

InvariantSpec: Benchmarking LLM Autoformalization of Conservation and Dissipation Laws into Verifiable State-Space Constraints

Hikaru Matsuoka

University of Nagasaki

RIKEN AIP

matsuokah@sun.ac.jp

Abstract

Mathematical descriptions of dynamical systems denote *conjunctions* of proof obligations—the advection–diffusion equation conserves mass *and* dissipates the L^2 norm—yet common task designs ask language models to summarize a system with a single structure label, a target that is wrong by construction for mixed conservation–dissipation systems. We recast the problem as *autoformalization*: translating natural language, equations, and stated assumptions into a typed formal specification whose atoms are machine-checkable matrix conditions on the generator of a linear state-space model ($A^\top c = 0$, $A^\top M + MA = 0$, $A^\top M + MA \preceq 0$, $A^\top J + JA = 0$). We introduce **InvariantSpec**, a benchmark of 12 dynamical systems $\times$ 5 linguistic variants $\times$ 3 modality conditions (180 inputs) with verifier-executable gold specifications, contrastive minimal pairs (boundary conditions, conserved vs. non-increasing, sentence order, quadratic vs. cubic invariants), and two systems whose correct answer is *not representable by a fixed linear generator*. A symbolic verifier checks every predicted obligation and returns typed counterexample feedback; verified specifications are then *compiled* into constrained generators, including a new parameterization that simultaneously conserves a linear quantity and dissipates a quadratic one. Across four frontier LLMs, single-label structure selection has exact specification match 0.00—a ceiling imposed by the answer space, not the models—while autoformalization with one round of verifier-guided refinement reaches up to 0.88 with soundness 0.98–1.00; the residual gap is dominated by unstated side conditions (assumption extraction ranges 0.23–0.93), not by obligation algebra. Downstream, one-step MSE does not predict long-rollout structural fidelity: spec-compiled generators hold every verified invariant to machine precision over 1000-step rollouts, while unconstrained models with equal or better MSE violate them by up to 10^{17} .

1 Introduction

Consider the periodic advection–diffusion equation $u_t + cu_x = \kappa u_{xx}$. Its physics is a conjunction of two facts of opposite algebraic type: the total mass $\int u dx$ is *exactly conserved* (an equality), while the L^2 norm $\int u^2 dx$ is *dissipated* (an inequality). A task that asks a model to choose *one* structure class for this system—“unitary” (norm-preserving) or “contractive” (norm-decaying)—has no correct answer: whichever label the model picks discards half of the mathematics. In pilot studies with such a design we repeatedly saw models identify *both* properties in their rationale (“*the system requires a combination of conservation (mass) and dissipation (L^2 norm)...*”) and then commit to one label. This is not only a model failure; it is a *task-design failure*: collapsing a specification into one label makes the gold standard itself unsound.

This paper treats the problem as *autoformalization* (Wu et al., 2022): translating mathematical natural language, equations, and stated assumptions into a typed formal specification that a machine can verify. Our target formalism is deliberately lightweight. For a linear state-space model (SSM) $\dot{h} = Ah$ with discrete map $\Phi_\Delta = \exp(\Delta A)$ —the computational core of the S4/S5/Mamba family (Gu et al., 2022; Smith et al., 2023; Gu and Dao, 2024; Dao and Gu, 2024)—every conservation or dissipation statement reduces to a *matrix condition on the generator*: conservation of a linear quantity $c^\top h$ is $A^\top c = 0$; conservation of a quadratic quantity $\frac{1}{2}h^\top Mh$ is the Lyapunov equality $A^\top M + MA = 0$; dissipation is the semidefinite inequality $A^\top M + MA \preceq 0$; preservation of a symplectic two-form J is $A^\top J + JA = 0$. These conditions are decidable by elementary linear algebra, so the verifier returns not just *false* but a *counterexample witness*: a concrete state at which the claimed invariant measurably changes—the verifier-grounded contract of formal mathemat-

ics (Jiang et al., 2023; Azerbayev et al., 2023) at the applied end of mathematical language (Meadows and Freitas, 2023; Huang et al., 2025).

The formal target also sharpens a distinction natural language habitually blurs: *symplecticity and energy conservation are different obligations*. The plausible claim “ $A = JS$ with S symmetric preserves $\frac{1}{2}h^\top Mh$ whenever $J^{-1}M$ commutes with S ” is false: for canonical J , $S = I$, $M = \text{diag}(1, 2)$ the premise holds trivially, yet the verifier exhibits a state with $\frac{d}{dt} \frac{1}{2}h^\top Mh = -\frac{1}{2} \neq 0$ (Proposition 1). Informal plausibility is exactly what a proof-obligation checker is for.

Contributions. (1) A typed specification language whose atoms are machine-checkable proof obligations for conservation, dissipation, two-form preservation, required assumptions, and *representability*—whether a fixed linear generator can realize the dynamics at all (§2). (2) INVARIANTSPEC: 12 systems $\times$ 5 linguistic variants $\times$ 3 modalities with verifier-executable gold specifications and contrastive minimal pairs (boundary conditions, polarity, sentence order, representability) (§3). (3) A symbolic verifier with counterexample search and a typed error taxonomy, used for evaluation and for one round of verifier-guided refinement, plus a *constraint compiler* from verified specifications to constrained generators—including one that provably conserves a linear quantity while dissipating a quadratic one, replacing the unsound “unitary or contractive” choice with sound constraint composition (§4). (4) An evaluation of four frontier LLMs with soundness/completeness metrics and cluster-bootstrap intervals, and a downstream SSM case study showing that one-step MSE does not predict long-rollout structural fidelity, whereas compiled constraints hold invariants at machine precision (§5–§6).

2 Typed Formal Specifications

Setting. We work with autonomous linear dynamics $\dot{h} = Ah$ on $\mathbb{R}^n$ (or $\mathbb{C}^n$), with exact discrete flow $\Phi_\Delta = \exp(\Delta A)$. A *specification* is a conjunction of atomic *proof obligations*, each tying a named physical quantity to a matrix condition on A (Table 1), together with a list of *assumptions* (side conditions such as boundary conditions) and a *representability* judgment.

All obligations are grounded by one calculation.

Obligation	Condition on A
preserve_linear(c)	$A^\top c = 0$
preserve_quadratic(M)	$A^\top M + MA = 0$
dissipate_quadratic(M)	$A^\top M + MA \preceq 0$
preserve_two_form(J)	$A^\top J + JA = 0$
preserve_complex_norm	$A^* + A = 0$
requires(a)	side condition a
unsupported(r)	no fixed linear A exists

Table 1: The specification language. Quantities are $c^\top h$, $\frac{1}{2}h^\top Mh$ ($M=M^\top$), a nondegenerate skew form J , and the complex norm. Arguments are *canonical form names* (mass, 12_norm, energy, ...) resolved by the verifier against the system’s discretization: grading is by the mathematical object, not string identity.

Theorem 1 (Quadratic quantities under linear flow). *Let $M = M^\top$. Then for all t ,*

$$\frac{d}{dt} \left(e^{tA^\top} M e^{tA} \right) = e^{tA^\top} \left(A^\top M + MA \right) e^{tA}.$$

Consequently (i) $A^\top M + MA = 0 \iff \Phi_\Delta^\top M \Phi_\Delta = M$ for all Δ ; (ii) $A^\top M + MA \preceq 0 \implies \Phi_\Delta^\top M \Phi_\Delta \preceq M$ for all $\Delta \geq 0$; and (iii) $A^\top c = 0 \iff \Phi_\Delta^\top c = c$ for all Δ .

Proof. The identity is the product rule. For (ii), fix v and note $t \mapsto v^\top e^{tA^\top} M e^{tA} v$ has nonpositive derivative pointwise, hence is nonincreasing; (i) and (iii) follow by integrating and by $\frac{d}{dt} e^{tA^\top} c = e^{tA^\top} A^\top c$. $\square$

Theorem 1 is what makes the DSL *verifiable*: each continuous-time obligation transports to an exact certificate for the discrete map used by the trained model, with no scheme-dependent error term. Conservation and dissipation live in one framework but with different logical strength (equality vs. semidefinite inequality), which the evaluation keeps separate throughout.

Symplecticity is not energy conservation. Natural language freely slides between “symplectic”, “Hamiltonian”, and “energy-conserving”. The formal picture is sharper.

Proposition 1. *Let $J^\top = -J$, $J^2 = -I$, $S = S^\top$, and $A = JS$. Then $A^\top J + JA = 0$ always (the two-form is preserved), and for symmetric M ,*

$$A^\top M + MA = 0 \iff MJS = SJM.$$

In particular $M = S$ is always conserved, but a generic M is not: for canonical J in $\mathbb{R}^2$, $S = I$, $M = \text{diag}(1, 2)$ we get $A^\top M + MA = \begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix} \neq 0$, and at the witness state

$v = \frac{1}{\sqrt{2}}(1, 1)^\top$ the “invariant” changes at rate $\frac{d}{dt} \frac{1}{2} v^\top M v = -\frac{1}{2}$.

Proof. $A^\top J + J A = -S J J + J J S = S - S = 0$. For the equivalence, $A^\top M + M A = -S J M + M J S$, which vanishes iff $M J S = S J M$. The instance is a direct computation. $\square$

The correct condition couples all three of M , J , S ; the plausible weaker premise “ $J^{-1}M$ commutes with S ” holds trivially above ($S = I$) while the conclusion fails. The verifier refutes such claims constructively with the witness v —the message the refinement loop (§4) feeds back. `preserve_two_form` and `preserve_quadratic` are therefore separate atoms; a Hamiltonian system lists both.

Representability. A specification finally judges whether the dynamics is realizable by a fixed linear generator at all: `exact_linear`, `approximate_linear`, or `not_representable`. Euler’s rigid-body equations ($\dot{\Pi} = \Pi \times \Omega$: state-dependent generator) and KdV (quadratic advection, cubic Hamiltonian) are `not_representable`: the honest answer is to *refuse the surrogate class*, removing the artifact of single-label designs in which nonlinear systems must be force-mapped onto some class (KdV $\rightarrow$ “unitary”) to be gradable at all. Models emit specifications as schema-validated JSON; Appendix A shows the advection–diffusion gold specification.

3 The InvariantSpec Benchmark

Systems and gold specifications. Table 2 lists the 12 base systems: Hamiltonian, dissipative, and—crucially—*mixed* systems whose specification is a nontrivial conjunction, plus two nonlinear systems whose gold answer is `not_representable`. PDEs are discretized on periodic or Dirichlet grids (central differences; $n \leq 32$); each system exposes its named forms (mass vector, energy matrix, graph Laplacian, canonical J), and obligations are graded *semantically*: a prediction matches gold iff both resolve to numerically proportional matrices under the same condition type. An oracle answering with gold specifications scores 1.000 on every metric—the noise floor is zero.

Linguistic variants and modalities. Each system is described in five registers (textbook, applied/engineering, terse lecture-note, verbose narra-

tive, and a statement-style variant used by the minimal pairs), each under three modality conditions—*equation-only*, *text-only*, and *full* (equation + text + explicit assumptions)—yielding $12 \times 5 \times 3 = 180$ inputs. Every text variant uniquely determines the system, so gold is recoverable in every condition; what varies is whether structure must be derived from symbols, extracted from prose, or reconciled across both.

Minimal pairs. Four contrastive pairs isolate one linguistic dimension each, holding everything else fixed:

Boundary condition. The same heat equation with “periodic” vs. “homogeneous Dirichlet” boundary conditions: mass conservation must appear in the first specification and be absent from the second.

Polarity. “the L^2 norm ... is conserved as well” vs. “... is non-increasing” in otherwise matched sentences (advection vs. advection–diffusion): one content word flips an equality obligation to an inequality.

Sentence order. “Mass is conserved, while the L^2 energy dissipates” vs. “The L^2 energy dissipates, although the mass remains constant”—the same system, so the outputs must be formally equivalent (probing the primacy effect seen in single-label pilots).

Representability. Matched sentences “conserves a *quadratic* energy ... *linear* vector field” vs. “conserves a *cubic* energy ... *nonlinear* vector field” (anisotropic oscillator vs. KdV): the invariant’s degree decides `exact_linear` vs. `not_representable`.

A pair is scored *consistent* only if the contrast is resolved correctly on both sides.

4 Autoformalization with a Verifier in the Loop

4.1 Three task designs

Baseline A: single structure label. The model picks one of $\{\textit{symplectic}, \textit{unitary}, \textit{contractive}, \textit{unconstrained}, \textit{vanilla}\}$. To grade this baseline in the same currency as the others, each label is mapped to the specification it implies (`unitary` $\rightarrow$ `preserve_quadratic( $L^2$ )`; `contractive` $\rightarrow$ `dissipate_quadratic( $L^2$ )`; `symplectic` $\rightarrow$ `preserve_two_form( $J$ )`; `unconstrained/vanilla` $\rightarrow$ no obligations). This makes the design’s ceiling

System	Gold obligations	Assumptions	Repr.
harmonic oscillator	pres_quad(energy), pres_2form(J)	—	exact
anisotropic oscillator	pres_quad(energy, non-Euclidean), pres_2form(J)	—	exact
damped oscillator	diss_quad(energy)	—	exact
periodic advection	pres_lin(mass), pres_quad(L^2)	periodic BC	exact
periodic heat	pres_lin(mass), diss_quad(L^2)	periodic BC	exact
Dirichlet heat	diss_quad(L^2); mass <i>not</i> conserved	Dirichlet BC	exact
periodic advection–diffusion	pres_lin(mass), diss_quad(L^2)	periodic BC	exact
wave equation	pres_quad(energy), pres_lin(momentum), pres_2form(J)	periodic BC	exact
linear Schrödinger	pres_complex_norm	periodic BC	exact
graph consensus	pres_lin(mean), diss_quad(disagreement)	connected graph	exact
free rigid body	—	—	not repr.
KdV	—	—	not repr.

Table 2: The 12 base systems and their gold specifications. Six involve more than one simultaneous obligation; two require refusing the linear surrogate class altogether. Every gold specification is validated by the symbolic verifier against the system’s discretized generator (all obligations pass; all discrete certificates of Theorem 1 hold to $< 10^{-12}$).

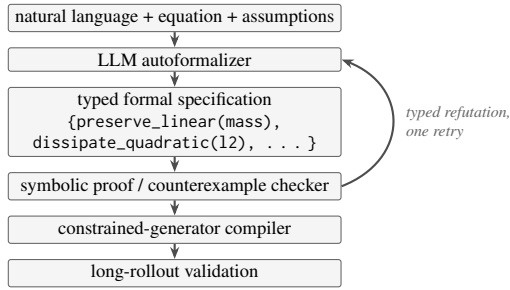


Figure 1: Pipeline. The LLM produces a typed specification; the verifier checks every obligation against the system’s discretized generator and either certifies it or returns a typed counterexample message for one refinement round; verified specifications compile into constrained generators whose long rollouts are validated numerically.

explicit: no label maps to a conjunction, so exact match on any multi-obligation system is impossible *regardless of model quality*.

Baseline B: direct autoformalization. The model emits the JSON specification directly. We use schema-constrained decoding (`response_format=json_schema`) whenever the provider supports it, so that interface failures are separated from semantic failures by construction rather than by parser forensics.

Proposed: autoformalization + verifier-guided refinement. The specification from Baseline B is checked by the verifier. On failure, the model receives a typed report and revises *once*. Example report on a polarity error: “[*polarity*] $A^\top M + MA = 0$ fails but $A^\top M + MA \preceq 0$ holds: the quadratic quantity ‘ L^2 norm’ is dissipated, not conserved (at the witness state, $\frac{d}{dt}Q =$

-3.1×10^{-1})—change preserve_quadratic to dissipate_quadratic.”

4.2 The verifier

Given the system’s discretized generator (mechanically derived from the equation—no gold labels consulted), the verifier resolves each named form, checks the matrix condition (equalities by residual norm, inequalities by extremal eigenvalue), and re-checks the discrete certificates of Theorem 1 on Φ_Δ ; small systems are re-checked in exact rational arithmetic (SymPy). Every failed check yields (i) a *witness state* at which the claimed invariant demonstrably changes and (ii) an error label from the taxonomy of §7: *quantity identification, polarity, algebraic typing, missing assumption, conjunction collapse*, or *representability*. The assumption-completeness check is *evaluation-only*: never included in refinement feedback, so the loop receives no annotation signal beyond the executable model of the system itself.

4.3 Metrics

Let $\hat{O}$ be the predicted and O^* the gold obligation set, matched semantically (§3). We report *soundness* = $|\{o \in \hat{O} : \text{verifier-true}\}|/|\hat{O}|$ and *completeness* = $|\{o \in O^* : \exists \hat{o} \text{ matching}\}|/|O^*|$, plus *exact specification match* (mutual matching, correct representability, all required assumptions stated), obligation-level macro- F_1 , assumption extraction accuracy, unsupported-case accuracy, minimal-pair consistency, verifier pass rate, and one-retry correction rate. Soundness is graded by the *verifier*, not by gold matching, so a true-but-unlisted obligation (e.g. L^2 conservation for the Schrödinger flow) is

never penalized as false. Paraphrases and modalities of one system are never treated as independent samples: we aggregate per base system and report 95% CIs from a cluster bootstrap over the 12 systems. Parse failures are tracked separately from all semantic metrics.

4.4 Compiling verified specifications

A verified specification is a conjunction of linear *equality* constraints and PSD *inequality* constraints on A , which we compile into a generator parameterization satisfying every obligation *for all parameter values*—structure is parameterized, not learned.

Proposition 2 (Constraint compilation). *Let $K^\top = -K$ and let R be arbitrary. (P1) If $M \succ 0$ and $A = M^{-1}K$ then $A^\top M + MA = 0$. (P2) If $M \succ 0$ and $A = M^{-1}(K - R^\top R)$ then $A^\top M + MA = -2R^\top R \preceq 0$. (P3) Let $P = I - \frac{cc^\top}{c^\top c}$ and $A = PKP - PR^\top RP$. Then $A^\top c = 0$ and $A^\top + A = -2PR^\top RP \preceq 0$: the linear quantity $c^\top h$ is conserved while $\|h\|^2$ is dissipated, simultaneously. (P3') More generally, for symmetric $G \succeq 0$, $A = (PKP - PR^\top RP)G$ conserves $c^\top h$ and dissipates $h^\top Gh$.*

Proof. (P1) $A^\top M + MA = K^\top M^{-1}M + MM^{-1}K = K^\top + K = 0$. (P2) identical, leaving $-2R^\top R$. (P3) $Pc = 0$ gives $A^\top c = 0$; the symmetric part of A is $-PR^\top RP$. (P3') $A^\top c = G(\cdots)^\top c = 0$ and $A^\top G + GA = G(B + B^\top)G = -2(RPG)^\top (RPG)$ with $B = PKP - PR^\top RP$. $\square$

P3 directly serves the mixed systems (heat, advection–diffusion, consensus) that single-label designs cannot express: mass is conserved *and* the L^2 norm is non-increasing, by construction. Pure-equality conjunctions (e.g. two-form *and* non-Euclidean energy) are compiled generically: each equality obligation is linear in A , so a basis of the joint nullspace is computed once by SVD and A is parameterized inside it; property-based tests confirm every compiled class satisfies its obligations for random parameters and contains the true generator. Specifications judged `not_representable` refuse compilation—the honest outcome for KdV and the rigid body.

5 LLM Experiments

Models and decoding. We evaluate four frontier models through a single API gateway: GPT-5.6-terra, Claude Sonnet 5, Gemini 3.6 Flash, and

DeepSeek V3.2 (temperature 0, low reasoning effort, schema-constrained JSON decoding). Each model answers all 180 inputs under all three task designs of §4.1; the refinement round is triggered only when the verifier rejects the first specification. Interface failures that survive schema-constrained decoding (provider errors, truncation) are reported as parse failures and excluded from semantic metrics; they are rare ($\leq 2.8\%$, zero for most cells).

Single-label designs have a semantic ceiling.

Table 3 quantifies the task-design failure that motivated this work. Graded through the specifications they imply, the single-label baselines reach soundness 0.73–0.76 but completeness only 0.39–0.40: a label can assert one true obligation while necessarily dropping the rest of the conjunction, and exact specification match is 0.000 for *every* model—not because the models are weak, but because the answer space cannot express the gold standard. The label space also cannot express *refusal*: unsupported-case accuracy is 0.00 across the board, since every available label asserts some structure for systems that have no linear realization.

Direct autoformalization recovers the conjunction.

Under the typed DSL, exact specification match rises to 0.55–0.85 (best: Gemini 3.6 Flash), soundness to 0.95–0.98, and completeness to 0.94–0.98. The verifier already certifies 91–96% of first-attempt specifications, and *every* model attains perfect unsupported-case accuracy: given a vocabulary for refusing, frontier models reliably recognize that Euler’s equations and KdV admit no fixed linear generator—the same systems that single-label designs force onto a least-wrong class.

Verifier refinement makes assertions true.

One round of typed counterexample feedback drives soundness to 0.98–1.00: after refinement, essentially every obligation asserted is *true*, and the *Fix* column shows that 70–100% of the few first-attempt verifier failures (7–10 per model out of 180) are fully repaired by a single retry. Exact match improves for three models (best 0.88, Gemini 3.6 Flash). The exception is instructive: GPT-5.6-terra’s exact match *decreases* slightly under refinement because rewriting occasionally drops assumption fields—which our no-leak design deliberately excludes from verifier feedback (§4.2)—trading a verifier-visible improvement for a verifier-invisible regression.

Model	Method	PF%	Exact match [95% CI]	Sound.	Compl.	mF ₁	Asm.	Unsup.	VPass	Fix
GPT-5.6-terra	single label	0.0	0.000 [0.00,0.00]	0.761	0.403	0.653	0.00	0.00	0.76	–
	direct	0.0	0.550 [0.35,0.76]	0.975	0.981	0.937	0.26	1.00	0.95	–
	+verifier	0.0	0.533 [0.34,0.75]	1.000	0.991	0.957	0.23	1.00	1.00	1.00
Claude Sonnet 5	single label	0.0	0.000 [0.00,0.00]	0.750	0.403	0.653	0.00	0.00	0.75	–
	direct	0.0	0.744 [0.58,0.89]	0.978	0.979	0.954	0.64	1.00	0.96	–
	+verifier	0.0	0.767 [0.63,0.90]	0.997	0.978	0.968	0.69	1.00	0.99	0.86
Gemini 3.6 Flash	single label	0.0	0.000 [0.00,0.00]	0.728	0.394	0.641	0.00	0.00	0.73	–
	direct	0.0	0.850 [0.73,0.96]	0.968	0.975	0.948	0.93	1.00	0.94	–
	+verifier	0.0	0.883 [0.78,0.97]	1.000	0.989	0.970	0.93	1.00	1.00	1.00
DeepSeek V3.2	single label	0.0	0.000 [0.00,0.00]	0.761	0.389	0.641	0.00	0.00	0.76	–
	direct	0.6	0.763 [0.63,0.89]	0.947	0.938	0.903	0.84	1.00	0.91	–
	+verifier	2.8	0.796 [0.69,0.90]	0.980	0.936	0.930	0.84	1.00	0.98	0.70

Table 3: Main results (180 inputs per cell; means of per-system means; cluster-bootstrap CIs over the 12 systems). PF = parse/interface failures (excluded from semantic metrics); mF₁ = obligation macro-F₁; Asm. = assumption extraction; Unsup. = accuracy on the two non-representable systems; VPass = verifier pass rate; Fix = failed first attempts corrected by one verifier round. Single-label rows are graded through the specification each label implies; their exact match is 0 *by construction* on multi-obligation systems.

Model	Exact match by modality			Minimal pairs (consistent?)			
	eq. only	text only	full	BC	polarity	order	repr.
GPT-5.6-terra	0.40	0.47	0.73	✓	✓	✓	✓
Claude Sonnet 5	0.75	0.63	0.92	✓	✓	✓	✓
Gemini 3.6 Flash	0.83	0.87	0.95	✓	✓	✓	✓
DeepSeek V3.2	0.71	0.73	0.95	✓	✓	✓	×

Table 4: Modality ablation and minimal-pair consistency for the verifier-refined method. A pair counts as consistent only if the contrast is resolved correctly on both sides (§3).

Assumptions, not algebra, separate the models.

The *Asm.* column exposes the largest capability gap: assumption extraction ranges from 0.93 (Gemini) down to 0.23 (GPT-5.6-terra), while obligation-level metrics are tightly clustered. For GPT-5.6-terra, 61 of 84 non-exact refined items have *perfect obligations* and fail only by omitting the boundary-condition side conditions their own specification depends on. Formalizing the mathematics is largely solved for these systems; formalizing its *context of validity* is not.

Modalities and minimal pairs. Table 4 separates reading mathematics from reading about mathematics. The full condition is best for every model (0.73–0.95 exact); equation-only is generally hardest (GPT-5.6-terra drops to 0.40), since invariants must be derived rather than extracted—e.g. noticing that the Dirichlet Laplacian does not annihilate the ones vector. With typed specifications, the minimal pairs are resolved almost uniformly: all models pass the boundary-condition, polarity, and sentence-order contrasts, and all pass representability under direct generation (DeepSeek regresses on it after refinement). Under the single-label design the boundary-condition and representability contrasts are *structurally unanswerable*

and every model fails them: the contrastive failures the old design exhibited were largely artifacts of its answer space. The pairs still matter—they pinpoint the one-token distinctions (*conserved* vs. *non-increasing*; *periodic* vs. *Dirichlet*) that decide between equality and inequality constraints downstream.

6 Downstream Case Study: Compiled Constraints in Trained SSMs

The specification is the contract a downstream surrogate must honor. We close the loop by training linear SSMs on the ten linear benchmark systems and measuring whether declared structure survives training and long rollouts. For each system, teacher trajectories come from the exact flow (32 initial conditions, 80 steps, $\Delta = 0.05$); three parameterizations train on one-step MSE (Adam, 200 epochs, float64): *unconstrained A*; *vanilla A* = $-\text{diag}(\exp \theta)$, the standard stability-first SSM rule; and *spec*, the gold specification compiled by Proposition 2. Rollouts run 1000 steps from 8 held-out initial conditions.

Obligation-appropriate error metrics. Conservation and dissipation must not share a “drift” number: drift is meaningless for a quantity that is *sup-*

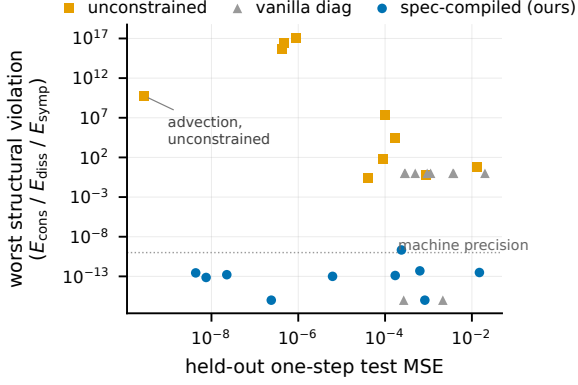


Figure 2: One-step test MSE does not predict long-rollout structural fidelity. Each point is one (system, parameterization); y is the worst violated obligation metric over a 1000-step rollout. Unconstrained models occupy the upper region at every MSE level (up to 1.2×10^{17}); spec-compiled models sit at machine precision across the entire MSE range.

posed to decrease. For a conserved quantity I , a dissipated quantity Q , and a two-form J we report

$$E_{\text{cons}} = \max_{t \leq T} \frac{|I(h_t) - I(h_0)|}{|I(h_0)| + \varepsilon},$$

$$E_{\text{diss}} = \frac{\sum_{t < T} [Q(h_{t+1}) - Q(h_t)]_+}{Q(h_0) + \varepsilon},$$

$$E_{\text{symp}} = \frac{\|\Phi^\top J \Phi - J\|_F}{\|J\|_F},$$

i.e. worst-case relative drift for equalities, accumulated *increase* for inequalities (decrease is physics, not error), and the discrete two-form defect of the learned map itself.

Results. Table 5 and Figures 2–3 give three findings. **(i) One-step MSE does not predict structure:** on periodic advection the unconstrained model attains the study’s best test MSE (2.8×10^{-10}) yet drifts mass by $\sim 10^5$ and L^2 by $\sim 10^9$ over 1000 steps; on periodic heat it reaches 4.6×10^{-7} MSE while *increasing* the supposedly dissipated L^2 norm by 2.5×10^{16} . **(ii) Compiled conjunctions cost little and can help:** spec-compiled models hold every equality obligation to $\leq 2.1 \times 10^{-10}$ and never violate an inequality ($E_{\text{diss}} = 0$), with test MSE within a small factor of unconstrained—on diffusive systems *better* (heat: 4.3×10^{-9} vs. 4.6×10^{-7}), because the constraint class excludes spurious expanding directions that one-step fitting cannot rule out. **(iii) Mixed conservation–dissipation is realized exactly:** Figure 3 shows both obligations of

advection–diffusion holding simultaneously—the regime the unitary/contractive dichotomy cannot express.

7 Error Analysis

Because every failed check carries a typed verifier label, error analysis is a count, not a reading exercise. Across all four models’ *direct* outputs (719 parseable specifications), the verifier and gold comparison yield 179 error events, with a distribution that inverts the usual expectations:

Missing assumption (119). By far the dominant class: obligations correct but the boundary-condition side condition unstated (67 events for GPT-5.6-terra alone, 6 for Gemini). Invisible to a free-form grader; caught by the typed field.

Algebraic typing (45). Quantities that do not type-check in the DSL. The recurrent pattern (18 events on Dirichlet heat) is physically astute: models want to assert that the *mass decays*, but a linear functional’s decay admits no semidefinite certificate, so the DSL rejects it. These errors mark a real expressiveness boundary of the formalism (see Limitations), not model confusion; the verifier’s message explains the boundary and suggests the certifiable alternative.

Conjunction collapse (10). Exactly one obligation of a multi-obligation gold specification emitted—the single-label habit persisting into the richer format, now rare rather than forced.

Quantity identification (3) and polarity (1).

Nearly extinct: models almost never claim a false invariant or confuse conserved with dissipated *when the answer space lets them assert both*; the polarity failures seen in pilot studies were overwhelmingly collapse artifacts.

Representability (1). One hallucination in the full run—a model declaring a frictionless mass–spring system `not_representable` because the spring “could be nonlinear”; the two-sided check (refusing linear systems is also an error) remains a necessary guard.

In aggregate: frontier models formalize the *algebra* of these systems nearly perfectly once the task is posed as autoformalization; residual failures concentrate in *pragmatics* (the context of validity) and in the *formalism’s expressiveness limits*—exactly

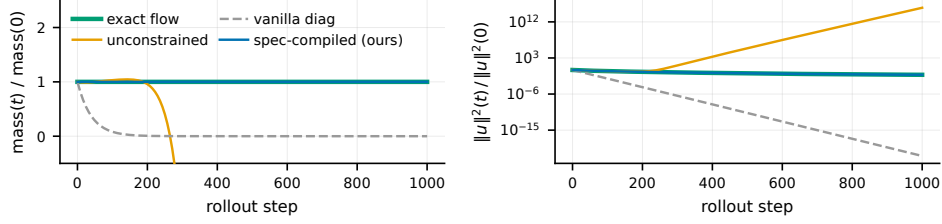


Figure 3: The two obligations of periodic advection–diffusion along a rollout. **Left:** mass ratio (equality); **right:** L^2 ratio, log scale (inequality). The spec-compiled generator (P3) tracks the exact flow; the unconstrained model loses mass catastrophically and grows L^2 without bound; the “stable” vanilla rule destroys both.

System	unconstrained		vanilla diag		spec-compiled (ours)	
	test MSE	worst viol.	test MSE	worst viol.	test MSE	worst viol.
harmonic osc.	4.0×10^{-5}	2.6×10^{-1}	1.1×10^{-3}	1.0×10^0	2.4×10^{-4}	2.1×10^{-10}
anisotropic osc.	1.3×10^{-2}	7.0×10^0	2.0×10^{-2}	1.0×10^0	1.5×10^{-2}	3.1×10^{-13}
damped osc.	8.9×10^{-4}	5.8×10^{-1}	2.1×10^{-3}	0	8.2×10^{-4}	0
advection (per.)	2.8×10^{-10}	5.8×10^9	9.3×10^{-4}	1.0×10^0	2.3×10^{-8}	1.6×10^{-13}
heat (per.)	4.6×10^{-7}	2.5×10^{16}	2.9×10^{-4}	1.0×10^0	4.3×10^{-9}	2.8×10^{-13}
heat (Dir.)	9.0×10^{-7}	1.2×10^{17}	2.7×10^{-4}	0	2.4×10^{-7}	0
adv.-diff. (per.)	4.1×10^{-7}	5.2×10^{15}	4.9×10^{-4}	1.0×10^0	7.5×10^{-9}	7.3×10^{-14}
wave (per.)	1.7×10^{-4}	2.5×10^4	3.8×10^{-3}	1.0×10^0	6.3×10^{-4}	5.0×10^{-13}
Schrödinger	1.0×10^{-4}	2.3×10^7	3.7×10^{-3}	1.0×10^0	1.7×10^{-4}	1.3×10^{-13}
consensus	9.1×10^{-5}	6.2×10^1	2.8×10^{-4}	1.0×10^0	6.2×10^{-6}	1.0×10^{-13}

Table 5: Downstream case study: held-out one-step test MSE vs. worst violated obligation metric ($E_{\text{cons}}/E_{\text{diss}}/E_{\text{symp}}$) over a 1000-step rollout. “0”: no dissipation obligation ever violated ($E_{\text{diss}} = 0$ exactly).

the classes a typed, verifier-grounded design makes attributable.

8 Related Work

Autoformalization and verifier-grounded mathematics. Translating informal mathematics into machine-checkable form spans proof autoformalization (Wu et al., 2022; Azerbayev et al., 2023), sketch-then-verify proving (Jiang et al., 2023), benchmarks such as miniF2F (Zheng et al., 2022), and mathematical language processing broadly (Meadows and Freitas, 2023). Our target language is far lighter than a proof assistant’s, but the contract is the same: outputs are graded by a checker, not by string overlap. Restricting to quadratic/linear invariants of linear flows buys complete decidability, counterexample witnesses, and—unusual for autoformalization—a *compiler* from verified statements to learnable model classes. Solver-verified modeling (Huang et al., 2025) and self-correcting simulation (Park et al., 2026) share the refinement pattern; we add typed obligations and obligation-level metrics.

Structure preservation in numerics and learning.

That conservation laws are properties of the discrete operator’s algebraic form, not of an optimization target, is classical: discrete variational deriva-

tive methods (Furihata, 1999; Furihata and Matsuo, 2010; Yaguchi et al., 2010), variational integrators (Marsden and West, 2001), and their neural inheritors (Greydanus et al., 2019; Lutter et al., 2019; Chen et al., 2020; Matsubara and Yaguchi, 2023; Tanaka et al., 2025; Khosrovian et al., 2025); soft-constrained alternatives fail in characterizable ways (Raissi et al., 2019; Krishnapriyan et al., 2021; Donti et al., 2021). For the S4/S5/Mamba family (Gu et al., 2022; Smith et al., 2023; Gu and Dao, 2024; Dao and Gu, 2024) the generator is an explicit matrix, which is what makes specification-to-parameterization compilation exact. Our contribution is the *constraint-composition* view: obligations, not classes, as the unit of structure, including simultaneous conservation–dissipation (Proposition 2).

9 Conclusion

Mathematical text about dynamical systems denotes conjunctions of proof obligations, not single structure labels. We made that observation operational—a typed specification language with decidable atoms, an executable gold standard, a verifier that refutes with witnesses and refines with typed feedback, and a compiler from verified specifications to generators that satisfy them for every

parameter value. Upstream, frontier LLMs recover most of the conjunction once the answer space can express it, and one verifier round makes their assertions essentially always true; downstream, one-step accuracy says nothing about thousand-step structure, while compiled obligations survive by construction. We hope INVARIANTSPEC serves as a small, fully checkable testbed at the interface of mathematical language and verified computation.

Limitations

Our systems are low-dimensional, autonomous, and linear (by design: the two nonlinear systems exist to be refused). Selective-scan, input-dependent SSM parameterizations are outside the present formal target; extending the obligation language to time-varying generators $A(x_t)$ —where the natural conditions become pointwise or averaged—is the main open direction. The benchmark’s 12 systems support cluster-bootstrap intervals but not fine-grained per-family comparisons, and its linguistic variants are author-written rather than corpus-mined, so lexical diversity is bounded by design choices we controlled for (register, order, polarity) rather than by natural variation. The DSL certifies conservation of linear quantities and conservation/dissipation of quadratic ones, but cannot certify decay of a linear functional (there is no semidefinite witness for $|e^\top h|$ shrinking); models bump into this boundary on Dirichlet systems (§7), and extending the language with weighted or squared surrogates is straightforward future work. The refinement loop requires the system’s executable discretization, which exists here by construction but must be synthesized (e.g. from the equation string) in fully open-ended settings; the loop runs for a single round with a single verifier, the minimal instance of verifier-guided repair. Finally, the evaluated models are accessed through a commercial gateway whose serving configuration we do not control, so absolute numbers should be read as a snapshot.

References

- Zhangir Azerbayev, Bartosz Piotrowski, Hailey Schoelkopf, Edward W. Ayers, Dragomir Radev, and Jeremy Avigad. 2023. ProofNet: Autoformalizing and formally proving undergraduate-level mathematics. *arXiv preprint arXiv:2302.12433*.
- Zhengdao Chen, Jianyu Zhang, Martin Arjovsky, and Léon Bottou. 2020. Symplectic recurrent neural networks. In *International Conference on Learning Representations (ICLR)*.
- Tri Dao and Albert Gu. 2024. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In *International Conference on Machine Learning (ICML)*.
- Priya L. Donti, David Rolnick, and J. Zico Kolter. 2021. DC3: A learning method for optimization with hard constraints. In *International Conference on Learning Representations (ICLR)*.
- Daisuke Furihata. 1999. Finite difference schemes for $\partial u/\partial t = (\partial/\partial x)^\alpha \delta G/\delta u$ that inherit energy conservation or dissipation property. *Journal of Computational Physics*, 156(1):181–205.
- Daisuke Furihata and Takayasu Matsuo. 2010. *Discrete Variational Derivative Method: A Structure-Preserving Numerical Method for Partial Differential Equations*. Numerical Analysis and Scientific Computing Series. Chapman & Hall/CRC.
- Samuel Greydanus, Misko Dzamba, and Jason Yosinski. 2019. Hamiltonian neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Albert Gu and Tri Dao. 2024. Mamba: Linear-time sequence modeling with selective state spaces. In *Conference on Language Modeling (COLM)*.
- Albert Gu, Karan Goel, and Christopher Ré. 2022. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations (ICLR)*.
- Xuhan Huang, Qingning Shen, Yan Hu, Anningzhe Gao, and Benyou Wang. 2025. LLMs for mathematical modeling: Towards bridging the gap between natural and mathematical languages. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 2678–2710.
- Albert Qiaochu Jiang, Sean Welleck, Jin Peng Zhou, Wenda Li, Jiacheng Liu, Mateja Jamnik, Timothée Lacroix, Yuhuai Wu, and Guillaume Lample. 2023. Draft, sketch, and prove: Guiding formal theorem provers with informal proofs. In *International Conference on Learning Representations (ICLR)*.
- Razmik Arman Khosrovian, Takaharu Yaguchi, Hiroaki Yoshimura, and Takashi Matsubara. 2025. Poisson-Dirac neural networks for modeling coupled dynamical systems across domains. In *International Conference on Learning Representations (ICLR)*.
- Aditi S. Krishnapriyan, Amir Gholami, Shandian Zhe, Robert M. Kirby, and Michael W. Mahoney. 2021. Characterizing possible failure modes in physics-informed neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Michael Lutter, Christian Ritter, and Jan Peters. 2019. Deep Lagrangian networks: Using physics as model prior for deep learning. In *International Conference on Learning Representations (ICLR)*.

- Jerrold E. Marsden and Matthew West. 2001. [Discrete mechanics and variational integrators](#). *Acta Numerica*, 10:357–514.
- Takashi Matsubara and Takaharu Yaguchi. 2023. FINDE: Neural differential equations for finding and preserving invariant quantities. In *International Conference on Learning Representations (ICLR)*.
- Jordan Meadows and André Freitas. 2023. [Introduction to mathematical language processing: Informal proofs, word problems, and supporting tasks](#). *Transactions of the Association for Computational Linguistics*, 11:1162–1184.
- Donggeun Park, Hyeonbin Moon, and Seunghwa Ryu. 2026. [A self-correcting multi-agent LLM framework for language-based physics simulation and explanation](#). *npj Artificial Intelligence*, 2:10.
- Maziar Raissi, Paris Perdikaris, and George E. Karniadakis. 2019. [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.
- Jimmy T. H. Smith, Andrew Warrington, and Scott W. Linderman. 2023. Simplified state space layers for sequence modeling. In *International Conference on Learning Representations (ICLR)*.
- Yusuke Tanaka, Takaharu Yaguchi, Tomoharu Iwata, and Naonori Ueda. 2025. Energy-consistent neural operators for Hamiltonian and dissipative partial differential equations. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*.
- Yuhuai Wu, Albert Qiaochu Jiang, Wenda Li, Markus Rabe, Charles Staats, Mateja Jamnik, and Christian Szegedy. 2022. Autoformalization with large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Takaharu Yaguchi, Takayasu Matsuo, and Masaaki Sugihara. 2010. [An extension of the discrete variational method to nonuniform grids](#). *Journal of Computational Physics*, 229(11):4382–4423.
- Kunhao Zheng, Jesse Michael Han, and Stanislas Polu. 2022. miniF2F: A cross-system benchmark for formal olympiad-level mathematics. In *International Conference on Learning Representations (ICLR)*.

A The Exchange Format

Gold specification of periodic advection–diffusion, in the JSON exchange format produced by the models (schema-constrained):

```
{
  "quantities": [
    {
      "name": "mass", "type": "linear",
      "behavior": "conserved", "form": "mass",
    },
    {
      "name": "L2 norm", "type": "quadratic",
      "behavior": "dissipated", "form": "L2_norm"
    }
  ],
  "assumptions": ["periodic boundary conditions"],
  "proof_obligations": ["A^T c = 0", "A^T + A <= 0"],
  "representability": "exact_linear"
}
```

The verifier resolves form names against the system’s discretization ($c = \mathbf{1}$, $M = I$ on the grid), checks $A^T c = 0$ and $A^T + A \preceq 0$, and transports both to the discrete map ($\Phi^T c = c$, $\Phi^T \Phi \preceq I$, Theorem 1).

B Verifier Details

Equalities are accepted at relative residual $< 10^{-10}$, inequalities at $\lambda_{\max} \leq 10^{-10}$ (scaled); all discrete certificates hold at $< 10^{-12}$ for the gold specifications. For failed equalities the witness is the extremal eigenvector v of the symmetrized residual $\text{sym}(A^T M + M A)$, reported with $\frac{d}{dt} \frac{1}{2} v^T M v$; for failed inequalities it is the eigenvector of the largest positive eigenvalue. Systems of dimension ≤ 8 with near-rational entries are re-checked in exact arithmetic (SymPy); the float and exact verdicts agreed in every instance.

C Training Details

Teacher trajectories use the exact flow $\Phi_\Delta = \exp(\Delta A)$ (float64), 32 training / 8 held-out initial conditions, 80 steps, $\Delta = 0.05$; smooth random Fourier fields (modes ≤ 4) for PDE systems, standard normal states otherwise. All parameterizations train with Adam (lr 5×10^{-3} , 200 epochs, full batch) on one-step MSE through `matrix_exp`. Rollouts are 1000 steps. State dimensions: oscillators 2–4, graph consensus 8, PDE systems 24 (12×2 for wave/Schrödinger). The LLM evaluation uses temperature 0, low reasoning effort, `max_tokens=4096`, and OpenRouter structured outputs; the complete evaluation (4 models $\times$ 3 methods $\times$ 180 inputs, plus refinement calls) cost under \$15.

D Additional Results

Figures 4 and 5 disaggregate Table 3 by method and by base system.

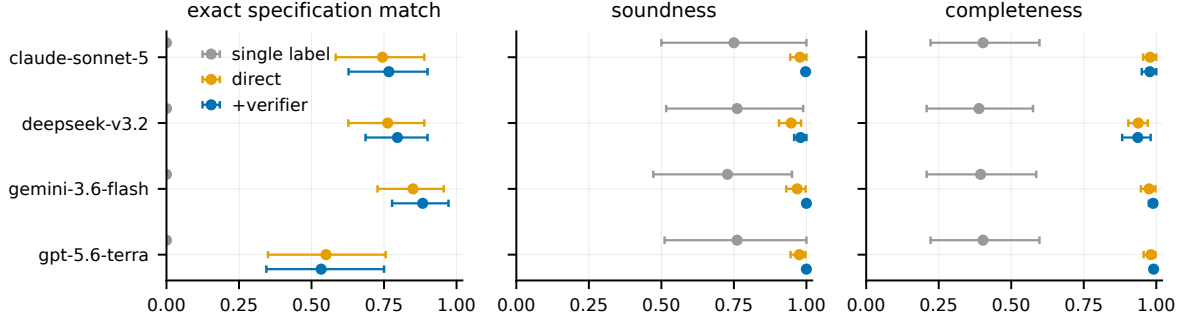


Figure 4: Exact specification match, soundness, and completeness per model and method, with cluster-bootstrap 95% CIs (gray: single label; orange: direct; blue: +verifier). The single-label baseline’s exact match is pinned at 0 by its answer space; soundness and completeness show the same ceiling pattern across all four models.

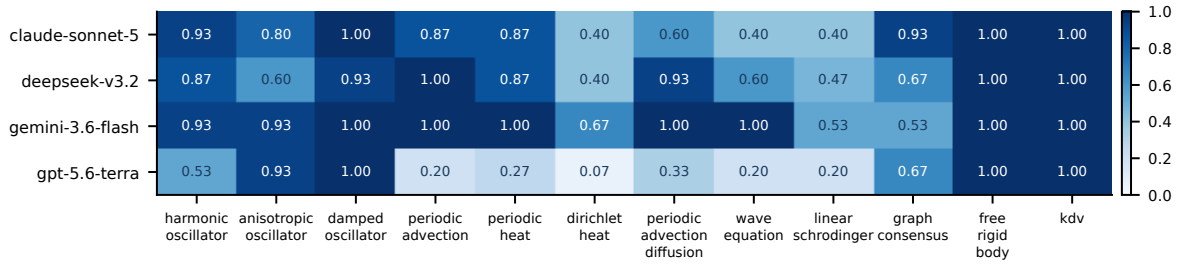


Figure 5: Exact specification match per base system (verifier-refined method; average over 5 variants $\times$ 3 modalities; parse failures count as 0). The hardest cells are systems whose specification validity depends on stated side conditions (periodic/Dirichlet boundary systems) rather than on obligation algebra.

E Benchmark Construction

Each system’s five text variants are written to be system-identifying under text-only presentation (register varies: textbook, applied, terse, narrative, statement-style). The oracle control (§3) feeds gold specifications through the full scoring pipeline and attains 1.000 on every metric, confirming that fingerprinting, semantic matching, and aggregation introduce no noise floor. A seeded noisy oracle with programmed corruption rates (collapse, polarity, dropped assumptions, representability flips) reproduces the intended metric decreases, validating metric sensitivity.