0	1	1	0	1	2	3
	same	1	1	0	0	2	0	1	0	0	0
	lower	3	3	4	4	1	3	3	3	2	1
DeepSeek	higher	3	3	2	4	3	2	4	3	2	3
	same	1	0	1	0	0	0	0	1	2	0
	lower	0	1	1	0	1	2	0	0	0	1
Gemma	higher	4	2	2	4	4	4	4	4	4	4
	same	0	1	0	0	0	0	0	0	0	0
	lower	0	1	2	0	0	0	0	0	0	0
samples vs plot: scale or closed form											
GLM	higher	1	0	1	1	0	2	2	3	0	0
	same	3	4	1	3	4	0	1	1	3	3
	lower	0	0	2	0	0	2	1	0	0	1
Qwen	higher	1	0	2	4	0	1	2	2	1	2
	same	0	1	0	0	1	0	0	1	3	1
	lower	3	3	2	0	3	3	2	1	0	1
DeepSeek	higher	3	1	1	4	0	2	2	1	0	3
	same	0	0	2	0	1	0	0	0	3	0
	lower	1	3	1	0	3	2	2	3	1	1
Gemma	higher	2	1	0	4	0	3	3	0	0	3
	same	0	0	2	0	3	0	1	0	2	0
	lower	2	3	2	0	1	1	0	4	2	1

Table 38 (continued).

Model	First input	KdV	K–S	Bu_v	Bu_i	VdP	A–C	Wave	Comp	Osc	Rat
samples vs plot: standard-form coefficients											
GLM	higher	0	0	1	0	0	0	0	1	0	0
	same	0	3	0	3	4	3	3	2	3	4
	lower	4	1	3	1	0	1	1	1	0	0
Qwen	higher	1	3	1	1	1	1	2	2	1	1
	same	1	0	2	3	1	0	1	1	3	2
	lower	2	1	1	0	2	3	1	1	0	1
DeepSeek	higher	2	3	2	0	0	1	2	0	0	2
	same	0	0	2	4	4	1	0	2	4	1
	lower	2	1	0	0	0	2	2	2	0	1
Gemma	higher	4	1	4	0	2	1	1	1	1	1
	same	0	0	0	2	1	2	1	2	1	2
	lower	0	3	0	2	1	1	2	1	2	1
derivatives vs none: values at points											
GLM	higher	3	4	4	0	1	3	1	2	1	4
	same	2	1	1	6	5	2	4	4	3	1
	lower	1	1	1	0	0	1	1	0	0	0
Qwen	higher	5	5	5	2	2	4	1	4	3	5
	same	0	0	1	0	3	1	1	1	0	1
	lower	1	1	0	4	1	1	4	1	3	0
DeepSeek	higher	5	4	4	2	0	1	3	5	3	3
	same	0	1	0	2	1	0	1	0	1	1
	lower	1	1	2	2	5	5	2	1	2	2
Gemma	higher	4	2	4	4	5	6	3	5	6	5
	same	1	0	0	2	0	0	3	1	0	1
	lower	1	4	2	0	1	0	0	0	0	0
derivatives vs none: tests or checks											
GLM	higher	1	2	2	1	0	3	2	3	1	1
	same	5	1	2	3	6	1	2	2	2	1
	lower	0	3	2	2	0	2	2	1	1	3
Qwen	higher	5	1	3	2	0	3	2	4	1	1
	same	0	1	0	1	1	0	1	0	0	1
	lower	1	4	3	3	5	3	3	2	5	4
DeepSeek	higher	2	2	3	4	3	1	3	2	4	1
	same	1	1	2	2	0	0	0	1	0	2
	lower	3	3	1	0	3	5	3	3	2	3
Gemma	higher	3	3	3	4	4	2	5	3	6	2
	same	3	1	1	2	1	0	1	2	0	2
	lower	0	2	2	0	1	4	0	1	0	2

Table 38 (continued).

Model	First input	KdV	K–S	Bu_v	Bu_i	VdP	A–C	Wave	Comp	Osc	Rat
derivatives vs none: scale or closed form											
GLM	higher	0	0	1	1	0	3	1	1	0	1
	same	5	6	2	5	6	0	1	1	4	1
	lower	1	0	3	0	0	3	4	4	0	3
Qwen	higher	3	0	4	0	2	2	1	0	0	1
	same	0	0	1	1	3	0	1	0	5	0
	lower	3	6	1	5	1	4	4	6	1	5
DeepSeek	higher	1	0	0	0	1	0	0	1	1	0
	same	0	0	2	4	1	1	2	0	2	1
	lower	5	6	4	2	4	5	4	5	3	5
Gemma	higher	2	0	0	0	0	0	0	0	0	0
	same	1	0	1	2	5	0	3	0	4	0
	lower	3	6	5	4	1	6	3	6	2	6
derivatives vs none: standard-form coefficients											
GLM	higher	3	0	1	0	0	0	0	0	0	0
	same	2	1	4	5	6	5	4	4	4	4
	lower	1	5	1	1	0	1	2	2	0	1
Qwen	higher	4	4	4	1	0	3	1	3	1	3
	same	0	1	1	5	5	0	2	1	4	2
	lower	2	1	1	0	1	3	3	2	1	1
DeepSeek	higher	1	1	2	0	0	2	2	0	0	2
	same	2	1	3	6	6	1	2	3	6	2
	lower	3	4	1	0	0	3	2	3	0	2
Gemma	higher	4	1	3	0	1	0	1	1	3	1
	same	0	1	0	3	3	1	1	2	2	1
	lower	2	4	3	3	2	5	4	3	1	4

Table 39: Discoveries made by the model's first candidate, and those made only by one of its alternatives; · marks no discovery.

Model	Cell	Made by	KdV	K–S	Bu_v	Bu_i	VdP	A–C	Wave	Comp	Osc	Rat
GLM	samples	first candidate	10	10	7	3	10	10	4	6	5	2
		alternative only	0	0	0	0	0	0	0	0	0	0
	control, $0.5\times$	first candidate	9	5	4	10	6	3	·	3	·	·
		alternative only	0	0	0	0	0	0	·	0	·	·
	control, $1\times$	first candidate	10	8	·	6	5	4	·	2	·	·
		alternative only	0	0	·	0	0	0	·	0	·	·
	control, $2\times$	first candidate	7	3	2	4	10	4	1	·	·	·
		alternative only	0	0	0	1	0	0	0	·	·	·
Qwen	samples	first candidate	10	8	6	·	10	7	0	6	3	·
		alternative only	0	2	1	·	0	0	3	0	0	·
	control, $0.5\times$	first candidate	10	7	5	9	6	·	0	5	·	·
		alternative only	0	2	0	1	0	·	2	1	·	·
	control, $1\times$	first candidate	9	6	2	7	5	·	·	4	·	·
		alternative only	0	4	0	0	1	·	·	4	·	·
	control, $2\times$	first candidate	9	8	6	8	9	·	·	4	·	·
		alternative only	1	0	1	1	0	·	·	2	·	·
DeepSeek	samples+derivs	first candidate	9	8	3	3	6	4	7	1	·	·
		alternative only	0	1	0	0	0	0	3	0	·	·
	control, $0.5\times$	first candidate	8	10	6	10	3	·	·	·	·	·
		alternative only	1	0	0	0	0	·	·	·	·	·
	control, $1\times$	first candidate	10	9	9	4	3	·	·	1	·	·
		alternative only	0	0	0	0	0	·	·	1	·	·
	control, $2\times$	first candidate	5	8	6	3	3	2	·	3	·	·
		alternative only	0	0	0	0	0	0	·	0	·	·
Gemma	plot+samples+derivs	first candidate	1	2	0	·	·	·	4	·	·	·
		alternative only	0	0	2	·	·	·	4	·	·	·
	control, $0.5\times$	first candidate	·	2	0	0	·	·	·	·	·	·
		alternative only	·	0	2	1	·	·	·	·	·	·
	control, $1\times$	first candidate	3	·	·	·	·	·	·	1	·	·
		alternative only	1	·	·	·	·	·	·	0	·	·
	control, $2\times$	first candidate	1	1	0	1	·	·	·	·	·	·
		alternative only	1	0	2	0	·	·	·	·	·	·

E.5 OPEN POOL

Each run is traced to the earliest stage at which a true term stopped: never conceived, when a term appears in no candidate and no reasoning; named but never written; written and then screened out; every true term surviving with no member holding exactly the true active set; and the true active set surviving with a coefficient more than 10% off. Each planned run is counted once, under the first way it failed, and the rest are the discoveries (Tables 40 to 44). Every run with no answer had a call cut at the provider's reply limit before it answered; nearly all of GLM's such runs sit in the configurations that carry the samples. On the oscillator, the Rational PDE and the Compound PDE the misses of Qwen, DeepSeek and Gemma are almost all never conceived, and a cell that never recovers a structure rarely has a run in which every true term reached a candidate. Appendix F sets the reasoning of these misses against that of the discoveries, system by system. On Inviscid Burgers most misses of GLM, Qwen and DeepSeek are written and then screened out, the equivalent forms of Appendix D.1 displacing the law. The arms are counted more coarsely, as never run, no answer, structure missed and coefficients off; a run never run is a request refused before any reply.

Table 40: GLM: failed runs per cell, by the stage at which a true term stopped: no answer; never conceived; named but never written; written, then dominated; on the front, but no member the law; the structure with coefficients off.

Configuration	Runs	KdV	K–S	Bu_v	Bu_i	VdP	A–C	Wave	Comp	Osc	Rat
samples	no answer	·	·	2	0	·	·	4	2	3	6
	never conceived	·	·	0	0	·	·	0	2	2	2
	named only	·	·	0	0	·	·	0	0	0	0
	dominated	·	·	1	7	·	·	1	0	0	0
	front lacks it	·	·	0	0	·	·	0	0	0	0
	coefficients off	·	·	0	0	·	·	1	0	0	0
plot	no answer	1	·	0	0	0	0	0	0	0	0
	never conceived	1	·	3	4	0	0	0	7	10	9
	named only	0	·	0	0	0	0	1	0	0	0
	dominated	0	·	1	2	0	3	0	0	0	0
	front lacks it	0	·	1	0	0	0	1	0	0	0
	coefficients off	0	·	3	0	1	0	7	3	0	1
plot+samples	no answer	·	·	2	0	·	0	7	3	8	7
	never conceived	·	·	0	0	·	0	0	1	2	2
	named only	·	·	0	0	·	0	0	0	0	0
	dominated	·	·	0	8	·	1	0	0	0	0
	front lacks it	·	·	0	0	·	0	0	0	0	0
	coefficients off	·	·	0	0	·	0	0	0	0	0
samples, 2 calls	no answer	0	0	2	0	0	1	7	3	4	8
	never conceived	1	1	0	0	0	0	0	4	1	1
	named only	0	0	0	0	0	0	0	0	0	0
	dominated	0	0	0	7	0	0	0	0	0	0
	front lacks it	0	0	0	0	0	0	0	0	0	0
	coefficients off	0	0	3	0	1	2	0	0	1	0
plot, 2 calls	no answer	0	·	0	0	·	0	0	0	2	0
	never conceived	0	·	0	2	·	0	0	10	8	9
	named only	0	·	1	1	·	0	0	0	0	0
	dominated	1	·	2	0	·	1	4	0	0	1
	front lacks it	0	·	0	0	·	0	4	0	0	0
	coefficients off	0	·	7	0	·	2	2	0	0	0
plot+samples, 2 calls	no answer	·	0	2	0	2	2	8	5	8	7
	never conceived	·	1	0	0	0	0	0	0	2	2
	named only	·	0	0	0	0	0	0	0	0	0
	dominated	·	0	0	5	1	1	0	0	0	0
	front lacks it	·	0	0	0	0	0	0	0	0	0
	coefficients off	·	0	2	0	0	0	1	1	0	0

Table 40 (continued).

Configuration	Runs	KdV	K–S	Bu_v	Bu_i	VdP	A–C	Wave	Comp	Osc	Rat
samples+derivs	no answer	·	·	1	0	0	1	2	4	10	5
	never conceived	·	·	0	0	0	0	0	1	0	4
	named only	·	·	0	0	0	0	0	0	0	0
	dominated	·	·	0	7	0	0	0	0	0	0
	front lacks it	·	·	0	0	0	0	0	0	0	0
	coefficients off	·	·	1	0	1	0	0	0	0	0
plot+derivs	no answer	0	·	0	0	·	0	1	0	0	0
	never conceived	0	·	0	0	·	2	0	7	10	10
	named only	0	·	1	0	·	0	0	0	0	0
	dominated	2	·	4	3	·	6	0	0	0	0
	front lacks it	0	·	1	1	·	0	4	0	0	0
	coefficients off	0	·	2	2	·	1	1	1	0	0
plot+samples+derivs	no answer	·	·	6	0	0	6	·	3	9	9
	never conceived	·	·	0	0	1	0	·	0	1	1
	named only	·	·	0	0	0	0	·	0	0	0
	dominated	·	·	0	8	0	0	·	0	0	0
	front lacks it	·	·	0	0	0	0	·	0	0	0
	coefficients off	·	·	0	0	1	0	·	0	0	0
samples+derivs, 2 calls	no answer	1	1	4	0	1	5	1	7	8	7
	never conceived	0	0	0	0	0	0	0	1	1	3
	named only	0	0	0	0	0	0	0	0	0	0
	dominated	0	0	0	7	0	0	0	0	0	0
	front lacks it	0	0	0	0	0	0	0	0	0	0
	coefficients off	0	0	1	0	2	0	0	1	0	0
plot+derivs, 2 calls	no answer	0	·	0	1	0	1	0	0	3	0
	never conceived	0	·	0	1	0	1	0	6	7	10
	named only	0	·	0	0	0	0	0	0	0	0
	dominated	0	·	5	0	0	0	0	0	0	0
	front lacks it	0	·	4	0	0	0	2	0	0	0
	coefficients off	2	·	0	3	1	8	4	4	0	0
plot+samples+derivs, 2 calls	no answer	0	·	6	0	4	9	6	9	10	10
	never conceived	0	·	0	0	1	0	0	1	0	0
	named only	1	·	0	0	0	0	0	0	0	0
	dominated	1	·	0	7	0	1	0	0	0	0
	front lacks it	0	·	0	0	0	0	0	0	0	0
	coefficients off	0	·	1	0	1	0	1	0	0	0

Table 41: Qwen: failed runs per cell, by the stage at which a true term stopped: no answer; never conceived; named but never written; written, then dominated; on the front, but no member the law; the structure with coefficients off.

Configuration	Runs	KdV	K–S	Bu_v	Bu_i	VdP	A–C	Wave	Comp	Osc	Rat
samples	no answer	·	·	0	0	·	0	0	0	0	0
	never conceived	·	·	0	0	·	0	0	2	7	10
	named only	·	·	0	0	·	0	0	0	0	0
	dominated	·	·	0	10	·	3	0	0	0	0
	front lacks it	·	·	0	0	·	0	5	1	0	0
	coefficients off	·	·	3	0	·	0	2	1	0	0
plot	no answer	·	0	0	0	0	0	0	0	0	0
	never conceived	·	2	4	5	4	1	0	10	10	10
	named only	·	0	0	0	0	0	2	0	0	0
	dominated	·	0	1	0	0	5	4	0	0	0
	front lacks it	·	0	0	0	0	0	1	0	0	0
	coefficients off	·	2	4	1	1	4	3	0	0	0
plot+samples	no answer	·	0	0	0	0	0	0	0	0	0
	never conceived	·	2	0	0	1	0	0	1	8	10
	named only	·	0	0	0	0	0	0	1	0	0
	dominated	·	0	0	8	0	2	2	0	0	0
	front lacks it	·	0	2	0	0	0	5	0	0	0
	coefficients off	·	0	5	0	0	3	1	1	0	0
samples, 2 calls	no answer	·	·	0	0	·	0	0	0	0	0
	never conceived	·	·	0	0	·	0	0	3	10	9
	named only	·	·	0	0	·	0	0	0	0	0
	dominated	·	·	0	10	·	1	1	0	0	0
	front lacks it	·	·	1	0	·	0	7	0	0	0
	coefficients off	·	·	7	0	·	6	1	0	0	0
plot, 2 calls	no answer	0	0	0	0	0	0	0	0	0	0
	never conceived	2	4	5	1	3	1	0	10	10	10
	named only	0	0	0	1	0	0	2	0	0	0
	dominated	0	1	0	0	0	0	1	0	0	0
	front lacks it	0	1	0	1	0	0	5	0	0	0
	coefficients off	0	1	5	3	0	9	2	0	0	0
plot+samples, 2 calls	no answer	·	0	0	0	0	0	0	0	0	0
	never conceived	·	1	1	0	1	0	0	3	10	10
	named only	·	1	0	0	0	0	0	1	0	0
	dominated	·	0	0	8	0	1	1	0	0	0
	front lacks it	·	0	1	0	0	0	6	0	0	0
	coefficients off	·	1	4	0	0	4	1	0	0	0

Table 41 (continued).

Configuration	Runs	KdV	K–S	Bu_v	Bu_i	VdP	A–C	Wave	Comp	Osc	Rat
samples+derivs	no answer	·	·	0	0	·	0	·	0	0	0
	never conceived	·	·	0	0	·	0	·	7	7	10
	named only	·	·	0	0	·	0	·	0	0	0
	dominated	·	·	1	8	·	7	·	1	0	0
	front lacks it	·	·	0	0	·	0	·	0	0	0
	coefficients off	·	·	4	0	·	0	·	0	0	0
plot+derivs	no answer	0	0	0	0	0	0	0	0	0	0
	never conceived	0	0	1	2	1	2	0	9	10	10
	named only	1	1	1	0	0	0	0	0	0	0
	dominated	0	0	4	0	1	4	0	1	0	0
	front lacks it	0	0	0	0	0	0	0	0	0	0
	coefficients off	1	0	4	4	4	4	5	0	0	0
plot+samples+derivs	no answer	0	0	0	0	0	0	0	0	0	0
	never conceived	1	1	2	0	0	0	0	5	10	10
	named only	0	1	0	0	0	0	0	0	0	0
	dominated	0	0	0	8	0	8	0	0	0	0
	front lacks it	1	0	3	0	0	0	1	0	0	0
	coefficients off	1	1	3	0	2	1	0	1	0	0
samples+derivs, 2 calls	no answer	·	0	0	0	0	0	0	0	1	0
	never conceived	·	1	0	0	0	0	0	4	9	10
	named only	·	0	0	0	0	0	1	1	0	0
	dominated	·	0	0	8	1	2	0	1	0	0
	front lacks it	·	0	2	0	0	0	1	0	0	0
	coefficients off	·	0	0	0	1	5	0	0	0	0
plot+derivs, 2 calls	no answer	0	0	0	0	0	0	0	0	0	0
	never conceived	0	1	0	1	2	2	0	8	10	10
	named only	0	1	3	0	1	0	0	1	0	0
	dominated	0	0	3	1	0	0	0	0	0	0
	front lacks it	0	0	0	0	0	0	6	0	0	0
	coefficients off	1	4	2	5	0	8	2	1	0	0
plot+samples+derivs, 2 calls	no answer	0	0	0	0	0	0	1	0	0	0
	never conceived	0	1	0	0	1	1	0	4	10	9
	named only	1	0	0	0	0	0	0	2	0	0
	dominated	0	0	3	9	1	3	0	1	0	0
	front lacks it	1	0	3	0	0	0	1	0	0	0
	coefficients off	0	0	2	0	1	6	0	2	0	0

Table 42: DeepSeek: failed runs per cell, by the stage at which a true term stopped: no answer; never conceived; named but never written; written, then dominated; on the front, but no member the law; the structure with coefficients off.

Configuration	Runs	KdV	K–S	Bu_v	Bu_i	VdP	A–C	Wave	Comp	Osc	Rat
samples	no answer	0	1	0	0	0	0	0	0	0	0
	never conceived	0	3	0	0	1	0	0	8	10	10
	named only	0	0	0	0	0	0	0	0	0	0
	dominated	0	0	1	10	0	4	3	1	0	0
	front lacks it	0	0	0	0	0	0	6	1	0	0
	coefficients off	1	3	6	0	2	3	0	0	0	0
plot	no answer	3	1	0	0	0	0	0	0	0	0
	never conceived	1	0	5	10	4	5	0	10	10	10
	named only	0	0	0	0	0	0	0	0	0	0
	dominated	4	0	0	0	0	2	2	0	0	0
	front lacks it	0	0	0	0	1	0	8	0	0	0
	coefficients off	0	2	5	0	4	3	0	0	0	0
plot+samples	no answer	0	0	0	2	·	0	0	1	0	0
	never conceived	1	0	1	0	·	0	0	4	10	10
	named only	0	0	0	0	·	0	0	1	0	0
	dominated	1	0	0	6	·	2	2	0	0	0
	front lacks it	0	1	0	0	·	0	3	0	0	0
	coefficients off	0	0	4	0	·	6	2	0	0	0
samples, 2 calls	no answer	0	0	0	0	0	0	1	0	0	0
	never conceived	0	3	2	0	1	0	0	8	10	10
	named only	0	1	0	0	0	0	1	0	0	0
	dominated	2	0	0	3	0	1	0	0	0	0
	front lacks it	0	0	0	0	0	0	6	1	0	0
	coefficients off	5	1	7	0	1	4	1	0	0	0
plot, 2 calls	no answer	0	0	0	0	0	3	0	0	0	0
	never conceived	2	0	2	7	8	1	1	10	10	10
	named only	0	0	0	0	0	1	4	0	0	0
	dominated	0	1	1	1	0	1	4	0	0	0
	front lacks it	0	0	0	0	0	1	1	0	0	0
	coefficients off	1	5	7	1	2	3	0	0	0	0
plot+samples, 2 calls	no answer	0	0	0	0	0	0	0	0	0	0
	never conceived	0	0	0	2	3	1	0	7	10	10
	named only	0	0	1	0	0	0	0	0	0	0
	dominated	0	0	0	1	0	2	0	0	0	0
	front lacks it	1	0	0	0	0	0	7	0	0	0
	coefficients off	4	3	9	2	0	5	0	2	0	0

Table 42 (continued).

Configuration	Runs	KdV	K–S	Bu_v	Bu_i	VdP	A–C	Wave	Comp	Osc	Rat
samples+derivs	no answer	0	0	0	0	0	0	·	1	0	0
	never conceived	1	0	1	0	2	3	·	6	10	10
	named only	0	1	0	0	0	0	·	1	0	0
	dominated	0	0	0	7	0	1	·	0	0	0
	front lacks it	0	0	1	0	0	0	·	1	0	0
	coefficients off	0	0	5	0	2	2	·	0	0	0
plot+derivs	no answer	0	0	0	0	0	1	0	0	0	1
	never conceived	0	4	0	2	3	6	1	10	10	9
	named only	0	0	0	0	0	0	1	0	0	0
	dominated	2	1	3	4	0	1	3	0	0	0
	front lacks it	3	2	0	0	0	0	1	0	0	0
	coefficients off	2	1	5	4	6	2	3	0	0	0
plot+samples+derivs	no answer	0	0	0	1	0	0	·	0	0	0
	never conceived	1	1	1	0	0	2	·	9	10	10
	named only	0	3	0	0	0	2	·	0	0	0
	dominated	3	1	1	4	0	0	·	0	0	0
	front lacks it	0	0	0	0	0	1	·	0	0	0
	coefficients off	0	0	4	0	3	5	·	1	0	0
samples+derivs, 2 calls	no answer	0	0	0	0	0	1	1	0	0	0
	never conceived	1	0	0	1	1	0	0	9	9	10
	named only	3	1	1	0	0	1	0	1	0	0
	dominated	1	0	1	2	0	2	0	0	0	0
	front lacks it	0	0	3	0	0	0	1	0	0	0
	coefficients off	1	1	2	1	2	4	0	0	1	0
plot+derivs, 2 calls	no answer	0	0	0	0	0	0	0	0	0	0
	never conceived	1	3	0	1	5	7	0	10	10	10
	named only	1	1	2	0	0	0	2	0	0	0
	dominated	2	2	2	2	1	0	0	0	0	0
	front lacks it	0	0	1	3	0	0	1	0	0	0
	coefficients off	3	1	4	3	2	3	1	0	0	0
plot+samples+derivs, 2 calls	no answer	0	0	0	0	2	2	1	1	1	0
	never conceived	3	0	2	0	4	3	0	9	9	10
	named only	3	4	1	0	0	0	1	0	0	0
	dominated	0	0	1	8	1	0	0	0	0	0
	front lacks it	0	0	1	0	0	0	0	0	0	0
	coefficients off	0	1	4	0	2	5	1	0	0	0

Table 43: Gemma: failed runs per cell, by the stage at which a true term stopped: no answer; never conceived; named but never written; written, then dominated; on the front, but no member the law; the structure with coefficients off.

Configuration	Runs	KdV	K–S	Bu_v	Bu_i	VdP	A–C	Wave	Comp	Osc	Rat
samples	no answer	0	0	0	0	0	0	0	0	0	0
	never conceived	0	9	7	7	8	1	0	10	10	10
	named only	0	1	2	0	0	0	3	0	0	0
	dominated	0	0	0	0	0	7	1	0	0	0
	front lacks it	0	0	1	0	0	0	0	0	0	0
	coefficients off	9	0	0	2	2	2	6	0	0	0
plot	no answer	0	0	0	0	0	0	0	0	0	0
	never conceived	2	3	2	10	9	0	0	7	10	10
	named only	1	0	1	0	1	0	0	0	0	0
	dominated	3	0	0	0	0	4	1	0	0	0
	front lacks it	0	0	0	0	0	0	4	0	0	0
	coefficients off	4	5	7	0	0	6	5	3	0	0
plot+samples	no answer	0	0	0	0	0	0	0	0	0	0
	never conceived	1	5	4	10	10	0	0	10	10	10
	named only	1	1	2	0	0	0	6	0	0	0
	dominated	2	0	1	0	0	8	0	0	0	0
	front lacks it	0	1	0	0	0	0	0	0	0	0
	coefficients off	6	2	3	0	0	2	4	0	0	0
samples, 2 calls	no answer	0	0	0	0	0	0	0	0	0	0
	never conceived	0	8	0	4	6	3	0	9	10	10
	named only	0	2	10	6	0	0	4	1	0	0
	dominated	0	0	0	0	1	0	2	0	0	0
	front lacks it	0	0	0	0	0	0	1	0	0	0
	coefficients off	10	0	0	0	3	7	3	0	0	0
plot, 2 calls	no answer	0	0	0	0	0	0	0	0	0	0
	never conceived	0	3	0	4	6	0	0	9	10	10
	named only	2	1	3	4	0	0	9	1	0	0
	dominated	0	1	0	0	3	0	0	0	0	0
	front lacks it	0	0	0	2	0	0	1	0	0	0
	coefficients off	7	2	7	0	1	10	0	0	0	0
plot+samples, 2 calls	no answer	0	0	0	0	0	0	0	0	0	0
	never conceived	0	7	1	8	5	0	0	9	10	10
	named only	2	0	5	1	3	0	3	0	0	0
	dominated	0	2	0	0	0	1	3	0	0	0
	front lacks it	0	0	0	0	0	0	1	0	0	0
	coefficients off	8	1	4	1	2	9	3	1	0	0

Table 43 (continued).

Configuration	Runs	KdV	K–S	Bu_v	Bu_i	VdP	A–C	Wave	Comp	Osc	Rat
samples+derivs	no answer	0	0	0	0	0	0	0	0	0	0
	never conceived	1	7	2	9	9	6	0	9	10	10
	named only	0	3	4	0	1	4	0	0	0	0
	dominated	0	0	2	0	0	0	0	0	0	0
	front lacks it	0	0	1	0	0	0	1	0	0	0
	coefficients off	9	0	0	0	0	0	1	0	0	0
plot+derivs	no answer	0	0	0	0	0	0	0	0	0	0
	never conceived	0	0	0	4	10	10	0	10	10	10
	named only	0	1	1	2	0	0	3	0	0	0
	dominated	3	0	4	1	0	0	0	0	0	0
	front lacks it	0	1	1	0	0	0	0	0	0	0
	coefficients off	6	6	1	1	0	0	6	0	0	0
plot+samples+derivs	no answer	0	0	0	0	0	0	0	0	0	0
	never conceived	0	0	1	8	5	10	0	7	10	10
	named only	0	1	0	0	1	0	0	3	0	0
	dominated	0	0	1	0	0	0	0	0	0	0
	front lacks it	0	0	2	0	0	0	1	0	0	0
	coefficients off	9	7	4	2	4	0	1	0	0	0
samples+derivs, 2 calls	no answer	0	0	0	0	0	0	0	0	0	0
	never conceived	0	7	0	6	5	8	0	10	10	10
	named only	0	2	4	2	2	1	1	0	0	0
	dominated	0	0	1	0	0	0	0	0	0	0
	front lacks it	0	0	1	0	0	0	2	0	0	0
	coefficients off	10	0	4	0	3	1	0	0	0	0
plot+derivs, 2 calls	no answer	0	0	0	0	0	0	0	0	0	0
	never conceived	0	0	0	3	8	10	0	8	10	10
	named only	0	0	2	2	1	0	4	1	0	0
	dominated	0	3	0	2	0	0	2	0	0	0
	front lacks it	0	0	2	0	0	0	2	0	0	0
	coefficients off	9	4	6	2	1	0	2	1	0	0
plot+samples+derivs, 2 calls	no answer	0	0	0	0	0	0	0	0	0	0
	never conceived	0	0	1	6	5	10	0	10	10	10
	named only	0	0	3	1	2	0	1	0	0	0
	dominated	0	0	0	0	1	0	0	0	0	0
	front lacks it	0	0	0	0	1	0	4	0	0	0
	coefficients off	9	7	6	2	1	0	0	0	0	0

Table 44: Prompt arms and the data-withheld arm: runs per cell never run, without an answer, missing the structure, and with the structure but coefficients off.

Arm	Runs	KdV	K–S	Bu_v	Bu_i	VdP	A–C	Wave	Comp	Osc	Rat
GLM (control: samples)											
control	never run	0	0	0	0	0	0	0	0	0	0
	no answer	0	0	2	0	0	0	4	2	3	6
	structure missed	0	0	1	7	0	0	1	2	2	2
	coefficients off	0	0	0	0	0	0	1	0	0	0
grammar to operands	never run	0	0	0	0	0	0	0	0	0	0
	no answer	1	2	2	0	2	2	8	7	10	7
	structure missed	0	0	0	1	0	0	0	0	0	3
	coefficients off	0	0	1	0	0	0	1	0	0	0
grammar deleted	never run	0	0	0	0	0	0	0	0	0	0
	no answer	1	2	5	0	0	0	7	3	4	8
	structure missed	0	2	1	1	3	1	0	0	6	2
	coefficients off	0	1	0	0	2	0	0	0	0	0
example abstract	never run	0	0	0	0	0	0	0	1	0	0
	no answer	0	8	7	0	0	2	7	6	4	7
	structure missed	0	0	0	7	0	0	0	0	2	2
	coefficients off	0	0	1	0	0	0	1	0	0	0
example poisoned	never run	0	0	0	0	0	0	0	0	0	0
	no answer	1	1	7	0	1	2	9	6	8	9
	structure missed	0	0	0	5	0	0	0	1	0	1
	coefficients off	0	0	0	0	0	0	0	0	0	0
example is answer	never run	0	0	0	0	0	0	0	0	0	0
	no answer	0	2	7	0	0	4	8	8	0	10
	structure missed	0	0	0	3	0	0	0	1	8	0
	coefficients off	0	0	0	0	7	0	0	0	0	0
data withheld	never run	0	0	0	0	0	0	0	0	0	0
	no answer	2	0	4	0	0	1	0	1	0	2
	structure missed	0	1	2	10	3	3	1	9	10	7
	coefficients off	0	0	2	0	7	3	9	0	0	1

Table 44 (continued).

Arm	Runs	KdV	K–S	Bu_v	Bu_i	VdP	A–C	Wave	Comp	Osc	Rat
Qwen (control: samples)											
control	never run	0	0	0	0	0	0	0	0	0	0
	no answer	0	0	0	0	0	0	0	0	0	0
	structure missed	0	0	0	10	0	3	5	3	7	10
	coefficients off	0	0	3	0	0	0	2	1	0	0
grammar to operands	never run	0	0	0	0	0	0	0	0	0	0
	no answer	0	0	0	0	0	0	0	0	0	0
	structure missed	0	0	0	9	0	3	4	2	9	10
	coefficients off	0	0	4	0	0	0	3	0	0	0
grammar deleted	never run	0	0	0	0	0	0	0	0	0	0
	no answer	0	0	0	0	0	0	0	0	0	0
	structure missed	0	0	2	3	0	4	5	2	8	10
	coefficients off	0	0	4	0	0	0	2	2	0	0
example abstract	never run	0	0	0	0	0	0	0	0	0	0
	no answer	0	0	0	0	0	0	0	0	0	0
	structure missed	0	1	0	9	1	3	3	3	7	9
	coefficients off	0	0	4	0	0	1	2	1	0	1
example poisoned	never run	0	0	0	0	0	0	0	0	0	0
	no answer	0	0	0	0	0	0	0	0	0	1
	structure missed	0	1	0	8	0	1	3	2	6	7
	coefficients off	0	0	4	0	0	1	3	0	0	0
example is answer	never run	0	0	0	0	0	0	0	0	0	0
	no answer	0	0	0	0	0	0	0	0	0	0
	structure missed	1	0	0	9	0	1	5	3	7	6
	coefficients off	1	0	7	0	0	0	3	0	0	1
data withheld	never run	0	0	0	0	0	0	0	0	0	0
	no answer	0	0	0	0	0	0	0	0	0	0
	structure missed	0	4	2	10	8	9	5	10	10	7
	coefficients off	3	0	7	0	2	1	5	0	0	3

Table 44 (continued).

Arm	Runs	KdV	K–S	Bu_v	Bu_i	VdP	A–C	Wave	Comp	Osc	Rat
DeepSeek (control: samples+derivs)											
control	never run	0	0	0	0	0	0	0	0	0	0
	no answer	0	0	0	0	0	0	0	1	0	0
	structure missed	1	1	2	7	2	4	0	8	10	10
	coefficients off	0	0	5	0	2	2	0	0	0	0
grammar to operands	never run	0	0	0	0	0	0	0	0	0	0
	no answer	1	6	3	3	6	1	5	0	7	0
	structure missed	0	0	4	3	0	6	1	9	3	10
	coefficients off	1	0	3	0	0	3	0	0	0	0
grammar deleted	never run	0	0	0	0	0	0	0	0	0	0
	no answer	3	0	1	1	0	3	1	0	1	0
	structure missed	1	0	1	0	0	3	1	8	9	10
	coefficients off	0	0	2	1	3	3	0	0	0	0
example abstract	never run	0	0	0	0	0	0	0	0	0	0
	no answer	5	1	3	0	7	1	0	1	5	0
	structure missed	2	0	1	8	2	6	0	8	5	10
	coefficients off	1	1	2	0	1	3	0	0	0	0
example poisoned	never run	0	0	0	0	0	0	0	0	0	0
	no answer	1	0	2	9	0	1	8	6	2	1
	structure missed	0	0	0	1	4	7	1	4	8	9
	coefficients off	1	0	6	0	0	1	0	0	0	0
example is answer	never run	0	0	0	0	0	0	0	0	0	0
	no answer	0	0	0	2	0	4	1	4	1	3
	structure missed	1	0	0	5	5	0	2	6	9	7
	coefficients off	0	0	1	0	1	3	0	0	0	0
data withheld	never run	0	0	0	0	0	0	0	0	0	0
	no answer	3	0	0	0	3	0	0	0	3	0
	structure missed	1	0	0	10	7	3	0	10	7	10
	coefficients off	6	0	1	0	0	3	10	0	0	0

Table 44 (continued).

Arm	Runs	KdV	K–S	Bu_v	Bu_i	VdP	A–C	Wave	Comp	Osc	Rat
Gemma (control: plot+samples+derivs)											
control	never run	0	0	0	0	0	0	0	0	0	0
	no answer	0	0	0	0	0	0	0	0	0	0
	structure missed	0	1	4	8	6	10	1	10	10	10
	coefficients off	9	7	4	2	4	0	1	0	0	0
grammar to operands	never run	3	0	0	0	0	0	0	0	0	0
	no answer	0	0	0	0	0	0	0	0	0	0
	structure missed	1	7	8	10	10	10	2	10	10	10
	coefficients off	5	1	2	0	0	0	3	0	0	0
grammar deleted	never run	0	0	0	0	0	0	0	0	0	0
	no answer	0	0	0	0	0	0	0	0	0	0
	structure missed	0	5	3	10	10	10	1	10	10	10
	coefficients off	9	3	5	0	0	0	8	0	0	0
example abstract	never run	3	0	0	0	0	0	0	0	0	0
	no answer	0	0	0	0	0	0	0	0	0	0
	structure missed	1	6	7	10	9	10	1	10	10	10
	coefficients off	6	1	3	0	1	0	5	0	0	0
example poisoned	never run	3	0	0	0	0	0	0	0	0	0
	no answer	0	0	0	0	0	0	0	0	0	0
	structure missed	0	3	8	9	10	10	3	10	10	10
	coefficients off	5	4	2	0	0	0	2	0	0	0
example is answer	never run	4	0	0	0	0	0	0	0	0	0
	no answer	0	0	0	0	0	0	0	0	0	0
	structure missed	0	3	5	8	3	7	0	9	10	10
	coefficients off	1	1	2	2	4	3	1	1	0	0
data withheld	never run	0	0	0	0	0	0	0	0	0	0
	no answer	0	0	0	0	0	0	0	0	0	0
	structure missed	0	10	1	6	9	10	0	10	10	10
	coefficients off	3	0	7	0	1	0	10	0	0	0

E.6 SELECTION

Tables 45 to 48 count, per cell, the discoveries of the screened set, of any of the five candidates, of the candidate with the lowest Discrepancy alone and of the one with the lowest Instability alone. Outside Inviscid Burgers the screened set keeps every discovery that one of the five candidates holds for GLM, Qwen and Gemma, and loses one for DeepSeek, on Allen–Cahn. For GLM, Qwen and DeepSeek, Instability alone loses most of the Allen–Cahn discoveries of the screened set, which Discrepancy alone mostly keeps; Discrepancy alone loses a few discoveries for GLM, Qwen and DeepSeek and none for Gemma, on no system more than twice in a cell. The screened set is small (Table 49): in the strongest cells it holds at most four candidates for every model. In the strongest cells every Allen–Cahn discovery of GLM, Qwen and DeepSeek shares its screened set with a wrong structure, the reaction-only form $u - u^3$ that Instability alone prefers; on the other systems this happens in at most three runs of any model.

Table 45: GLM: equations discovered by the screened set, by any of the five candidates, by the candidate with the lowest Discrepancy and by the one with the lowest Instability; · marks none.

Configuration	Discovered by	KdV	K–S	Bu_v	Bu_i	VdP	A–C	Wave	Comp	Osc	Rat
samples	screened set	10	10	7	3	10	10	4	6	5	2
	any candidate	10	10	7	10	10	10	4	6	5	2
	lowest Discrepancy	10	10	7	3	10	10	4	5	5	2
	lowest Instability	10	10	7	3	10	0	4	6	5	2
plot	screened set	8	10	2	4	9	7	1	·	·	·
	any candidate	8	10	2	6	9	7	1	·	·	·
	lowest Discrepancy	8	10	2	4	9	7	1	·	·	·
	lowest Instability	8	10	2	4	9	0	1	·	·	·
plot+samples	screened set	10	10	8	2	10	9	3	6	·	1
	any candidate	10	10	8	10	10	9	3	6	·	1
	lowest Discrepancy	10	10	8	2	10	9	3	6	·	1
	lowest Instability	10	10	8	1	10	0	3	6	·	1
samples, 2 calls	screened set	9	9	5	3	9	7	3	3	4	1
	any candidate	9	9	5	10	9	7	3	3	4	1
	lowest Discrepancy	9	9	5	2	9	7	3	3	4	1
	lowest Instability	9	9	5	3	9	1	3	3	4	1
plot, 2 calls	screened set	9	10	·	7	10	7	·	·	·	·
	any candidate	9	10	·	7	10	7	·	·	·	·
	lowest Discrepancy	9	10	·	7	10	7	·	·	·	·
	lowest Instability	6	10	·	7	10	4	·	·	·	·
plot+samples, 2 calls	screened set	10	9	6	5	7	7	1	4	·	1
	any candidate	10	9	6	10	7	7	1	4	·	1
	lowest Discrepancy	10	9	6	4	7	7	1	4	·	1
	lowest Instability	10	9	5	5	7	2	1	3	·	1
samples+derivs	screened set	10	10	8	3	9	9	8	5	·	1
	any candidate	10	10	8	10	9	9	8	5	·	1
	lowest Discrepancy	10	10	8	3	9	9	8	5	·	1
	lowest Instability	10	10	8	3	9	1	8	5	·	1
plot+derivs	screened set	8	10	2	4	10	1	4	2	·	·
	any candidate	8	10	2	7	10	1	4	2	·	·
	lowest Discrepancy	8	10	2	4	10	1	4	2	·	·
	lowest Instability	8	10	2	4	10	0	4	2	·	·
plot+samples+derivs	screened set	10	10	4	2	8	4	10	7	·	·
	any candidate	10	10	4	10	8	4	10	7	·	·
	lowest Discrepancy	10	10	4	2	8	4	10	7	·	·
	lowest Instability	10	10	4	2	8	0	10	7	·	·
samples+derivs, 2 calls	screened set	9	9	5	3	7	5	9	1	1	·
	any candidate	9	9	5	10	7	5	9	1	1	·
	lowest Discrepancy	9	9	5	3	7	5	9	1	1	·
	lowest Instability	9	9	5	3	7	2	9	1	1	·
plot+derivs, 2 calls	screened set	8	10	1	5	9	·	4	·	·	·
	any candidate	8	10	1	5	9	·	4	·	·	·
	lowest Discrepancy	8	10	1	3	9	·	4	·	·	·
	lowest Instability	7	10	1	5	9	·	4	·	·	·
plot+samples+derivs, 2 calls	screened set	8	10	3	3	4	·	3	·	·	·
	any candidate	8	10	3	10	4	·	3	·	·	·
	lowest Discrepancy	8	10	3	2	4	·	3	·	·	·
	lowest Instability	8	10	3	3	4	·	3	·	·	·

Table 46: Qwen: equations discovered by the screened set, by any of the five candidates, by the candidate with the lowest Discrepancy and by the one with the lowest Instability; · marks none.

Configuration	Discovered by	KdV	K–S	Bu_v	Bu_i	VdP	A–C	Wave	Comp	Osc	Rat
samples	screened set	10	10	7	0	10	7	3	6	3	·
	any candidate	10	10	7	10	10	7	3	6	3	·
	lowest Discrepancy	10	10	7	0	10	5	3	6	3	·
	lowest Instability	10	10	7	0	10	0	3	6	3	·
plot	screened set	10	6	1	4	5	·	·	·	·	·
	any candidate	10	6	1	4	5	·	·	·	·	·
	lowest Discrepancy	10	6	1	3	5	·	·	·	·	·
	lowest Instability	10	6	1	4	5	·	·	·	·	·
plot+samples	screened set	10	8	3	2	9	5	2	7	2	·
	any candidate	10	8	3	10	9	5	2	7	2	·
	lowest Discrepancy	9	8	3	2	9	5	2	7	2	·
	lowest Instability	9	8	3	2	9	0	2	7	2	·
samples, 2 calls	screened set	10	10	2	0	10	3	1	7	·	1
	any candidate	10	10	2	10	10	3	1	7	·	1
	lowest Discrepancy	10	10	2	0	10	3	0	7	·	1
	lowest Instability	9	10	2	0	10	2	1	6	·	1
plot, 2 calls	screened set	8	3	·	4	7	·	·	·	·	·
	any candidate	8	3	·	4	7	·	·	·	·	·
	lowest Discrepancy	8	3	·	4	7	·	·	·	·	·
	lowest Instability	8	3	·	4	7	·	·	·	·	·
plot+samples, 2 calls	screened set	10	7	4	2	9	5	2	6	·	·
	any candidate	10	7	4	10	9	5	2	6	·	·
	lowest Discrepancy	10	7	4	2	9	5	2	6	·	·
	lowest Instability	10	7	3	2	8	3	2	6	·	·
samples+derivs	screened set	10	10	5	2	10	3	10	2	3	·
	any candidate	10	10	5	10	10	3	10	2	3	·
	lowest Discrepancy	10	10	5	2	10	3	10	2	2	·
	lowest Instability	10	10	5	2	10	0	10	2	3	·
plot+derivs	screened set	8	9	·	4	4	·	5	·	·	·
	any candidate	8	9	·	4	4	·	5	·	·	·
	lowest Discrepancy	8	9	·	4	4	·	5	·	·	·
	lowest Instability	8	9	·	4	4	·	5	·	·	·
plot+samples+derivs	screened set	7	7	2	2	8	1	9	4	·	·
	any candidate	7	7	2	10	8	1	9	4	·	·
	lowest Discrepancy	7	7	2	2	8	1	9	4	·	·
	lowest Instability	7	7	2	2	8	0	9	4	·	·
samples+derivs, 2 calls	screened set	10	9	8	2	8	3	8	4	·	·
	any candidate	10	9	8	10	8	3	8	4	·	·
	lowest Discrepancy	10	9	8	1	8	3	8	4	·	·
	lowest Instability	10	9	6	2	8	0	8	4	·	·
plot+derivs, 2 calls	screened set	9	4	2	3	7	·	2	·	·	·
	any candidate	9	4	2	3	7	·	2	·	·	·
	lowest Discrepancy	9	4	2	3	7	·	2	·	·	·
	lowest Instability	8	4	2	2	7	·	2	·	·	·
plot+samples+derivs, 2 calls	screened set	8	9	2	1	7	·	8	1	·	1
	any candidate	8	9	2	10	7	·	8	1	·	1
	lowest Discrepancy	8	9	2	1	7	·	8	1	·	1
	lowest Instability	6	9	2	1	6	·	8	1	·	1

Table 47: DeepSeek: equations discovered by the screened set, by any of the five candidates, by the candidate with the lowest Discrepancy and by the one with the lowest Instability; · marks none.

Configuration	Discovered by	KdV	K–S	Bu_v	Bu_i	VdP	A–C	Wave	Comp	Osc	Rat
samples	screened set	9	3	3	0	7	3	1	·	·	·
	any candidate	9	3	3	10	7	3	1	·	·	·
	lowest Discrepancy	9	3	3	0	6	3	1	·	·	·
	lowest Instability	9	3	3	0	7	0	1	·	·	·
plot	screened set	2	7	·	·	1	·	·	·	·	·
	any candidate	2	7	·	·	1	·	·	·	·	·
	lowest Discrepancy	2	7	·	·	1	·	·	·	·	·
	lowest Instability	2	7	·	·	1	·	·	·	·	·
plot+samples	screened set	8	9	5	2	10	2	3	4	·	·
	any candidate	8	9	5	8	10	2	3	4	·	·
	lowest Discrepancy	8	9	5	2	10	2	3	4	·	·
	lowest Instability	8	9	5	2	10	1	3	4	·	·
samples, 2 calls	screened set	3	5	1	7	8	5	1	1	·	·
	any candidate	3	5	1	10	8	6	1	1	·	·
	lowest Discrepancy	3	5	1	7	8	5	1	1	·	·
	lowest Instability	3	5	1	7	7	4	1	1	·	·
plot, 2 calls	screened set	7	4	·	1	·	·	·	·	·	·
	any candidate	7	4	·	1	·	·	·	·	·	·
	lowest Discrepancy	7	4	·	1	·	·	·	·	·	·
	lowest Instability	4	4	·	1	·	·	·	·	·	·
plot+samples, 2 calls	screened set	5	7	·	5	7	2	3	1	·	·
	any candidate	5	7	·	6	7	2	3	1	·	·
	lowest Discrepancy	5	7	·	4	7	2	3	1	·	·
	lowest Instability	5	7	·	5	6	1	3	0	·	·
samples+derivs	screened set	9	9	3	3	6	4	10	1	·	·
	any candidate	9	9	3	10	6	4	10	1	·	·
	lowest Discrepancy	7	9	3	3	6	4	10	1	·	·
	lowest Instability	8	9	3	3	6	0	10	1	·	·
plot+derivs	screened set	3	2	2	0	1	·	1	·	·	·
	any candidate	3	2	2	1	1	·	1	·	·	·
	lowest Discrepancy	3	2	2	0	1	·	1	·	·	·
	lowest Instability	3	2	2	0	1	·	1	·	·	·
plot+samples+derivs	screened set	6	5	4	5	7	·	10	·	·	·
	any candidate	6	5	4	9	7	·	10	·	·	·
	lowest Discrepancy	6	5	4	5	7	·	9	·	·	·
	lowest Instability	6	5	4	5	7	·	10	·	·	·
samples+derivs, 2 calls	screened set	4	8	3	6	7	2	8	·	·	·
	any candidate	4	8	3	8	7	2	8	·	·	·
	lowest Discrepancy	4	8	3	6	7	2	8	·	·	·
	lowest Instability	3	8	3	6	5	2	8	·	·	·
plot+derivs, 2 calls	screened set	3	3	1	1	2	·	6	·	·	·
	any candidate	3	3	1	1	2	·	6	·	·	·
	lowest Discrepancy	3	3	1	1	2	·	6	·	·	·
	lowest Instability	3	3	1	1	2	·	4	·	·	·
plot+samples+derivs, 2 calls	screened set	4	5	1	2	1	·	7	·	·	·
	any candidate	4	5	1	10	1	·	7	·	·	·
	lowest Discrepancy	4	5	1	2	1	·	7	·	·	·
	lowest Instability	3	5	1	2	1	·	7	·	·	·

Table 48: Gemma: equations discovered by the screened set, by any of the five candidates, by the candidate with the lowest Discrepancy and by the one with the lowest Instability; · marks none.

Configuration	Discovered by	KdV	K–S	Bu_v	Bu_i	VdP	A–C	Wave	Comp	Osc	Rat
samples	screened set	1	·	·	1	·	·	·	·	·	·
	any candidate	1	·	·	1	·	·	·	·	·	·
	lowest Discrepancy	1	·	·	1	·	·	·	·	·	·
	lowest Instability	1	·	·	1	·	·	·	·	·	·
plot	screened set	·	2	·	·	·	·	·	·	·	·
	any candidate	·	2	·	·	·	·	·	·	·	·
	lowest Discrepancy	·	2	·	·	·	·	·	·	·	·
	lowest Instability	·	2	·	·	·	·	·	·	·	·
plot+samples	screened set	·	1	·	·	·	·	·	·	·	·
	any candidate	·	1	·	·	·	·	·	·	·	·
	lowest Discrepancy	·	1	·	·	·	·	·	·	·	·
	lowest Instability	·	1	·	·	·	·	·	·	·	·
samples, 2 calls		·	·	·	·	·	·	·	·	·	·
plot, 2 calls	screened set	1	3	·	·	·	·	·	·	·	·
	any candidate	1	3	·	·	·	·	·	·	·	·
	lowest Discrepancy	1	3	·	·	·	·	·	·	·	·
	lowest Instability	1	3	·	·	·	·	·	·	·	·
plot+samples, 2 calls		·	·	·	·	·	·	·	·	·	·
samples+derivs	screened set	·	·	1	1	·	·	8	1	·	·
	any candidate	·	·	1	1	·	·	8	1	·	·
	lowest Discrepancy	·	·	1	1	·	·	8	1	·	·
	lowest Instability	·	·	1	1	·	·	8	1	·	·
plot+derivs	screened set	1	2	3	2	·	·	1	·	·	·
	any candidate	1	2	3	2	·	·	1	·	·	·
	lowest Discrepancy	1	2	3	2	·	·	1	·	·	·
	lowest Instability	1	2	3	2	·	·	1	·	·	·
plot+samples+derivs	screened set	1	2	2	·	·	·	8	·	·	·
	any candidate	1	2	2	·	·	·	8	·	·	·
	lowest Discrepancy	1	2	2	·	·	·	8	·	·	·
	lowest Instability	1	2	2	·	·	·	8	·	·	·
samples+derivs, 2 calls	screened set	·	1	·	2	·	·	7	·	·	·
	any candidate	·	1	·	2	·	·	7	·	·	·
	lowest Discrepancy	·	1	·	2	·	·	7	·	·	·
	lowest Instability	·	1	·	2	·	·	5	·	·	·
plot+derivs, 2 calls	screened set	1	3	·	1	·	·	·	·	·	·
	any candidate	1	3	·	1	·	·	·	·	·	·
	lowest Discrepancy	1	3	·	1	·	·	·	·	·	·
	lowest Instability	1	3	·	1	·	·	·	·	·	·
plot+samples+derivs, 2 calls	screened set	1	3	·	1	·	·	5	·	·	·
	any candidate	1	3	·	1	·	·	5	·	·	·
	lowest Discrepancy	1	3	·	1	·	·	5	·	·	·
	lowest Instability	1	3	·	1	·	·	5	·	·	·

Table 49: Screened set per model and system: mean (maximum) number of candidates over the runs that answered; runs with a discovery whose screened set also holds a wrong structure, and runs with a discovery; discoveries whose first candidate is itself a discovery on the screened set, on or off it, and all discoveries; · marks none.

Model			KdV	K–S	Bu_v	Bu_i	VdP	A–C	Wave	Comp	Osc	Rat
GLM	size, strongest		1.0	1.0	1.2	1.1	1.0	2.0	1.3	1.6	1.3	1.5
			(1)	(1)	(2)	(2)	(1)	(2)	(3)	(4)	(3)	(2)
	size, main sweep		1.1	1.0	1.3	1.2	1.1	1.7	1.4	1.4	1.9	1.8
			(3)	(4)	(3)	(3)	(3)	(4)	(4)	(4)	(4)	(5)
	wrong structure too, strongest	with a wrong one	0	0	0	0	0	10	0	1	0	0
		with a discovery	10	10	7	3	10	10	4	6	5	2
	wrong structure too, main sweep	with a wrong one	4	0	1	6	0	56	0	2	0	0
		with a discovery	109	117	51	44	102	66	50	34	10	6
	first candidate, strongest	on screened set	10	10	7	3	8	10	4	6	5	2
		on or off it	10	10	7	3	10	10	4	6	5	2
		discoveries	10	10	7	3	10	10	4	6	5	2
Qwen	size, strongest		1.3	1.0	1.1	1.4	1.0	1.9	1.8	1.4	2.1	1.8
			(2)	(1)	(2)	(2)	(1)	(3)	(3)	(3)	(4)	(3)
	size, main sweep		1.2	1.2	1.7	1.2	1.2	1.5	1.6	1.6	2.1	1.6
			(4)	(3)	(4)	(3)	(3)	(3)	(4)	(4)	(4)	(4)
	wrong structure too, strongest	with a wrong one	0	0	0	·	0	7	0	0	0	·
		with a discovery	10	10	7	·	10	7	3	6	3	·
	wrong structure too, main sweep	with a wrong one	6	0	3	2	2	22	1	1	1	0
		with a discovery	110	92	36	26	94	27	50	37	8	2
	first candidate, strongest	on screened set	10	8	6	·	10	7	0	3	3	·
		on or off it	10	8	6	·	10	7	0	6	3	·
		discoveries	10	10	7	·	10	7	3	6	3	·
DeepSeek	size, strongest		1.3	1.1	1.4	1.1	1.0	1.7	1.0	2.4	2.0	1.7
			(2)	(2)	(3)	(2)	(1)	(3)	(1)	(4)	(3)	(3)
	size, main sweep		1.4	1.5	1.5	1.3	1.5	1.6	1.4	1.9	1.9	1.7
			(4)	(4)	(4)	(3)	(5)	(4)	(4)	(4)	(4)	(4)
	wrong structure too, strongest	with a wrong one	3	0	0	0	0	4	0	0	·	·
		with a discovery	9	9	3	3	6	4	10	1	·	·
	wrong structure too, main sweep	with a wrong one	8	0	0	1	5	10	3	1	·	·
		with a discovery	63	67	23	32	57	18	50	7	·	·
	first candidate, strongest	on screened set	9	8	3	2	6	4	7	1	·	·
		on or off it	9	8	3	3	6	4	7	1	·	·
		discoveries	9	9	3	3	6	4	10	1	·	·
Gemma	size, strongest		2.0	1.8	1.3	1.3	1.6	2.2	1.3	1.4	1.8	1.4
			(3)	(3)	(3)	(3)	(3)	(4)	(3)	(3)	(3)	(2)
	size, main sweep		1.9	2.0	1.7	1.6	1.8	2.1	1.6	1.8	2.0	1.7
			(4)	(4)	(3)	(4)	(4)	(5)	(4)	(4)	(4)	(4)
	wrong structure too, strongest	with a wrong one	0	0	0	·	·	·	0	·	·	·
		with a discovery	1	2	2	·	·	·	8	·	·	·
	wrong structure too, main sweep	with a wrong one	0	0	0	0	·	·	2	0	·	·
		with a discovery	6	17	6	8	·	·	29	1	·	·
	first candidate, strongest	on screened set	1	2	0	·	·	·	4	·	·	·
		on or off it	1	2	0	·	·	·	4	·	·	·
		discoveries	1	2	2	·	·	·	8	·	·	·

E.7 Discovery or recall

Table 50: Structures recovered and equations discovered, out of 10 planned runs, in each model's strongest configuration and with the data withheld. The strongest configuration shows GLM and Qwen the numerical samples, DeepSeek the samples with derivative fields, and Gemma the plot with the samples and derivative fields. Inviscid Burgers is scored on its true form (Appendix D.1); bold marks the most discoveries on a system.

	GLM		Qwen		DeepSeek		Gemma	
System	Struct.	Disc.	Struct.	Disc.	Struct.	Disc.	Struct.	Disc.
With the data (strongest configuration)								
Korteweg–de Vries	10	**10**	10	**10**	9	9	10	1
Kuramoto–Sivashinsky	10	**10**	10	**10**	9	9	9	2
Viscous Burgers	7	**7**	10	**7**	8	3	6	2
Inviscid Burgers	3	**3**	0	0	3	**3**	2	0
Van der Pol	10	**10**	10	**10**	8	6	4	0
Allen–Cahn	10	**10**	7	7	6	4	0	0
Wave	5	4	5	3	10	**10**	9	8
Compound PDE	6	**6**	7	**6**	1	1	0	0
Forced damped osc.	5	**5**	3	3	0	0	0	0
Rational PDE	2	**2**	0	0	0	0	0	0
Data withheld								
Korteweg–de Vries	8	**8**	10	7	6	0	10	7
Kuramoto–Sivashinsky	9	9	6	6	10	**10**	0	0
Viscous Burgers	4	2	8	1	10	**9**	9	2
Inviscid Burgers	0	0	0	0	0	0	4	**4**
Van der Pol	7	0	2	0	0	0	1	0
Allen–Cahn	6	3	1	0	7	**4**	0	0
Wave	9	0	5	0	10	0	10	0
Compound PDE	0	0	0	0	0	0	0	0
Forced damped osc.	0	0	0	0	0	0	0	0
Rational PDE	1	0	3	0	0	0	0	0

Without the data. Recall is a property of a model on a system (Table 50). The structure-level test is sharper than the law-level one (§4.3).

With the data. The structural recoveries that are not discoveries mostly have a coefficient more than 10% off rather than just outside the bound, and Gemma has the most of them (Table 15). In a few cells the data displace a law the model writes without them, that is, the model discovers it in fewer runs with the data than without: on no system for GLM or Qwen and on two systems each for DeepSeek and Gemma, most clearly DeepSeek on Viscous Burgers and Gemma on Korteweg–de Vries, by six runs each.

E.8 Prompt ablations

Tables 51 to 54 give the discoveries and structural recoveries of each strongest cell under each arm. Qwen answers every planned run in all but one arm cell, and Gemma every run it was sent. For both, cutting the grammar to its operands lowers the discoveries on two systems, and deleting it lowers them on three systems for Qwen and on one for Gemma; the misleading example helps more systems than it hurts, and the example that is the answer raises Gemma's discoveries on five systems and lowers them on none. For GLM and DeepSeek every arm raises the runs with no answer on most systems (Table 44).

Terms written outside the grammar. The control grammar builds a term as a product or quotient of factors, each an operand, an integer power of one, its sine or cosine, or the constant 1, and asks for a time derivative on the left. Table 55 checks every written term against it. In the strongest cells such terms are rare and decide no discovery. For every model most appear when the grammar is deleted. There, a third of GLM's such candidates, about half of DeepSeek's, three quarters of Qwen's and

Table 51: GLM under the prompt arms, on its strongest cell (samples): structural recoveries and equations discovered.

Arm		KdV	K–S	Bu_v	Bu_i	VdP	A–C	Wave	Comp	Osc	Rat
control	structures	10	10	7	3	10	10	5	6	5	2
	discovered	10	10	7	3	10	10	4	6	5	2
grammar to operands	structures	9	8	8	9	8	8	2	3	0	0
	discovered	9	8	7	9	8	8	1	3	0	0
grammar deleted	structures	9	6	4	9	7	9	3	7	0	0
	discovered	9	5	4	9	5	9	3	7	0	0
example abstract	structures	10	2	3	3	10	8	3	3	4	1
	discovered	10	2	2	3	10	8	2	3	4	1
example poisoned	structures	9	9	3	5	9	8	1	3	2	0
	discovered	9	9	3	5	9	8	1	3	2	0
example is answer	structures	10	8	3	7	10	6	2	1	2	0
	discovered	10	8	3	7	3	6	2	1	2	0

Table 52: Qwen under the prompt arms, on its strongest cell (samples): structural recoveries and equations discovered.

Arm		KdV	K–S	Bu_v	Bu_i	VdP	A–C	Wave	Comp	Osc	Rat
control	structures	10	10	10	0	10	7	5	7	3	0
	discovered	10	10	7	0	10	7	3	6	3	0
grammar to operands	structures	10	10	10	1	10	7	6	8	1	0
	discovered	10	10	6	1	10	7	3	8	1	0
grammar deleted	structures	10	10	8	7	10	6	5	8	2	0
	discovered	10	10	4	7	10	6	3	6	2	0
example abstract	structures	10	9	10	1	9	7	7	7	3	1
	discovered	10	9	6	1	9	6	5	6	3	0
example poisoned	structures	10	9	10	2	10	9	7	8	4	2
	discovered	10	9	6	2	10	8	4	8	4	2
example is answer	structures	9	10	10	1	10	9	5	7	3	4
	discovered	8	10	3	1	10	9	2	7	3	3

none of Gemma's carry a true term in a compact form such as $(u^2)_x$ or $(1-u^2)\,u_t$, which the harness does not evaluate. The few that decide a discovery are left-hand sides without a time derivative, in runs that also write the law inside the grammar.

Table 53: DeepSeek under the prompt arms, on its strongest cell (the samples with derivative fields): structural recoveries and equations discovered.

Arm		KdV	K–S	Bu_v	Bu_i	VdP	A–C	Wave	Comp	Osc	Rat
control	structures	9	9	8	3	8	6	10	1	0	0
	discovered	9	9	3	3	6	4	10	1	0	0
grammar to operands	structures	9	4	3	4	4	3	4	1	0	0
	discovered	8	4	0	4	4	0	4	1	0	0
grammar deleted	structures	6	10	8	9	10	4	8	2	0	0
	discovered	6	10	6	8	7	1	8	2	0	0
example abstract	structures	3	9	6	2	1	3	10	1	0	0
	discovered	2	8	4	2	0	0	10	1	0	0
example poisoned	structures	9	10	8	0	6	2	1	0	0	0
	discovered	8	10	2	0	6	1	1	0	0	0
example is answer	structures	9	10	10	3	5	6	7	0	0	0
	discovered	9	10	9	3	4	3	7	0	0	0

Table 54: Gemma under the prompt arms, on its strongest cell (the plot with the samples and derivative fields): structural recoveries and equations discovered.

Arm		KdV	K–S	Bu_v	Bu_i	VdP	A–C	Wave	Comp	Osc	Rat
control	structures	10	9	6	2	4	0	9	0	0	0
	discovered	1	2	2	0	0	0	8	0	0	0
grammar to operands	structures	6	3	2	0	0	0	8	0	0	0
	discovered	1	2	0	0	0	0	5	0	0	0
grammar deleted	structures	10	5	7	0	0	0	9	0	0	0
	discovered	1	2	2	0	0	0	1	0	0	0
example abstract	structures	6	4	3	0	1	0	9	0	0	0
	discovered	0	3	0	0	0	0	4	0	0	0
example poisoned	structures	7	7	2	1	0	0	7	0	0	0
	discovered	2	3	0	1	0	0	5	0	0	0
example is answer	structures	6	7	5	2	7	3	10	1	0	0
	discovered	5	6	3	0	3	0	9	0	0	0

Table 55: Terms outside the control grammar per cell of 10 planned runs: candidates carrying one, those whose such terms are true terms, those on the screened set, and discoveries decided by such a candidate; · marks none.

Arm	Count	KdV	K–S	Bu_v	Bu_i	VdP	A–C	Wave	Comp	Osc	Rat
GLM (samples)											
control	candidates	·	·	1	·	·	·	·	·	·	·
	true terms	·	·	0	·	·	·	·	·	·	·
	screened	·	·	0	·	·	·	·	·	·	·
	decides	·	·	0	·	·	·	·	·	·	·
grammar to operands	candidates	1	1	5	·	1	·	·	·	·	2
	true terms	1	1	5	·	1	·	·	·	·	2
	screened	1	1	0	·	0	·	·	·	·	0
	decides	0	1	0	·	0	·	·	·	·	0
grammar deleted	candidates	·	·	4	5	4	2	·	2	8	·
	true terms	·	·	3	2	3	0	·	0	0	·
	screened	·	·	0	0	0	0	·	0	0	·
	decides	·	·	0	0	0	0	·	0	0	·
example abstract	candidates	·	·	·	1	·	·	·	·	·	·
	true terms	·	·	·	1	·	·	·	·	·	·
	screened	·	·	·	1	·	·	·	·	·	·
	decides	·	·	·	0	·	·	·	·	·	·
example poisoned		·	·	·	·	·	·	·	·	·	·
example is answer	candidates	·	·	·	·	·	·	·	·	6	·
	true terms	·	·	·	·	·	·	·	·	0	·
	screened	·	·	·	·	·	·	·	·	0	·
	decides	·	·	·	·	·	·	·	·	0	·
Qwen (samples)											
control		·	·	·	·	·	·	·	·	·	·
grammar to operands	candidates	4	·	1	2	·	1	·	·	·	2
	true terms	4	·	1	2	·	1	·	·	·	2
	screened	4	·	0	2	·	0	·	·	·	0
	decides	0	·	0	0	·	0	·	·	·	0
grammar deleted	candidates	1	·	7	7	1	2	·	·	1	4
	true terms	1	·	6	3	1	2	·	·	0	4
	screened	1	·	0	3	0	0	·	·	0	0
	decides	0	·	0	0	0	0	·	·	0	0
example abstract	candidates	·	·	·	·	·	·	1	·	·	·
	true terms	·	·	·	·	·	·	0	·	·	·
	screened	·	·	·	·	·	·	0	·	·	·
	decides	·	·	·	·	·	·	0	·	·	·
example poisoned		·	·	·	·	·	·	·	·	·	·
example is answer		·	·	·	·	·	·	·	·	·	·

Table 55 (continued).

Arm	Count	KdV	K–S	Bu_v	Bu_i	VdP	A–C	Wave	Comp	Osc	Rat
DeepSeek (samples+derivs)											
control		·	·	·	·	·	·	·	·	·	·
grammar to operands		·	·	·	·	·	·	·	·	·	·
grammar deleted	candidates	·	2	1	2	·	5	1	6	2	·
	true terms	·	2	1	1	·	4	0	1	0	·
	screened	·	2	0	0	·	0	0	0	0	·
	decides	·	2	0	0	·	0	0	0	0	·
example abstract		·	·	·	·	·	·	·	·	·	·
example poisoned		·	·	·	·	·	·	·	·	·	·
example is answer		·	·	·	·	·	·	·	·	·	·
Gemma (plot+samples+derivs)											
control	candidates	·	·	·	1	·	·	·	·	·	1
	true terms	·	·	·	1	·	·	·	·	·	0
	screened	·	·	·	1	·	·	·	·	·	0
	decides	·	·	·	0	·	·	·	·	·	0
grammar to operands	candidates	·	·	·	·	·	·	·	·	5	3
	true terms	·	·	·	·	·	·	·	·	0	0
	screened	·	·	·	·	·	·	·	·	3	0
	decides	·	·	·	·	·	·	·	·	0	0
grammar deleted	candidates	·	·	·	·	1	·	·	·	24	·
	true terms	·	·	·	·	0	·	·	·	0	·
	screened	·	·	·	·	0	·	·	·	10	·
	decides	·	·	·	·	0	·	·	·	0	·
example abstract	candidates	·	2	1	·	·	·	·	·	·	·
	true terms	·	2	1	·	·	·	·	·	·	·
	screened	·	0	0	·	·	·	·	·	·	·
	decides	·	0	0	·	·	·	·	·	·	·
example poisoned	candidates	·	3	·	·	·	·	·	2	·	·
	true terms	·	3	·	·	·	·	·	0	·	·
	screened	·	0	·	·	·	·	·	0	·	·
	decides	·	0	·	·	·	·	·	0	·	·
example is answer	candidates	·	·	·	·	·	·	·	·	·	1
	true terms	·	·	·	·	·	·	·	·	·	0
	screened	·	·	·	·	·	·	·	·	·	0
	decides	·	·	·	·	·	·	·	·	·	0

E.9 DISCOVERY UNDER NOISE

Table 56 gives the discoveries and structural recoveries of each strongest cell from clean data to twice the noise limit, and Table 57 the runs with no answer: under noise, a reply cut at the provider's limit leaves more than half of GLM's runs with no answer, a twelfth of DeepSeek's and almost none of Qwen's or Gemma's. Refitted on the draw the models were given, the sparse regression recovers the structure on most systems at half the limit, on three at the limit and on none at twice the limit (Table 58). On Inviscid Burgers the gains under noise come from breaking the tie between the law and its equivalent forms (Appendix D.1), not from reading through the noise.

Table 56: Structures recovered and equations discovered, out of 10 planned runs, in each model's strongest cell from clean data to twice the noise limit.

Model	Noise		KdV	K–S	Bu_v	Bu_i	VdP	A–C	Wave	Comp	Osc	Rat
GLM	clean	structures	10	10	7	3	10	10	5	6	5	2
		discovered	10	10	7	3	10	10	4	6	5	2
	0.5×	structures	9	5	4	10	6	3	0	3	0	0
		discovered	9	5	4	10	6	3	0	3	0	0
	1×	structures	10	8	0	6	5	4	0	2	0	0
		discovered	10	8	0	6	5	4	0	2	0	0
	2×	structures	7	3	5	5	10	4	1	0	0	0
		discovered	7	3	2	5	10	4	1	0	0	0
Qwen	clean	structures	10	10	10	0	10	7	5	7	3	0
		discovered	10	10	7	0	10	7	3	6	3	0
	0.5×	structures	10	9	9	10	9	1	5	8	0	1
		discovered	10	9	5	10	6	0	2	6	0	0
	1×	structures	10	10	10	7	9	1	6	10	0	2
		discovered	9	10	2	7	6	0	0	8	0	0
	2×	structures	10	8	10	9	10	0	2	10	0	0
		discovered	10	8	7	9	9	0	0	6	0	0
DeepSeek	clean	structures	9	9	8	3	8	6	10	1	0	0
		discovered	9	9	3	3	6	4	10	1	0	0
	0.5×	structures	9	10	10	10	9	7	9	2	0	0
		discovered	9	10	6	10	3	0	0	0	0	0
	1×	structures	10	10	10	4	5	4	4	3	0	0
		discovered	10	9	9	4	3	0	0	2	0	0
	2×	structures	6	8	9	3	6	6	6	3	0	0
		discovered	5	8	6	3	3	2	0	3	0	0
Gemma	clean	structures	10	9	6	2	4	0	9	0	0	0
		discovered	1	2	2	0	0	0	8	0	0	0
	0.5×	structures	9	6	2	1	0	0	4	1	0	0
		discovered	0	2	2	1	0	0	0	0	0	0
	1×	structures	8	8	5	0	0	1	4	1	0	0
		discovered	4	0	0	0	0	0	0	1	0	0
	2×	structures	6	4	5	1	1	0	6	1	0	0
		discovered	2	1	2	1	0	0	0	0	0	0

Table 57: Runs without an answer out of 10 planned runs in each model's strongest cell, without noise (the main sweep) and at 0.5×, 1× and 2× the noise limit. A dagger marks cells in which most of those runs were cut before 131,072 completion tokens.

Level	KdV	K–S	Bu_v	Bu_i	VdP	A–C	Wave	Comp	Osc	Rat
GLM, none	·	·	2	·	·	·	4	2	3	$6^\dagger$
0.5×	1	5	5	·	4	7	10	5	10	$9^\dagger$
1×	·	1	10	·	5	6	10	$8^\dagger$	10	$9^\dagger$
2×	1	7	1	·	·	4	9	$9^\dagger$	5	$10^\dagger$
Qwen, none	·	·	·	·	·	·	·	·	·	·
0.5×	·	·	·	·	·	·	·	·	·	·
1×	·	·	·	·	1	·	·	·	·	·
2×	·	·	·	·	·	·	·	·	·	·
DeepSeek, none	·	·	·	·	·	·	·	1	·	·
0.5×	1	·	·	·	·	·	1	·	4	·
1×	·	·	·	·	4	2	·	1	2	1
2×	2	·	1	3	1	1	·	·	·	·
Gemma, none	·	·	·	·	·	·	·	·	·	·
0.5×	·	·	·	·	·	·	·	·	·	·
1×	·	·	·	·	·	·	·	·	·	·
2×	·	·	·	·	·	·	·	·	·	·

Coefficients. Table 58 separates the coefficients from the structure. Across the three noise levels GLM places every coefficient within 10% in nearly every structure it recovers, Qwen in three quarters, DeepSeek in two thirds and Gemma in a fifth, and nearly all of these are exact. On the Wave equation Qwen, DeepSeek and Gemma recover the structure in noisy runs but almost never its coefficient, and GLM rarely recovers the structure.

Table 58: Coefficients under noise, each model on its strongest cell: runs that recover the structure, those of them with every coefficient within 10%, and those with the exact coefficients. The first rows give the sparse regression refitted on the draw the models were given: its ε_∞ where it recovers the structure, · where it does not.

Level	Runs	KdV	K–S	Bu_v	Bu_i	VdP	A–C	Wave	Comp	Osc	Rat
regression, 0.5×		.021	·	.0013	$5.3{\cdot}10^{-5}$	·	.22	.026	.0024	.0086	$2.1{\cdot}10^{-4}$
regression, 1×		·	.89	·	$5.4{\cdot}10^{-5}$	·	·	·	·	.061	·
regression, 2×		·	·	·	·	·	·	·	·	·	·
GLM, none	structure	10	10	7	3	10	10	5	6	5	2
	within 10%	10	10	7	3	10	10	4	6	5	2
	exact	10	10	7	3	10	10	4	5	5	2
0.5×	structure	9	5	4	10	6	3	·	3	·	·
	within 10%	9	5	4	10	6	3	·	3	·	·
	exact	9	5	4	10	6	3	·	3	·	·
1×	structure	10	8	·	6	5	4	·	2	·	·
	within 10%	10	8	·	6	5	4	·	2	·	·
	exact	10	8	·	6	5	4	·	2	·	·
2×	structure	7	3	5	5	10	4	1	·	·	·
	within 10%	7	3	2	5	10	4	1	·	·	·
	exact	7	3	2	5	10	4	1	·	·	·
Qwen, none	structure	10	10	10	·	10	7	5	7	3	·
	within 10%	10	10	7	·	10	7	3	6	3	·
	exact	10	10	7	·	10	7	3	6	3	·
0.5×	structure	10	9	9	10	9	1	5	8	·	1
	within 10%	10	9	5	10	6	0	2	6	·	0
	exact	10	9	5	10	6	0	2	6	·	0
1×	structure	10	10	10	7	9	1	6	10	·	2
	within 10%	9	10	2	7	6	0	0	8	·	0
	exact	9	10	1	7	5	0	0	6	·	0
2×	structure	10	8	10	9	10	·	2	10	·	·
	within 10%	10	8	7	9	9	·	0	6	·	·
	exact	10	8	7	9	8	·	0	4	·	·
DeepSeek, none	structure	9	9	8	3	8	6	10	1	·	·
	within 10%	9	9	3	3	6	4	10	1	·	·
	exact	9	9	3	3	5	3	10	1	·	·
0.5×	structure	9	10	10	10	9	7	9	2	·	·
	within 10%	9	10	6	10	3	0	0	0	·	·
	exact	9	10	6	10	3	0	0	0	·	·
1×	structure	10	10	10	4	5	4	4	3	·	·
	within 10%	10	9	9	4	3	0	0	2	·	·
	exact	10	9	8	4	3	0	0	1	·	·
2×	structure	6	8	9	3	6	6	6	3	·	·
	within 10%	5	8	6	3	3	2	0	3	·	·
	exact	5	8	6	3	3	2	0	3	·	·
Gemma, none	structure	10	9	6	2	4	·	9	·	·	·
	within 10%	1	2	2	0	0	·	8	·	·	·
	exact	1	2	2	0	0	·	8	·	·	·
0.5×	structure	9	6	2	1	·	·	4	1	·	·
	within 10%	0	2	2	1	·	·	0	0	·	·
	exact	0	2	2	1	·	·	0	0	·	·
1×	structure	8	8	5	·	·	1	4	1	·	·
	within 10%	4	0	0	·	·	0	0	1	·	·
	exact	4	0	0	·	·	0	0	1	·	·
2×	structure	6	4	5	1	1	·	6	1	·	·
	within 10%	2	1	2	1	0	·	0	0	·	·
	exact	2	1	2	1	0	·	0	0	·	·

Saying that the data are noisy. A reply says the data are noisy when its reasoning attributes noise to the samples or to the supplied fields; calling a negligible value noise, or the reply's own differencing error, does not count. Table 59 gives the counts. On clean data almost no reply says it. Under noise, every model's replies say it in at most a seventh of the runs; GLM's replies that say it nearly all discover the law, and the other models' mostly do not. GLM says it mainly on Van der Pol and the oscillator, where many of its replies estimate the noise level and, on Van der Pol, read the

coefficients from global features instead of pointwise differences; DeepSeek's and Gemma's replies say it of the supplied derivative fields.

Table 59: Replies that say the data are noisy, with those of them that discover the equation in parentheses, and replies with reasoning, in each model's strongest cell from clean data to twice the noise limit.

Model	Noise	Replies	KdV	K–S	Bu_v	Bu_i	VdP	A–C	Wave	Comp	Osc	Rat
GLM	none	say noisy	0	0	1 (1)	0	2 (2)	0	0	0	0	0
		with reasoning	10	10	8	10	10	10	6	8	7	4
	$0.5\times$	say noisy	0	0	1 (1)	1 (1)	5 (5)	1 (1)	–	0	–	0
		with reasoning	9	5	5	10	6	3	–	5	–	1
	$1\times$	say noisy	0	0	–	1 (1)	5 (5)	2 (2)	–	0	–	0
		with reasoning	10	9	–	10	5	4	–	2	–	1
	$2\times$	say noisy	0	0	0	0	9 (9)	2 (2)	0	0	4 (0)	–
		with reasoning	9	3	9	10	10	6	1	1	5	–
Qwen	none	say noisy	0	0	0	0	0	0	0	0	0	0
		with reasoning	10	10	10	10	10	10	10	10	10	10
	$0.5\times$	say noisy	0	0	0	0	0	0	0	0	0	0
		with reasoning	10	10	10	10	10	10	10	10	10	10
	$1\times$	say noisy	0	0	1 (0)	0	0	1 (0)	0	0	2 (0)	0
		with reasoning	10	10	10	10	9	10	10	10	10	10
	$2\times$	say noisy	0	0	0	0	0	0	1 (0)	0	3 (0)	0
		with reasoning	10	10	10	10	10	10	10	10	10	10
DeepSeek	none	say noisy	0	0	0	0	0	0	0	0	0	0
		with reasoning	10	10	10	8	10	10	10	9	10	10
	$0.5\times$	say noisy	0	0	1 (1)	0	2 (0)	2 (0)	0	1 (0)	0	0
		with reasoning	9	10	10	10	10	10	9	10	6	10
	$1\times$	say noisy	0	0	1 (1)	0	1 (1)	0	1 (0)	0	0	0
		with reasoning	10	10	10	10	6	8	10	9	8	9
	$2\times$	say noisy	0	0	1 (0)	1 (1)	4 (1)	3 (0)	2 (0)	0	2 (0)	0
		with reasoning	8	10	9	7	9	9	10	10	10	10
Gemma	none	say noisy	0	0	0	0	0	0	0	0	0	0
		with reasoning	10	10	10	10	10	10	10	10	10	10
	$0.5\times$	say noisy	0	0	0	0	3 (0)	0	5 (0)	0	0	0
		with reasoning	10	10	10	10	10	10	10	10	10	10
	$1\times$	say noisy	0	1 (0)	0	0	5 (0)	3 (0)	2 (0)	0	1 (0)	0
		with reasoning	10	10	10	10	10	10	10	10	10	10
	$2\times$	say noisy	0	5 (1)	0	2 (0)	4 (0)	4 (0)	7 (0)	0	1 (0)	0
		with reasoning	10	10	10	10	10	10	10	10	10	10

Reasoning. Tables 60 to 63 give the move families at each level. From clean data to twice the limit, readings of values at single points become less frequent for GLM, Qwen and DeepSeek on most systems, and readings of scales more frequent for Qwen and DeepSeek.

Table 60: GLM, strongest cell from clean data to twice the noise limit: replies using each family of moves, among the replies of the runs that discover the equation and among the other replies.

Family	Replies	KdV	K–S	Bu_v	Bu_i	VdP	A–C	Wave	Comp	Osc	Rat
noise none											
replies	discovering	10	10	7	3	10	10	4	6	5	2
	other	0	0	1	7	0	0	2	2	2	2
names the law	discovering	10	10	7	3	10	10	4	6	2	0
	other	0	0	1	7	0	0	2	2	0	1
takes coefficients from a standard form	discovering	·	4	·	·	·	·	·	0	·	·
	other	·	0	·	·	·	·	·	1	·	·
reads values at points	discovering	6	8	7	3	10	10	4	6	5	2
	other	0	0	1	7	0	0	2	2	1	1
tests constancy or verifies	discovering	10	8	7	3	10	9	3	6	4	2
	other	0	0	1	6	0	0	1	1	2	2
reads a scale or closed form	discovering	10	10	6	3	10	10	4	6	5	2
	other	0	0	1	7	0	0	1	2	2	2
rules out by a signature	discovering	9	9	7	3	7	8	4	5	4	2
	other	0	0	0	7	0	0	1	2	2	2
balances magnitudes	discovering	·	1	1	·	·	·	0	1	·	·
	other	·	0	0	·	·	·	1	0	·	·
acknowledges or suppresses noise	discovering	3	·	3	·	4	·	·	0	2	1
	other	0	·	0	·	0	·	·	1	0	0
noise $0.5\times$											
replies	discovering	9	5	4	10	6	3	0	3	0	0
	other	0	0	1	0	0	0	0	2	0	1
names the law	discovering	9	5	4	10	6	3	·	3	·	0
	other	0	0	0	0	0	0	·	1	·	1
takes coefficients from a standard form	discovering	·	1	·	·	·	·	·	0	·	·
	other	·	0	·	·	·	·	·	1	·	·
reads values at points	discovering	6	4	4	10	6	2	·	3	·	·
	other	0	0	0	0	0	0	·	2	·	·
tests constancy or verifies	discovering	9	4	4	9	6	3	·	2	·	0
	other	0	0	1	0	0	0	·	2	·	1
reads a scale or closed form	discovering	9	5	3	10	6	3	·	0	·	0
	other	0	0	1	0	0	0	·	2	·	1
rules out by a signature	discovering	9	5	4	10	4	3	·	3	·	0
	other	0	0	1	0	0	0	·	1	·	1
balances magnitudes	discovering	·	·	·	1	1	·	·	1	·	·
	other	·	·	·	0	0	·	·	0	·	·
acknowledges or suppresses noise	discovering	5	1	1	2	5	1	·	·	·	·
	other	0	0	0	0	0	0	·	·	·	·

Table 60 (continued).

Family	Replies	KdV	K–S	Bu_v	Bu_i	VdP	A–C	Wave	Comp	Osc	Rat
noise 1×											
replies	discovering	10	8	0	6	5	4	0	2	0	0
	other	0	1	0	4	0	0	0	0	0	1
names the law	discovering	10	8	·	6	5	4	·	2	·	0
	other	0	1	·	4	0	0	·	0	·	1
takes coefficients from a standard form	discovering	1	1	·	·	·	·	·	1	·	·
	other	0	1	·	·	·	·	·	0	·	·
reads values at points	discovering	5	4	·	6	5	4	·	2	·	0
	other	0	1	·	4	0	0	·	0	·	1
tests constancy or verifies	discovering	10	7	·	6	5	4	·	2	·	0
	other	0	0	·	4	0	0	·	0	·	1
reads a scale or closed form	discovering	10	8	·	6	5	2	·	2	·	0
	other	0	1	·	4	0	0	·	0	·	1
rules out by a signature	discovering	10	8	·	6	3	3	·	2	·	0
	other	0	1	·	4	0	0	·	0	·	1
balances magnitudes		·	·	·	·	·	·	·	·	·	·
acknowledges or suppresses noise	discovering	4	·	·	2	5	2	·	·	·	·
	other	0	·	·	0	0	0	·	·	·	·
noise 2×											
replies	discovering	7	3	2	5	10	4	1	0	0	0
	other	2	0	7	5	0	2	0	1	5	0
names the law	discovering	7	3	2	5	10	4	1	0	·	·
	other	2	0	5	2	0	2	0	1	·	·
takes coefficients from a standard form	discovering	3	·	0	·	·	0	·	·	·	·
	other	0	·	1	·	·	1	·	·	·	·
reads values at points	discovering	4	2	0	5	10	4	1	·	0	·
	other	1	0	4	1	0	1	0	·	4	·
tests constancy or verifies	discovering	7	2	2	5	9	4	1	0	0	·
	other	2	0	5	4	0	2	0	1	4	·
reads a scale or closed form	discovering	7	3	2	5	10	4	1	0	0	·
	other	2	0	7	5	0	2	0	1	5	·
rules out by a signature	discovering	5	3	1	5	6	2	1	0	0	·
	other	2	0	7	4	0	2	0	1	5	·
balances magnitudes	discovering	1	·	1	·	·	·	·	·	·	·
	other	0	·	0	·	·	·	·	·	·	·
acknowledges or suppresses noise	discovering	2	·	1	1	9	2	·	·	0	·
	other	0	·	1	0	0	0	·	·	4	·

Table 61: Qwen, strongest cell from clean data to twice the noise limit: replies using each family of moves, among the replies of the runs that discover the equation and among the other replies.

Family	Replies	KdV	K–S	Bu_v	Bu_i	VdP	A–C	Wave	Comp	Osc	Rat
noise none											
replies	discovering	10	10	7	0	10	7	3	6	3	0
	other	0	0	3	10	0	3	7	4	7	10
names the law	discovering	10	10	7	0	10	3	3	2	·	0
	other	0	0	3	10	0	1	5	3	·	8
takes coefficients from a standard form	discovering	4	4	·	·	·	0	1	0	·	0
	other	0	0	·	·	·	1	0	2	·	1
reads values at points	discovering	6	4	3	0	10	7	3	6	3	0
	other	0	0	2	10	0	3	7	4	6	9
tests constancy or verifies	discovering	4	3	3	0	6	2	0	5	2	0
	other	0	0	1	2	0	1	3	2	2	5
reads a scale or closed form	discovering	9	6	5	0	7	4	2	3	3	0
	other	0	0	2	10	0	1	4	4	7	8
rules out by a signature	discovering	9	8	6	0	9	7	2	6	1	0
	other	0	0	2	8	0	2	7	3	5	8
balances magnitudes	discovering	·	7	3	·	2	1	1	·	·	0
	other	·	0	0	·	0	1	2	·	·	2
acknowledges or suppresses noise	discovering	2	·	·	0	1	·	·	·	0	·
	other	0	·	·	1	0	·	·	·	1	·
noise $0.5\times$											
replies	discovering	10	9	5	10	6	0	2	6	0	0
	other	0	1	5	0	4	10	8	4	10	10
names the law	discovering	10	9	5	9	6	0	2	1	·	0
	other	0	1	5	0	2	8	4	1	·	8
takes coefficients from a standard form	discovering	1	7	2	·	0	0	0	·	0	0
	other	0	0	0	·	1	1	2	·	1	1
reads values at points	discovering	4	4	4	9	4	0	2	6	0	0
	other	0	1	3	0	2	7	7	3	9	4
tests constancy or verifies	discovering	2	3	1	6	2	0	0	2	0	0
	other	0	0	2	0	1	1	2	3	2	2
reads a scale or closed form	discovering	8	7	5	9	6	0	0	2	0	0
	other	0	1	5	0	4	9	7	1	10	10
rules out by a signature	discovering	8	7	5	9	4	0	1	6	0	0
	other	0	1	4	0	3	8	6	4	7	9
balances magnitudes	discovering	·	7	2	·	·	0	2	·	0	0
	other	·	1	2	·	·	2	6	·	2	2
acknowledges or suppresses noise	discovering	3	·	·	1	·	·	·	·	·	·
	other	0	·	·	0	·	·	·	·	·	·

Table 61 (continued).

Family	Replies	KdV	K–S	Bu_v	Bu_i	VdP	A–C	Wave	Comp	Osc	Rat
noise 1×											
replies	discovering	9	10	2	7	6	0	0	8	0	0
	other	1	0	8	3	3	10	10	2	10	10
names the law	discovering	9	10	2	7	6	0	0	2	·	0
	other	1	0	7	3	2	9	8	1	·	7
takes coefficients from a standard form	discovering	1	4	1	·	1	·	0	2	·	0
	other	0	0	1	·	0	·	2	0	·	2
reads values at points	discovering	6	4	0	6	6	0	0	6	0	0
	other	1	0	2	2	3	6	5	1	8	6
tests constancy or verifies	discovering	3	1	0	2	2	0	0	1	0	0
	other	0	0	4	2	1	4	3	1	4	3
reads a scale or closed form	discovering	9	7	1	7	6	0	0	5	0	0
	other	0	0	6	3	3	6	7	1	10	9
rules out by a signature	discovering	9	10	2	6	4	0	0	7	0	0
	other	1	0	6	3	3	10	8	1	5	5
balances magnitudes	discovering	·	6	0	·	·	0	0	1	·	0
	other	·	0	2	·	·	4	2	1	·	1
acknowledges or suppresses noise	discovering	3	·	0	·	·	0	·	·	0	·
	other	1	·	1	·	·	3	·	·	2	·
noise 2×											
replies	discovering	10	8	7	9	9	0	0	6	0	0
	other	0	2	3	1	1	10	10	4	10	10
names the law	discovering	10	8	7	9	9	0	0	3	0	0
	other	0	2	3	1	1	9	7	2	1	7
takes coefficients from a standard form	discovering	4	5	1	·	0	0	0	·	·	0
	other	0	2	0	·	1	1	3	·	·	1
reads values at points	discovering	3	0	1	9	8	0	0	6	0	0
	other	0	2	0	1	1	5	7	3	6	6
tests constancy or verifies	discovering	3	3	2	3	5	0	0	3	0	0
	other	0	0	1	0	0	3	1	2	3	1
reads a scale or closed form	discovering	10	6	7	9	9	0	0	5	0	0
	other	0	2	3	1	1	7	5	1	10	10
rules out by a signature	discovering	7	7	5	5	3	0	0	6	0	0
	other	0	2	2	1	1	9	7	4	5	8
balances magnitudes	discovering	·	6	3	·	1	0	0	·	·	0
	other	·	1	0	·	0	3	5	·	·	3
acknowledges or suppresses noise	discovering	2	0	·	·	·	0	0	·	0	·
	other	0	1	·	·	·	1	1	·	4	·

Table 62: DeepSeek, strongest cell from clean data to twice the noise limit: replies using each family of moves, among the replies of the runs that discover the equation and among the other replies.

Family	Replies	KdV	K–S	Bu_v	Bu_i	VdP	A–C	Wave	Comp	Osc	Rat
noise none											
replies	discovering	9	9	3	3	6	4	10	1	0	0
	other	1	1	7	7	4	6	0	8	10	10
names the law	discovering	9	9	3	3	6	4	9	1	0	0
	other	1	1	6	7	4	5	0	5	2	4
takes coefficients from a standard form	discovering	3	4	·	·	·	1	1	·	·	0
	other	0	0	·	·	·	2	0	·	·	1
reads values at points	discovering	5	4	2	2	6	2	7	1	0	0
	other	1	1	1	7	4	4	0	8	5	8
tests constancy or verifies	discovering	8	1	3	2	3	·	3	1	0	0
	other	0	1	0	5	1	·	0	4	3	1
reads a scale or closed form	discovering	6	4	1	3	3	2	2	0	0	0
	other	0	1	2	7	3	1	0	3	9	6
rules out by a signature	discovering	7	6	2	1	5	2	9	1	0	0
	other	0	1	4	6	2	4	0	7	7	8
balances magnitudes	discovering	1	2	0	·	·	0	2	·	0	0
	other	0	0	2	·	·	1	0	·	2	3
acknowledges or suppresses noise	discovering	0	·	1	1	1	·	3	·	·	·
	other	1	·	0	2	0	·	0	·	·	·
noise $0.5\times$											
replies	discovering	9	10	6	10	3	0	0	0	0	0
	other	0	0	4	0	7	10	9	10	6	10
names the law	discovering	9	10	6	10	2	0	0	0	0	0
	other	0	0	4	0	7	10	9	8	4	9
takes coefficients from a standard form	discovering	6	5	2	1	·	0	0	0	·	0
	other	0	0	0	0	·	2	1	1	·	1
reads values at points	discovering	5	1	3	8	2	0	0	0	0	0
	other	0	0	2	0	5	7	9	6	2	6
tests constancy or verifies	discovering	3	·	4	5	1	0	0	0	·	0
	other	0	·	0	0	4	1	2	4	·	4
reads a scale or closed form	discovering	5	2	4	9	2	0	0	0	0	0
	other	0	0	3	0	7	3	3	6	6	8
rules out by a signature	discovering	7	5	5	5	1	0	0	0	0	0
	other	0	0	2	0	2	4	7	8	1	3
balances magnitudes	discovering	·	2	2	1	0	0	0	0	·	0
	other	·	0	0	0	1	2	2	1	·	3
acknowledges or suppresses noise	discovering	·	·	·	5	0	0	·	0	0	·
	other	·	·	·	0	2	2	·	1	1	·

Table 62 (continued).

Family	Replies	KdV	K–S	Bu_v	Bu_i	VdP	A–C	Wave	Comp	Osc	Rat
noise 1×											
replies	discovering	10	9	9	4	3	0	0	2	0	0
	other	0	1	1	6	3	8	10	7	8	9
names the law	discovering	10	9	9	4	3	0	0	1	·	0
	other	0	1	1	6	2	8	7	5	·	8
takes coefficients from a standard form	discovering	5	5	1	·	·	0	·	·	·	0
	other	0	1	0	·	·	2	·	·	·	1
reads values at points	discovering	2	·	3	3	1	0	0	2	0	0
	other	0	·	0	4	1	4	7	5	3	4
tests constancy or verifies	discovering	5	1	6	2	1	0	0	0	0	0
	other	0	0	0	2	0	3	1	5	1	4
reads a scale or closed form	discovering	9	1	3	4	3	0	0	2	0	0
	other	0	0	1	6	2	4	2	5	5	6
rules out by a signature	discovering	5	5	4	3	1	0	0	1	0	0
	other	0	1	0	3	2	2	4	5	2	3
balances magnitudes	discovering	·	3	2	·	0	0	0	0	·	·
	other	·	1	0	·	1	6	4	3	·	·
acknowledges or suppresses noise	discovering	1	1	1	0	1	·	0	·	·	·
	other	0	0	0	2	0	·	3	·	·	·
noise 2×											
replies	discovering	5	8	6	3	3	2	0	3	0	0
	other	3	2	3	4	6	7	10	7	10	10
names the law	discovering	5	8	5	3	3	2	0	2	0	0
	other	3	2	2	4	4	7	8	6	5	7
takes coefficients from a standard form	discovering	1	4	0	·	·	1	0	0	·	·
	other	0	1	1	·	·	2	1	2	·	·
reads values at points	discovering	2	2	4	3	0	2	0	3	·	0
	other	1	0	1	4	4	0	6	6	·	7
tests constancy or verifies	discovering	4	·	0	2	1	0	0	1	0	0
	other	1	·	1	3	2	1	2	4	1	5
reads a scale or closed form	discovering	4	2	2	3	3	1	0	0	0	0
	other	3	0	1	4	5	3	2	6	9	8
rules out by a signature	discovering	3	4	5	2	0	2	0	2	0	0
	other	2	0	1	3	4	5	6	5	2	5
balances magnitudes	discovering	1	1	·	·	·	1	0	0	0	0
	other	0	2	·	·	·	2	6	1	1	1
acknowledges or suppresses noise	discovering	·	·	1	2	1	1	0	0	0	0
	other	·	·	1	1	3	5	2	1	2	2

Table 63: Gemma, strongest cell from clean data to twice the noise limit: replies using each family of moves, among the replies of the runs that discover the equation and among the other replies.

Family	Replies	KdV	K–S	Bu_v	Bu_i	VdP	A–C	Wave	Comp	Osc	Rat
noise none											
replies	discovering	1	2	2	0	0	0	8	0	0	0
	other	9	8	8	10	10	10	2	10	10	10
names the law	discovering	1	2	2	·	0	0	8	0	·	0
	other	9	8	8	·	9	10	1	10	·	10
takes coefficients from a standard form	discovering	0	2	0	·	·	·	·	0	0	·
	other	3	5	1	·	·	·	·	3	1	·
reads values at points	discovering	1	1	2	0	0	0	8	0	0	0
	other	6	4	6	10	10	9	2	9	8	10
tests constancy or verifies	discovering	0	1	1	0	0	0	3	0	0	0
	other	4	3	5	10	4	1	1	9	9	4
reads a scale or closed form	discovering	1	·	0	0	0	·	·	·	0	0
	other	7	·	2	6	10	·	·	·	10	1
rules out by a signature	discovering	1	2	2	0	0	0	6	0	0	0
	other	9	6	6	7	8	5	2	8	3	9
balances magnitudes	discovering	1	1	1	·	0	0	6	0	0	0
	other	5	5	3	·	1	9	1	1	2	2
acknowledges or suppresses noise	discovering	·	1	·	0	·	0	·	0	·	0
	other	·	2	·	3	·	1	·	1	·	1
noise $0.5\times$											
replies	discovering	0	2	2	1	0	0	0	0	0	0
	other	10	8	8	9	10	10	10	10	10	10
names the law	discovering	0	2	2	0	0	0	0	0	0	0
	other	10	8	8	1	6	10	9	10	2	10
takes coefficients from a standard form	discovering	0	2	1	·	·	0	·	·	·	0
	other	4	4	4	·	·	1	·	·	·	1
reads values at points	discovering	0	0	1	1	0	0	0	0	0	0
	other	8	3	5	9	10	10	10	10	4	9
tests constancy or verifies	discovering	0	0	0	1	0	0	0	0	0	0
	other	9	3	4	5	6	2	1	6	3	5
reads a scale or closed form	discovering	0	·	0	·	0	·	·	·	0	·
	other	8	·	5	·	10	·	·	·	10	·
rules out by a signature	discovering	0	1	0	0	0	0	0	0	0	0
	other	8	5	5	8	9	3	7	10	4	10
balances magnitudes	discovering	0	2	1	0	·	0	0	0	0	0
	other	3	6	5	2	·	8	7	3	2	3
acknowledges or suppresses noise	discovering	·	0	·	1	0	·	0	0	·	·
	other	·	1	·	2	4	·	4	1	·	·

Table 63 (continued).

Family	Replies	KdV	K–S	Bu_v	Bu_i	VdP	A–C	Wave	Comp	Osc	Rat
noise 1×											
replies	discovering	4	0	0	0	0	0	0	1	0	0
	other	6	10	10	10	10	10	10	9	10	10
names the law	discovering	4	0	0	0	0	0	0	1	0	0
	other	6	10	10	1	8	10	8	9	1	9
takes coefficients from a standard form	discovering	2	0	0	·	·	0	·	·	0	0
	other	2	6	4	·	·	1	·	·	2	1
reads values at points	discovering	4	0	0	0	0	0	0	1	0	0
	other	3	3	7	10	6	7	10	9	8	10
tests constancy or verifies	discovering	2	0	0	0	0	0	0	1	·	0
	other	3	3	3	3	4	2	3	6	·	8
reads a scale or closed form	discovering	3	·	0	0	0	·	·	0	0	0
	other	3	·	6	1	10	·	·	1	10	2
rules out by a signature	discovering	4	0	0	0	0	0	0	1	0	0
	other	5	10	5	9	9	3	9	9	3	10
balances magnitudes	discovering	2	0	0	0	0	0	0	0	0	0
	other	3	6	6	5	1	5	5	1	3	5
acknowledges or suppresses noise	discovering	·	0	·	0	0	0	0	·	0	·
	other	·	6	·	3	6	3	2	·	1	·
noise 2×											
replies	discovering	2	1	2	1	0	0	0	0	0	0
	other	8	9	8	9	10	10	10	10	10	10
names the law	discovering	2	1	2	1	0	0	0	0	0	0
	other	8	9	8	1	6	10	9	10	1	10
takes coefficients from a standard form	discovering	2	0	0	·	·	0	·	·	·	·
	other	2	5	1	·	·	1	·	·	·	·
reads values at points	discovering	2	1	1	1	0	0	0	0	0	0
	other	7	5	8	8	9	9	10	10	4	10
tests constancy or verifies	discovering	1	1	1	1	0	0	0	0	0	0
	other	5	1	4	5	5	3	4	3	7	6
reads a scale or closed form	discovering	2	·	0	0	0	·	·	0	0	·
	other	6	·	4	1	10	·	·	1	10	·
rules out by a signature	discovering	2	1	2	1	0	0	0	0	0	0
	other	7	8	2	4	9	1	6	9	4	10
balances magnitudes	discovering	0	1	0	·	0	0	0	0	0	0
	other	5	4	3	·	2	8	9	3	1	5
acknowledges or suppresses noise	discovering	·	1	·	0	0	0	0	0	0	0
	other	·	4	·	6	3	4	7	2	1	1

F PER-SYSTEM REASONING

This appendix takes each system in turn and asks what the successful reasoning does that the unsuccessful reasoning does not, and complements the counts of Appendix E.4. Counts are out of the 10 planned runs of one cell of the main sweep, and the reasoning quoted is the text the model wrote before its answer. Of the 15 point readings quoted, 12 match the values their run's prompt carried to the precision stated, and the others are off in the last digit or in the time they are placed at.

F.1 THE FORCED DAMPED OSCILLATOR

On $u_{tt} = -\sin(2t)\,u_t - 4u + 1.5t$ the successful runs proceed by subtraction. They fix the frequency from the spacing of the extrema, which gives $-4u$, form the residual $R = u_{tt} + 4u$ by finite differences, and observe that it equals $1.5t$ at the extrema, where $u_t = 0$, and departs from it elsewhere: the missing term carries u_t as a factor. One GLM run then checks that $(1.5t - R)/(u_t \sin 2t)$ stays near 1 across all phases; another reaches the term from a component at frequency 4 in the residual, which a damping modulated at frequency 2 produces. Qwen's successes follow the first route.

The failures break where this chain breaks. Gemma writes $-4u$ in most of its best guesses but never forms a residual, and elsewhere misreads the period or reads the growth as constant anti-damping. A GLM miss assumes that any coefficient of u_t must be constant, rules damping out because the amplitude ratios vary, and fits the residual as exogenous forcing; DeepSeek's misses read the growing envelope as resonant forcing. Some Qwen misses come closest, noting residuals "largest when u_t is large" and then attributing them to u_t^2 without dividing through. A picture shows the period, the midline slope and the growing envelope, the evidence the failures stop at, but not the pointwise residual the successes need: GLM recovers the structure in 5 runs from the samples and in none from the plot.

F.2 THE VAN DER POL OSCILLATOR

The successful runs read the period first ($T \approx 2\pi$, unchanged as the amplitude grows), which fixes the stiffness at -1, form $r = u_{tt} + u$, and read it where the nonlinear part switches off: at a zero crossing $u \approx 0$, r equals μu_t alone, and one GLM run reports $r = 0.348$ at $u_t = 1.744$ and $r = -0.224$ at $u_t = -1.120$, both giving $\mu = 0.200$. The decisive step tests whether the coefficient is constant: another GLM run forms $(u_{tt} + u)/[(1 - u^2)u_t]$ near the zero crossings with both signs of u_t, finds that it "clusters tightly at 0.198–0.203", and rejects a Rayleigh term because the effective damping tracks $1 - u^2$ and not u_t^2.

The failures skip a measurement. Gemma names the family but sets the coefficients from a misread period and a balance argument; a DeepSeek run reads a slight lengthening of the period as a softening cubic and drops the restoring term. The plot hides the saturation of the amplitude: Qwen recovers the structure in every run from the samples and in 6 of 10 from the plot, where the failing runs read the saturating growth as unbounded. Six GLM runs write the law multiplied through by u, $u\,u_{tt} = -u^2 + 0.2\,u\,u_t - 0.2\,u^3 u_t$, which fits marginally better on both objectives and survives the screen whenever a run returns both forms; dividing out the common factor makes it the same equation (Appendix D).

F.3 THE WAVE EQUATION

The runs that estimate the derivatives themselves pivot on one observation: the crest rises before it falls, which rules out diffusion and forces a second time derivative. With the derivative fields supplied, runs read the ratio u_{tt}/u_{xx} directly. Recovery then turns on testing whether that ratio is constant: GLM's cleanest run reads it at the crest and near the boundary, finds it "in 0.037–0.043 at every location", and confirms 0.04 by summing sine modes.

Structure and coefficient fail separately. A GLM run reaches the Wave structure through the crest's oscillation frequency, misreads it and returns 0.0625; Gemma's reasoning stays diffusive, and wherever the derivative fields are withheld it recovers the structure without the coefficient, once dismissing the decisive rise of the crest as noise. Other failures read the ratio at a single point or keep a first-order left side. Supplying the derivative fields hands over the second difference: from the samples in one call, DeepSeek goes from 1 recovered structure to 10 and Qwen from 5 to 10.

F.4 THE INVISCID BURGERS EQUATION

The data are a straight ramp, $u = x/(t-2)$, that steepens as it grows, and the successful reasoning measures the growth rate rather than assuming it constant. Working from the samples, GLM reads the amplitude row by row at intermediate times ("0.5 ($t = 0$), 0.50505 ($t = 0.02 = 1/1.98$), 0.58824 ($t = 0.3 = 1/1.7$)"), concludes that the amplification is algebraic, $f(t) = 1/(2-t)$, fixes the coefficient at -1 from the slope law $a' = -a^2$, and checks it pointwise; Qwen and DeepSeek mostly read the first and last rows and state $u = -x/(2-t)$ directly. Every run of GLM, Qwen and DeepSeek in its strongest configuration writes $u_t = -u\,u_x$ with its coefficient within 10%, and most also write forms that equal it on this solution, which the screen then keeps in its place (Appendix D.1).

Most failing runs read the growth off the $t = 0$ and $t = 1$ edges alone, where one doubling cannot tell growth from coordinate forcing: GLM and DeepSeek return $u_t = 0.69\,u$ from the plot, and Qwen reads the colour bands as a profile translating at constant speed. Gemma fails the same way with the samples in hand, fitting $u = 500t - 0.5x$ after checking one corner. The plot hides the intermediate rows: DeepSeek writes the law in none of its runs from the plot and in every run from the samples.

F.5 THE VISCOUS BURGERS EQUATION

Success on $u_t = -u\,u_x + 0.1\,u_{xx}$ turns on whether the pulse speed is fixed or changing. GLM's successful runs time the crest over separate intervals and tie its falling speed to the falling amplitude, which rules out linear advection–diffusion; they read the viscosity at the crest, where $u_x = 0$ and the decay is $2\nu u$ alone, and check that $(u_t - 0.1\,u_{xx})/(-u\,u_x)$ "clusters at 1.0–1.15 across points". The failures divide one displacement by one time, a single speed that cannot separate $-c\,u_x$ from $-u\,u_x$, most of them from the plot. The samples afford comparisons at the crest and the flanks and the plot only ridge tracking: Qwen recovers the structure in all 10 runs from the samples and in 5 from the plot. Gemma reverses this, seeing a pulse that stays in place in the samples and settling on pure diffusion.

F.6 THE KORTEWEG–DE VRIES EQUATION

The field is a two-soliton state, and nearly every successful GLM run measures the speed and the amplitude of each pulse and asks whether one relation fits both. Both obey $c = 2A$; the soliton relation $c = \alpha A/3$ read backwards gives $\alpha = 6$ from each pulse, and the two widths give $\beta = 1$. The second pulse is the check that the scaling is not an accident of one hump. The other models' successes also work from a single pulse, from relations between the derivatives, or from the standard coefficients. The failures break at the measurement: a GLM run from the plot misreads the slow pulse's speed and, reasoning correctly from the wrong number, selects the modified equation; DeepSeek and Gemma see both pulses in the plot move at the same speed and return linear advection. DeepSeek recovers the structure in all 10 runs from the samples and in 2 from the plot. Gemma names the right family from the samples, measures a speed that implies $\alpha \approx 10$, and falls back on the textbook value $\alpha = 6$.

F.7 THE ALLEN–CAHN EQUATION

The true law $u_t = 5u - 5u^3 + 10^{-4}u_{xx}$ makes the structure easy and the coefficients hard: every model but Gemma proposes a bistable reaction–diffusion law in most of its runs, from the pinned fronts and the ± 1 plateaus, and what separates the cells is whether the coefficients are measured. The successful runs read each where the competing term vanishes. At the centre, where $u = 0$ at $t = 0$, the reaction vanishes and $u_t = D\,u_{xx}$ gives $D \approx 10^{-4}$; in a lobe, where diffusion is negligible, a fit of the logistic solution gives the rate 5, checked at other locations.

The failures set a coefficient by assumption. DeepSeek fixes the rate at 1 because $u = \pm 1$ are the stable states ("the reaction terms u and $-u^3$ have coefficients 1 and -1"), although saturation at ± 1 fixes the ratio of the two coefficients, never their scale; a Qwen run assumes $D \approx 0.01$ and solves the lobe balance for an inflated rate. The plot shows the centre notch but not its value, and GLM's runs from the plot fall back on the width of the fronts and miss D by one or two orders of magnitude.

F.8 THE KURAMOTO–SIVASHINSKY EQUATION

GLM's successful runs work through the selected wavelength. One run measures the late-time cell size, matches it to $2\pi\sqrt{2|c_4/c_2|}$ to read the ratio of the two linear coefficients, rules out Viscous Burgers because the amplitude grows past its initial range, and takes the advection coefficient from the steepening time, rejecting 6 because a front would form far earlier than the data show. Some of Qwen's successes reach the same place through the growth rate $k^2 - k^4$. The failures stop at different points: a GLM miss reads the oscillatory train behind the front as dispersive and settles on Korteweg–de Vries, and a DeepSeek miss takes a sampling artefact for a physical period and ends with the right structure and a coefficient fifty times too large. The input shifts attention in opposite directions: Gemma recovers the structure from the plot, which displays the cells, and never from the samples, where Qwen recovers it in every run.

F.9 THE COMPOUND PDE

The law $u_t = u_x^2 + u\,u_{xx}$ is nonlinear diffusion written out, $(u\,u_x)_x$, and the runs that recover it ask whether an apparent diffusivity is constant. GLM's clearest success reads u_t/u_{xx} at the crest, where $u_x = 0$, at four times, and finds that it tracks u ("0.636 against 0.636, 0.474 against 0.477"): the diffusivity is u. DeepSeek and Qwen arrive from the boundary, where the field rises although $u_{xx} < 0$, which no positive constant diffusivity allows. The failures see the same variation and average it away, settling on a constant diffusivity from a half-life, or explain the slowing decay by a cubic sink. Some recoveries return linear diffusion as the best guess and carry $(u\,u_x)_x$ only as an alternative candidate.

F.10 THE RATIONAL PDE

The law $u_t = 0.5\,u_{xx} - 2u_x/x$ hides a drift whose speed depends on position, and every GLM recovery from the samples refuses to treat that speed as constant. One run inverts $u_t = \nu u_{xx} - c(x)\,u_x$ station by station, reads $c \approx 1.87, 1.69, 1.55, 1.03$ at $x \approx 1.07, 1.18, 1.29, 1.92$, and concludes that "$c(x) \approx 2/x$, while $\nu \approx 0.50$ stays constant"; another regresses u_t on u_{xx} and u_x/x at six stations and obtains 0.503 and -1.998. The failures that see the drift stop one step earlier: the failing runs of Qwen and DeepSeek from the samples all lead with a drift of constant speed or of speed proportional to u, some built on a tilted eigenmode whose envelope exponent comes from a single pair of points, and a GLM miss finds the slope varying and fits a drift linear in x. Gemma reports a symmetric profile and never looks for drift. Qwen recovers the law only in two-call runs, both of which carry the true form as an alternative.

G TIME AND COST

What each cell took and cost (Tables 64 and 65), as a mean and a standard deviation over its 10 runs, runs with no answer included. Configurations are abbreviated: s samples, p plot, +d derivatives supplied, 2c the answer split across two calls.

Table 64: Wall time in seconds per run, mean (standard deviation) over the 10 runs of each cell.

Configuration	KdV	K–S	Bu_v	Bu_i	VdP	A–C	Wave	Comp	Osc	Rat
GLM										
s	662	949	970	360	704	1085	1523	1232	1408	824
	(269)	(513)	(774)	(105)	(201)	(1201)	(1061)	(1598)	(1019)	(306)
p	717	405	262	578	419	765	672	307	746	342
	(436)	(198)	(86)	(330)	(234)	(291)	(298)	(154)	(487)	(151)
p+s	418	561	1172	247	697	763	1847	1304	1692	1852
	(171)	(509)	(1012)	(76)	(568)	(415)	(1101)	(769)	(946)	(1662)
s, 2c	1315	1640	1847	330	1286	2069	2673	1967	2446	1985
	(672)	(1745)	(1163)	(74)	(550)	(1505)	(1992)	(1225)	(1325)	(1130)
p, 2c	1022	962	550	674	702	1363	833	375	1591	464
	(507)	(552)	(283)	(468)	(195)	(539)	(402)	(194)	(707)	(210)
p+s, 2c	1235	1720	2857	465	2483	2395	2705	2333	2351	1959
	(675)	(1041)	(1503)	(199)	(1466)	(1145)	(942)	(609)	(1534)	(1048)
s +d	332	874	833	375	763	572	690	813	1474	973
	(189)	(489)	(333)	(108)	(381)	(485)	(440)	(374)	(653)	(249)
p +d	576	303	447	930	847	1301	1197	965	1204	1221
	(308)	(151)	(299)	(268)	(257)	(592)	(720)	(557)	(649)	(514)
p+s +d	733	732	1773	389	953	2476	1074	2393	2004	1250
	(191)	(771)	(707)	(218)	(687)	(1816)	(425)	(1186)	(1728)	(1059)
s +d, 2c	1280	1827	2591	670	1402	3351	2945	2187	2604	1827
	(776)	(2009)	(874)	(525)	(483)	(3361)	(1960)	(1000)	(1275)	(783)
p +d, 2c	945	1078	1161	1317	871	2496	2135	1701	2010	1722
	(478)	(626)	(328)	(635)	(294)	(1311)	(1198)	(985)	(1159)	(671)
p+s +d, 2c	1520	2031	2492	830	2582	1234	2608	2070	2283	2090
	(579)	(1525)	(1743)	(357)	(1304)	(1111)	(1492)	(1093)	(1065)	(756)
Qwen										
s	297	356	520	228	331	608	725	576	745	500
	(140)	(113)	(215)	(58)	(132)	(137)	(270)	(157)	(287)	(130)
p	182	201	195	283	199	172	296	143	234	106
	(121)	(96)	(111)	(104)	(80)	(64)	(179)	(43)	(106)	(30)
p+s	250	292	539	186	386	463	607	350	594	385
	(109)	(71)	(99)	(80)	(136)	(172)	(116)	(116)	(149)	(98)
s, 2c	681	675	1074	374	823	870	957	874	1074	752
	(211)	(224)	(233)	(128)	(226)	(147)	(219)	(200)	(159)	(215)
p, 2c	283	371	256	453	348	381	290	408	501	231
	(114)	(91)	(81)	(126)	(158)	(109)	(86)	(133)	(127)	(61)
p+s, 2c	671	561	754	267	643	673	1093	726	908	695
	(205)	(124)	(144)	(57)	(122)	(125)	(194)	(171)	(206)	(142)
s +d	342	305	466	242	532	590	387	457	686	515
	(106)	(111)	(102)	(36)	(201)	(243)	(138)	(101)	(165)	(104)
p +d	185	202	206	375	222	263	195	291	380	209
	(86)	(105)	(104)	(196)	(78)	(95)	(102)	(126)	(228)	(81)
p+s +d	222	253	338	241	576	352	316	425	612	504
	(107)	(85)	(148)	(36)	(142)	(135)	(76)	(170)	(82)	(105)
s +d, 2c	795	636	893	369	1065	801	996	968	1340	957
	(177)	(176)	(274)	(134)	(227)	(246)	(227)	(225)	(313)	(210)
p +d, 2c	432	414	435	533	441	369	453	316	718	370
	(139)	(113)	(100)	(230)	(126)	(145)	(101)	(138)	(197)	(71)
p+s +d, 2c	530	476	637	271	866	712	638	709	951	531
	(165)	(201)	(245)	(81)	(165)	(308)	(192)	(90)	(168)	(196)

Table 64 (continued).

Configuration	KdV	K–S	Bu_v	Bu_i	VdP	A–C	Wave	Comp	Osc	Rat
DeepSeek										
s	166	199	319	197	205	210	549	337	306	463
	(49)	(158)	(172)	(103)	(211)	(102)	(224)	(181)	(168)	(146)
p	238	172	118	436	228	245	75	71	256	119
	(285)	(204)	(126)	(211)	(125)	(112)	(35)	(33)	(103)	(43)
p+s	212	399	256	278	230	627	375	337	398	345
	(97)	(209)	(58)	(245)	(54)	(269)	(220)	(138)	(96)	(131)
s, 2c	389	362	531	155	460	578	589	468	509	443
	(155)	(65)	(211)	(106)	(357)	(266)	(353)	(149)	(302)	(211)
p, 2c	247	300	115	253	193	438	230	225	221	402
	(129)	(184)	(61)	(170)	(99)	(194)	(210)	(191)	(116)	(437)
p+s, 2c	372	318	341	221	419	325	425	531	788	290
	(219)	(106)	(149)	(75)	(368)	(115)	(73)	(274)	(339)	(63)
s +d	224	79	183	285	536	132	234	285	347	329
	(75)	(54)	(57)	(67)	(425)	(29)	(48)	(151)	(129)	(189)
p +d	263	401	138	330	276	164	128	129	221	195
	(163)	(194)	(50)	(135)	(116)	(187)	(51)	(93)	(60)	(232)
p+s +d	199	253	229	152	291	607	257	274	477	263
	(103)	(201)	(45)	(181)	(100)	(338)	(136)	(85)	(126)	(89)
s +d, 2c	572	231	410	222	1004	530	429	518	646	419
	(361)	(79)	(135)	(165)	(419)	(135)	(206)	(97)	(358)	(172)
p +d, 2c	512	276	213	563	225	320	360	290	374	255
	(267)	(127)	(102)	(280)	(87)	(73)	(166)	(149)	(316)	(57)
p+s +d, 2c	346	311	593	195	836	477	589	594	978	455
	(171)	(125)	(537)	(52)	(302)	(168)	(163)	(201)	(422)	(288)
Gemma										
s	41	68	36	56	24	24	24	43	24	20
	(24)	(55)	(16)	(67)	(5)	(6)	(15)	(31)	(3)	(10)
p	27	30	32	45	17	29	38	37	35	9
	(11)	(15)	(10)	(24)	(6)	(19)	(15)	(23)	(17)	(3)
p+s	62	51	42	35	32	41	50	19	44	23
	(10)	(22)	(28)	(18)	(27)	(18)	(29)	(4)	(32)	(10)
s, 2c	53	51	58	87	35	59	50	57	41	41
	(28)	(20)	(30)	(77)	(8)	(9)	(12)	(15)	(11)	(9)
p, 2c	60	98	66	50	42	57	66	55	44	138
	(35)	(58)	(31)	(36)	(20)	(34)	(39)	(35)	(25)	(370)
p+s, 2c	74	70	61	96	64	108	68	78	77	75
	(18)	(17)	(19)	(65)	(34)	(54)	(20)	(50)	(67)	(43)
s +d	56	77	47	57	24	29	35	48	27	32
	(37)	(60)	(21)	(39)	(4)	(6)	(23)	(6)	(7)	(7)
p +d	37	34	33	22	25	73	39	24	44	233
	(18)	(17)	(24)	(10)	(22)	(95)	(32)	(7)	(30)	(632)
p+s +d	168	101	147	303	59	139	175	109	31	85
	(57)	(97)	(54)	(492)	(41)	(70)	(84)	(38)	(17)	(53)
s +d, 2c	188	167	127	153	46	112	112	115	44	98
	(80)	(41)	(38)	(63)	(21)	(53)	(32)	(33)	(9)	(51)
p +d, 2c	72	73	49	58	54	57	59	82	66	185
	(34)	(57)	(21)	(30)	(24)	(19)	(24)	(63)	(31)	(451)
p+s +d, 2c	261	251	251	243	153	289	216	177	116	338
	(83)	(114)	(56)	(64)	(84)	(142)	(107)	(72)	(44)	(524)

Table 65: Cost in US cents per run, mean (standard deviation) over the 10 runs of each cell.

Configuration	KdV	K–S	Bu_v	Bu_i	VdP	A–C	Wave	Comp	Osc	Rat
GLM										
s	2.4 (1.4)	1.5 (0.6)	2.5 (2.0)	0.6 (0.3)	0.9 (0.3)	2.8 (1.3)	4.3 (2.4)	2.6 (2.0)	3.8 (2.0)	4.2 (1.9)
p	0.8 (0.5)	1.1 (0.6)	0.4 (0.1)	0.7 (0.3)	0.6 (0.4)	1.4 (1.1)	0.6 (0.3)	0.5 (0.2)	1.4 (0.9)	0.4 (0.2)
p+s	1.0 (0.5)	1.8 (1.2)	3.6 (1.7)	0.6 (0.4)	0.7 (0.3)	3.5 (1.4)	4.4 (2.0)	3.3 (2.7)	3.8 (2.3)	4.1 (2.6)
s, 2c	3.3 (1.5)	2.7 (0.9)	4.6 (1.9)	1.2 (0.3)	2.4 (0.5)	3.5 (2.0)	6.0 (1.4)	5.4 (2.2)	5.5 (2.7)	5.9 (2.5)
p, 2c	3.3 (1.5)	1.1 (0.8)	0.9 (0.6)	2.1 (1.0)	1.9 (1.4)	2.2 (0.9)	1.0 (0.4)	0.5 (0.2)	4.1 (1.9)	0.9 (0.6)
p+s, 2c	3.1 (2.0)	2.6 (1.4)	5.4 (1.9)	1.3 (0.6)	4.6 (2.5)	5.3 (1.6)	6.0 (1.7)	7.3 (2.0)	4.9 (2.3)	5.8 (2.9)
s +d	2.7 (1.1)	2.1 (0.7)	4.2 (1.6)	1.2 (1.0)	1.9 (1.4)	2.8 (1.3)	4.1 (2.4)	4.4 (1.8)	5.0 (0.9)	4.7 (0.7)
p +d	0.9 (0.3)	1.0 (0.4)	1.1 (0.7)	1.6 (1.1)	1.0 (0.4)	2.3 (0.9)	2.0 (1.8)	1.9 (1.4)	2.0 (0.7)	1.6 (0.7)
p+s +d	1.8 (0.3)	3.4 (1.2)	6.2 (1.4)	1.4 (0.9)	1.0 (0.5)	5.6 (1.7)	3.7 (1.3)	6.2 (1.7)	6.0 (1.9)	6.1 (1.8)
s +d, 2c	5.9 (1.9)	4.0 (1.5)	6.2 (2.8)	4.0 (1.0)	4.8 (1.8)	7.3 (2.3)	6.8 (3.2)	9.3 (5.4)	6.2 (2.3)	7.6 (2.5)
p +d, 2c	2.6 (1.8)	1.3 (0.8)	1.6 (0.8)	2.9 (1.6)	2.7 (1.5)	2.6 (1.9)	3.1 (2.0)	2.7 (1.9)	4.6 (2.5)	2.4 (1.0)
p+s +d, 2c	4.8 (1.6)	4.0 (1.3)	6.5 (3.0)	3.1 (0.9)	5.5 (1.5)	6.4 (7.0)	6.7 (2.9)	7.1 (1.8)	6.3 (1.9)	9.3 (5.4)
Qwen										
s	1.3 (0.6)	1.7 (0.6)	2.9 (1.1)	1.0 (0.2)	1.6 (0.4)	2.5 (0.6)	2.8 (0.9)	2.6 (0.7)	3.7 (0.9)	1.9 (0.5)
p	0.6 (0.2)	0.6 (0.2)	0.5 (0.2)	0.9 (0.3)	0.7 (0.2)	0.6 (0.1)	0.6 (0.2)	0.4 (0.1)	0.8 (0.2)	0.5 (0.1)
p+s	1.3 (0.5)	1.5 (0.5)	3.2 (0.6)	0.8 (0.2)	1.9 (0.6)	2.5 (0.7)	3.1 (0.5)	2.1 (0.6)	3.0 (0.8)	2.5 (0.7)
s, 2c	3.2 (0.7)	3.6 (1.1)	5.2 (0.5)	1.7 (0.6)	3.7 (0.9)	4.5 (0.6)	5.3 (0.9)	4.7 (0.6)	4.8 (0.4)	3.9 (1.0)
p, 2c	1.2 (0.5)	1.6 (0.4)	1.0 (0.3)	1.6 (0.6)	1.4 (0.3)	1.7 (0.5)	1.2 (0.3)	1.3 (0.5)	2.1 (0.6)	1.0 (0.3)
p+s, 2c	3.4 (0.8)	3.5 (0.7)	4.1 (0.7)	1.6 (0.2)	3.2 (0.7)	4.2 (0.8)	5.1 (0.6)	4.6 (0.9)	4.2 (0.5)	4.0 (0.4)
s +d	2.2 (0.7)	1.7 (0.7)	2.6 (0.9)	1.2 (0.4)	2.1 (0.8)	3.4 (1.5)	2.2 (1.4)	2.7 (0.6)	3.6 (0.8)	2.3 (0.7)
p +d	0.8 (0.4)	0.7 (0.4)	0.8 (0.4)	1.2 (0.6)	0.8 (0.3)	1.0 (0.3)	0.8 (0.4)	0.9 (0.3)	1.4 (1.0)	0.9 (0.3)
p+s +d	2.3 (1.0)	1.7 (1.1)	2.8 (1.2)	1.4 (0.8)	2.7 (0.8)	3.2 (1.4)	2.7 (1.0)	3.6 (0.9)	3.5 (0.6)	4.1 (0.7)
s +d, 2c	6.1 (1.1)	5.8 (1.0)	6.9 (2.2)	3.5 (0.9)	5.3 (1.0)	7.1 (1.3)	7.1 (0.9)	6.3 (1.2)	6.1 (1.0)	6.2 (0.9)
p +d, 2c	1.7 (0.6)	1.7 (0.5)	1.7 (0.4)	1.9 (0.5)	1.9 (0.5)	1.6 (0.6)	1.8 (0.4)	1.5 (0.7)	2.7 (0.8)	1.5 (0.3)
p+s +d, 2c	7.0 (1.2)	6.3 (1.2)	7.7 (1.3)	3.3 (0.7)	4.3 (1.0)	7.8 (0.8)	7.3 (0.8)	7.2 (0.6)	5.5 (0.6)	6.1 (1.3)

Table 65 (continued).

Configuration	KdV	K–S	Bu_v	Bu_i	VdP	A–C	Wave	Comp	Osc	Rat
DeepSeek										
s	2.7 (0.8)	3.2 (2.2)	3.8 (1.9)	1.4 (0.6)	2.1 (0.8)	4.0 (2.4)	5.1 (1.9)	4.5 (1.6)	4.1 (1.7)	3.1 (0.8)
p	3.2 (3.8)	2.4 (2.3)	0.7 (0.2)	1.4 (0.5)	1.0 (0.3)	1.2 (0.6)	1.0 (0.6)	0.8 (0.2)	1.5 (0.6)	1.7 (0.7)
p+s	3.1 (1.3)	2.1 (0.8)	4.2 (0.9)	3.3 (3.1)	3.5 (0.8)	3.0 (0.8)	4.1 (1.4)	5.6 (1.8)	6.2 (1.5)	5.6 (2.0)
s, 2c	4.7 (1.6)	5.1 (0.9)	5.5 (1.2)	1.9 (0.4)	3.6 (0.7)	7.0 (2.0)	5.9 (1.5)	5.8 (1.4)	6.0 (1.3)	6.5 (0.8)
p, 2c	2.4 (0.9)	2.8 (0.9)	1.1 (0.4)	2.5 (0.6)	1.9 (1.3)	5.4 (2.6)	2.0 (0.8)	1.9 (0.4)	2.4 (0.9)	2.3 (1.2)
p+s, 2c	4.2 (1.8)	4.5 (1.8)	4.8 (2.0)	2.2 (0.7)	3.2 (1.1)	6.0 (1.8)	4.9 (0.7)	6.4 (2.7)	5.6 (2.1)	4.9 (1.4)
s +d	4.2 (2.4)	2.8 (1.9)	4.4 (2.6)	1.9 (0.4)	4.3 (1.8)	5.6 (2.2)	4.2 (1.4)	4.2 (2.5)	5.0 (2.2)	3.7 (1.4)
p +d	1.5 (0.5)	1.2 (0.3)	2.1 (0.8)	3.7 (1.4)	1.3 (0.2)	2.5 (2.2)	1.8 (0.8)	2.1 (0.8)	3.3 (0.9)	2.2 (2.3)
p+s +d	3.8 (1.8)	3.7 (0.9)	4.3 (1.7)	2.8 (2.8)	4.6 (1.6)	3.9 (1.4)	3.8 (0.5)	7.1 (2.7)	7.0 (2.3)	4.9 (1.9)
s +d, 2c	9.5 (2.4)	9.2 (2.3)	8.3 (2.5)	4.6 (1.0)	6.6 (2.0)	13.4 (2.5)	8.3 (1.6)	7.7 (2.1)	7.5 (1.5)	9.1 (2.2)
p +d, 2c	2.5 (0.8)	2.1 (0.6)	2.6 (0.7)	3.5 (1.1)	2.7 (0.8)	3.7 (0.8)	3.0 (0.8)	2.8 (1.3)	3.0 (0.3)	3.0 (0.9)
p+s +d, 2c	7.9 (3.5)	7.4 (1.7)	7.5 (2.2)	4.3 (1.0)	6.7 (1.5)	11.9 (3.9)	7.5 (2.3)	9.8 (2.9)	6.6 (2.3)	7.6 (1.7)
Gemma										
s	0.3 (0.1)	0.5 (0.6)	0.3 (0.1)	0.2 (0.1)	0.1 (0.0)	0.2 (0.1)	0.2 (0.1)	0.2 (0.1)	0.1 (0.0)	0.5 (0.6)
p	0.0 (0.0)	0.1 (0.1)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.1 (0.1)	0.0 (0.0)	0.1 (0.1)
p+s	0.4 (0.0)	0.5 (0.6)	0.3 (0.0)	0.3 (0.0)	0.2 (0.2)	0.5 (0.6)	0.3 (0.1)	0.2 (0.0)	0.1 (0.0)	0.5 (0.6)
s, 2c	0.7 (0.6)	1.5 (1.3)	0.5 (0.1)	0.5 (0.1)	0.2 (0.0)	0.7 (0.6)	0.5 (0.1)	0.6 (0.6)	0.2 (0.0)	0.5 (0.1)
p, 2c	0.1 (0.0)	0.1 (0.1)	0.1 (0.0)	0.2 (0.1)	0.1 (0.0)	0.1 (0.1)	0.1 (0.0)	0.1 (0.0)	0.1 (0.1)	0.1 (0.1)
p+s, 2c	0.6 (0.1)	1.3 (0.9)	0.9 (0.8)	0.8 (0.6)	0.2 (0.0)	0.6 (0.0)	0.5 (0.0)	1.1 (1.2)	0.3 (0.2)	1.1 (0.9)
s +d	1.0 (0.4)	4.5 (4.7)	1.0 (0.4)	0.8 (0.2)	0.1 (0.0)	0.8 (0.2)	0.9 (0.4)	0.6 (0.1)	0.1 (0.0)	0.7 (0.1)
p +d	0.1 (0.0)	0.1 (0.1)	0.1 (0.0)	0.1 (0.0)	0.1 (0.1)	0.1 (0.0)	0.1 (0.1)	0.0 (0.0)	0.1 (0.0)	0.1 (0.0)
p+s +d	1.7 (0.4)	9.4 (3.8)	2.2 (3.2)	1.2 (0.2)	0.1 (0.0)	3.4 (4.1)	1.3 (0.0)	1.0 (0.1)	0.2 (0.0)	4.0 (5.9)
s +d, 2c	2.8 (0.3)	8.7 (6.9)	2.5 (0.4)	2.4 (0.5)	0.3 (0.0)	4.3 (4.2)	2.4 (0.4)	2.0 (0.2)	0.3 (0.0)	2.8 (2.6)
p +d, 2c	0.2 (0.1)	0.2 (0.1)	0.2 (0.1)	0.1 (0.0)	0.1 (0.0)	0.1 (0.0)	0.1 (0.1)	0.1 (0.1)	0.1 (0.1)	0.1 (0.1)
p+s +d, 2c	4.0 (3.0)	6.7 (6.9)	3.7 (3.1)	5.8 (4.4)	0.3 (0.0)	3.9 (3.1)	7.6 (4.6)	11.6 (5.8)	0.5 (0.4)	15.7 (5.8)