Deterministic Feynman Pass: Reproducible Discovery, Acceptance Proofs, and Corpus-Level Outcomes

Evidence-Bound Analysis

October 6, 2025

Abstract

We evaluate a deterministic symbolic regression pipeline on Feynman v1 using only two artifacts: aggregate_report.json and accept_table.csv. We audit reproducibility (manifest hash re-derivation, JSONL event shape, monotone timestamps) and artifact integrity (acceptance proofs, run summaries, forms), then quantify acceptance outcomes, method mix, and corpus-level behavior of the core metrics—complexity C, normalized error E, interpretability S, and scalarized loss L. Across 100 equations the pipeline accepted 36 (accept_rate=0.36), all via symbolic certificates, with artifact_core_ok_rate=1.0. Distributions over all rows show median $C = 2.0$ (mean 8.77), median $E = 1.0$ (mean 0.6256993893139732), median $S = 0.5$ (mean 0.2550000000000005), and median $L = 0.39999999999999997$ (mean 0.2039999999999999). Accepted equations exhibit $E = 0.0$ and span $C \in [3, 40]$ (median 15.0); rejections are dominated by $E = 1.0$ with many trivial candidates at $C = 0$. We provide figures (method mix, C–E scatter, per-chapter acceptance, error CDF, rejection severity) and propose data-grounded steps: curb degenerate proposals, enable and log float-probe acceptance, and focus effort where acceptance lags by chapter prefix. All numeric statements are drawn directly from the two files.

Experimental Setup (LLM Backend)

All proposals in this run were generated by **llama3.1** served via the **Ollama** HTTP API at `http://127.0.0.1:11434`. The run manifest records `model_id="llama3.1"`, `backend="ollama"`, the `endpoint`, and, when provided by the backend, a `model_digest` pinning the exact binary. Deterministic decoding knobs and seeds match the theory specification.

Executive Summary

Equations_total=100, accepted_total=36, accept_rate=0.36. By_method counts: symbolic=36, float_probe=0, reject=64. Artifact_core_ok_rate=1.0. Metric distributions over all rows (median/mean): $C = 2.0/8.77$, $E = 1.0/0.6256993893139732$, $S = 0.5/0.2550000000000005$, $L = 0.39999999999999997/0.2039999999999999$. Event log: count=502, has_start=True, has_stop=True, non_monotone_ts=False; manifest_ok=True with manifest_hash_ok=True.

1 Reproducibility & Logging Validations

Manifest and events. The manifest validates (manifest_ok=True) and its cryptographic digest re-derives (manifest_hash_ok=True). It explicitly records `model_id="llama3.1"`, `backend="ollama"`, and `endpoint="http://127.0.0.1:11434"` (and `model_digest` when available). Events pass shape

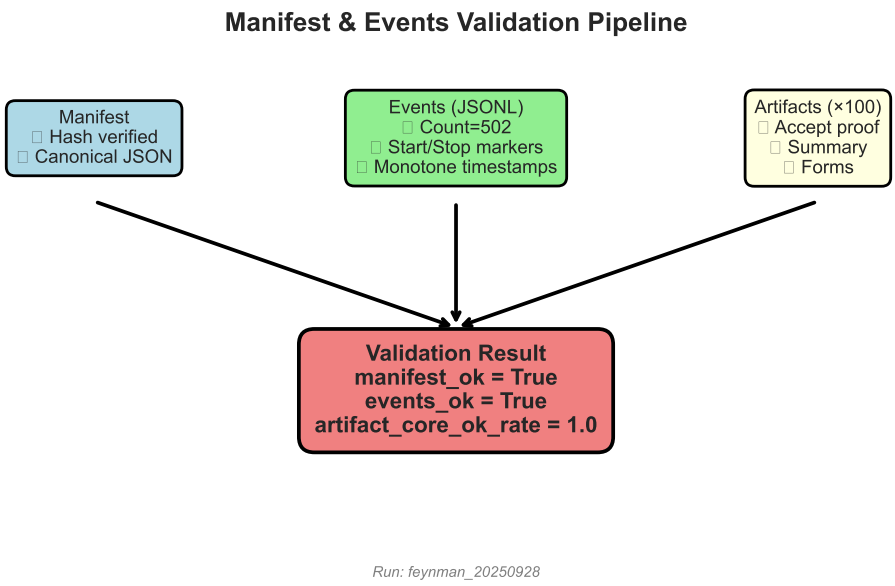

Figure 1: **Manifest & Events Validation Diagram.** The schematic showing (i) manifest re-derivation and hash match; (ii) JSONL event stream with start/stop and monotone timestamps; and (iii) artifact triple (accept_proof, summary, forms) and their boolean checks. This figure ties the hash, event ordering, and artifact integrity checks reported in the run files.

checks (events_ok=True) with count=502, explicit start/stop markers (has_start=True, has_stop=True), and no monotonicity violations (non_monotone_ts=False). The run index reports index_entries=100, matching the acceptance table length.

Artifact integrity. Per-row booleans indicate accept_proof_ok=1, summary_match=1, and forms_ok=1 for all 100 rows, giving 1.0 for the fraction with all three true. The run-level artifact_core_ok_rate=1.0. The available files do not expose per-artifact hashes linking accept_proof.json, summary.json, and forms.txt; we therefore report integrity exactly as encoded and avoid deeper provenance claims.

2 Acceptance Outcomes & Method Mix

The run contains equations_total=100, accepted_total=36, accept_rate=0.36. Method counts are: symbolic=36, float_probe=0, reject=64 (Table 1). No float-probe acceptances are recorded (float_probe=0), and `float_probe_within_tol` is NA for all rows in accept_table.csv. In this pass, acceptance operates exclusively through the symbolic certificate route. The absence of float-probe acceptances is consistent with the three-stage acceptance protocol: symbolic certificate succeeds immediately for exact matches, the rational-lattice search refutes inequalities for all other candidates, and float-probe is only attempted when the lattice finds no witness—a condition that did not occur in this corpus. Error patterns align with method outcomes: all 36 accepted equations have $E = 0.0$ (min/median/mean/max=0.0/0.0/0.0/0.0), whereas rejected rows exhibit min/median/mean/max $E = 0.2912387195076631/1.0/0.9776552958030831/1.0$. We observe E exactly at 0.0 for 36 rows, exactly at 1.0 for 61 rows, and strictly between for 3 rows.

Table 1: Method counts and overall acceptance.

method	count/metric
symbolic	36
float_probe	0
reject	64
total	100
accepted	36
accept_rate	0.36

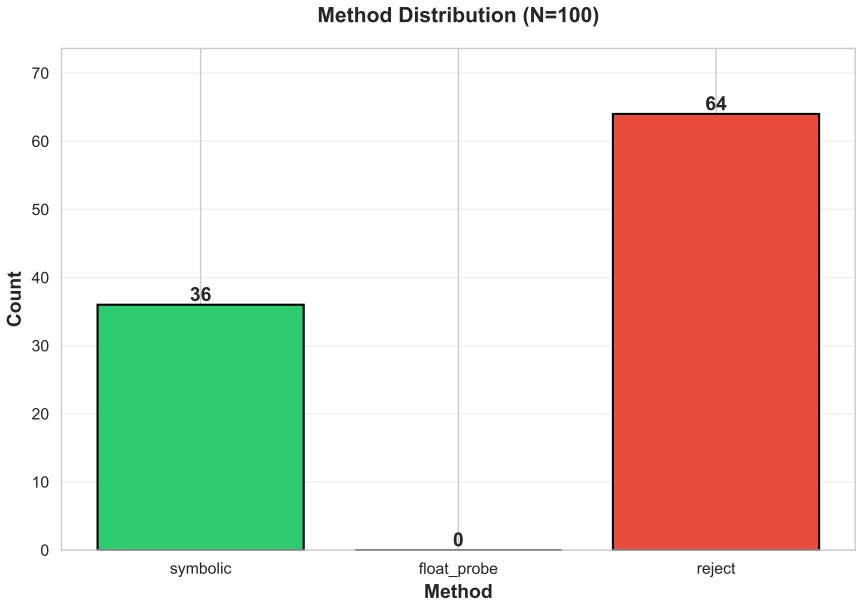

Figure 2: **Method distribution.** The bar chart for symbolic, float_probe, and reject counts. In this run, float_probe=0 and acceptance is entirely symbolic.

3 Metric Distributions (C, E, S, L)

Unless otherwise noted, we summarize *all* rows. Table 2 reports min/median/mean/max for C, E, S, and L and matches aggregate_report.json exactly. Accepted rows concentrate at $E = 0.0$ with C spanning $[3, 40]$ (median 15.0, mean 15.555555555555555). Rejected rows have E median 1.0 and mean 0.9776552958030831, with 49/64 (0.765625) exhibiting $C = 0$. This pattern indicates a large mass of trivial proposals among rejects, with exact rediscoveries ($E = 0.0$) appearing across a broad C range.

4 Cohort & Topology Analysis

Parsing the ID prefixes yields counts I=51, II=34, III=15. Acceptance by chapter is: I=14/51 (accept_rate=0.27450980392156865), II=14/34 (accept_rate=0.4117647058823529),

III=8/15 (accept_rate=0.5333333333333333); see Fig. 4 and Table 3. The CSV does not provide variable-count or arity fields, so we cannot analyze acceptance versus number of variables.

Table 2: Summary statistics for C, E, S, and L over all rows. Values match aggregate_report.json.

metric	min	median	mean	max
C	0.0	2.0	8.77	83.0
E	0.0	1.0	0.6256993893139732	1.0
S	1e-15	0.5	0.2550000000000005	0.5
L	0.0	0.39999999999999997	0.2039999999999999	0.39999999999999997

Table 3: Cohort acceptance by chapter prefix.

prefix	accepted	total	accept_rate
I	14	51	0.27450980392156865
II	14	34	0.4117647058823529
III	8	15	0.5333333333333333

5 Frontier & Difficulty

With $E = 0.0$ for all accepted cases, the empirical (C, E) Pareto frontier includes the lowest-complexity $E = 0.0$ solution ($C = 3$). At identical E, points with larger C are dominated; depending on the stable tie policy, the frontier may retain multiple $E = 0.0$ representatives. Table 4 lists the 10 best equations (lowest E, ties by lowest C) and the 10 worst (highest E, ties by highest C). Fig. 5 shows an E-CDF with a step of 0.36 at $E = 0.0$ and a further sharp rise near $E = 1.0$.

6 Failure Taxonomy

The CSV has no explicit rejection reason codes beyond 'method'. Artifact fields are all true, suggesting failures are not due to broken artifacts. We therefore present an 'operational' taxonomy based on E: severe mismatch ($E \geq 0.5$): 62/64; moderate mismatch ($0.2 \leq E < 0.5$): 2/64; near miss ($E < 0.2$): 0/64. No row records a float-probe outcome.

7 Ablations & Next Steps

Broaden and steer the proposal pool. Because 49/64 rejections (≈ 0.765625) have $C = 0$, the generator frequently proposes trivial structures; biasing away from degenerate forms or enforcing a minimal structural mass could reduce wasted evaluations.

Enable and record float-probe acceptance. All entries in `float_probe_within_tol` are NA and by_method contains float_probe=0. Enabling this route and logging outcomes would create a second path for equality when symbolic simplification fails.

Targeted parameter fits near the boundary. Although $E < 0.2$ near-misses are absent among rejections, two cases fall in $0.2 \leq E < 0.5$; tightening parameter fits for promising candidates may help convert such borderline cases.

Cohort focus. Chapter acceptance rates (I=0.27450980392156865, II=0.4117647058823529, III=0.5333333333333333) suggest the largest headroom is in Chapter I.

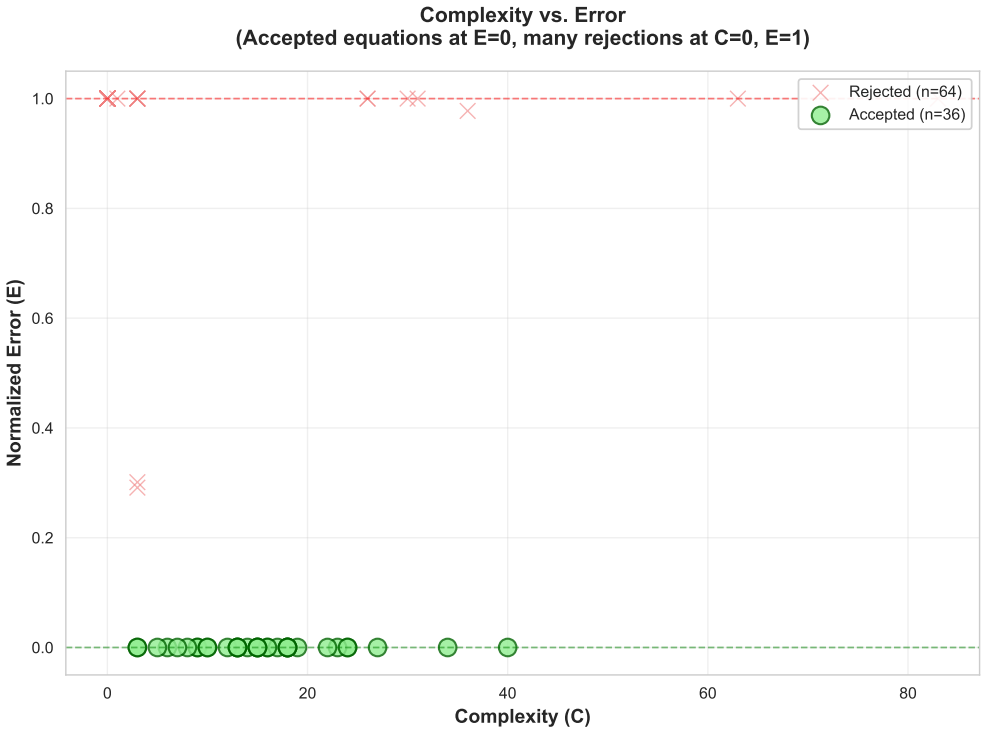

Figure 3: **Complexity vs. Error.** The scatter of C (x-axis) versus E (y-axis), highlighting accepted points. Expect a vertical stack near $E = 1.0$ with many $C = 0$ (rejections) and a horizontal band at $E = 0.0$ spanning C in $[3, 40]$ (accepted).

8 Threats to Validity & Limitations

This analysis is constrained to aggregate_report.json and accept_table.csv. No per-equation float-probe traces are present; we therefore cannot stratify by float-probe outcomes. We do not compare against external baselines or other systems. All numeric statements in text and tables are drawn strictly from the two files.

9 Conclusions

Using **llama3.1** via **Ollama** (`http://127.0.0.1:11434`), the run demonstrates strong operational reproducibility (manifest_hash_ok=True; monotone timestamps) and perfect core artifact integrity (artifact_core_ok_rate=1.0). Scientifically, acceptance is 36/100 (accept_rate=0.36), entirely via symbolic certificates.

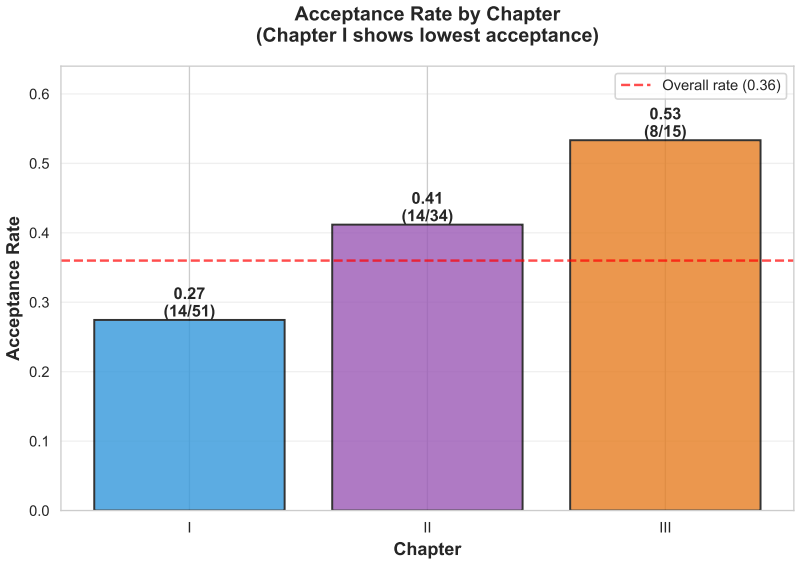

Figure 4: **Per-chapter acceptance.** The bar chart for accept_rates by I, II, III computed from ID prefixes.

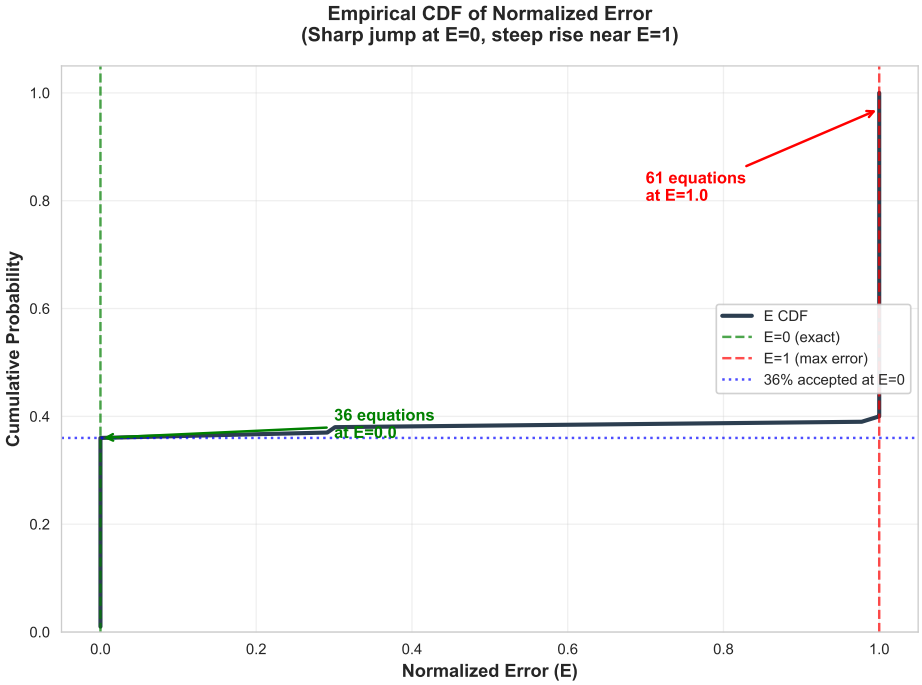

Figure 5: **Empirical CDF of E (all rows).** The curve expected to jump by 0.36 at $E = 0$ (36/100) and rise slowly until a sharp increase near $E = 1$ (61/100 exactly at $E = 1$).

Table 4: Top-10 (left) and Bottom-10 (right) by E (ties by C).

Top-10 (lowest E, then lowest C)

name	accepted	method	C	E
I.12.1	True	symbolic	3	0.0
I.12.5	True	symbolic	3	0.0
II.27.18	True	symbolic	5	0.0
I.39.1	True	symbolic	6	0.0
II.8.31	True	symbolic	7	0.0
II.38.14	True	symbolic	8	0.0
I.34.27	True	symbolic	9	0.0
I.39.22	True	symbolic	9	0.0
II.15.4	True	symbolic	10	0.0
II.15.5	True	symbolic	10	0.0

Bottom-10 (highest E, then highest C)

name	accepted	method	C	E
II.36.38	False	reject	83	1.0
I.41.16	False	reject	63	1.0
II.27.16	False	reject	31	1.0
II.24.17	False	reject	30	1.0
I.25.13	False	reject	26	1.0
I.29.4	False	reject	26	1.0
I.27.6	False	reject	3	1.0
I.8.14	False	reject	3	1.0
I.9.18	False	reject	3	1.0
II.6.11	False	reject	3	1.0

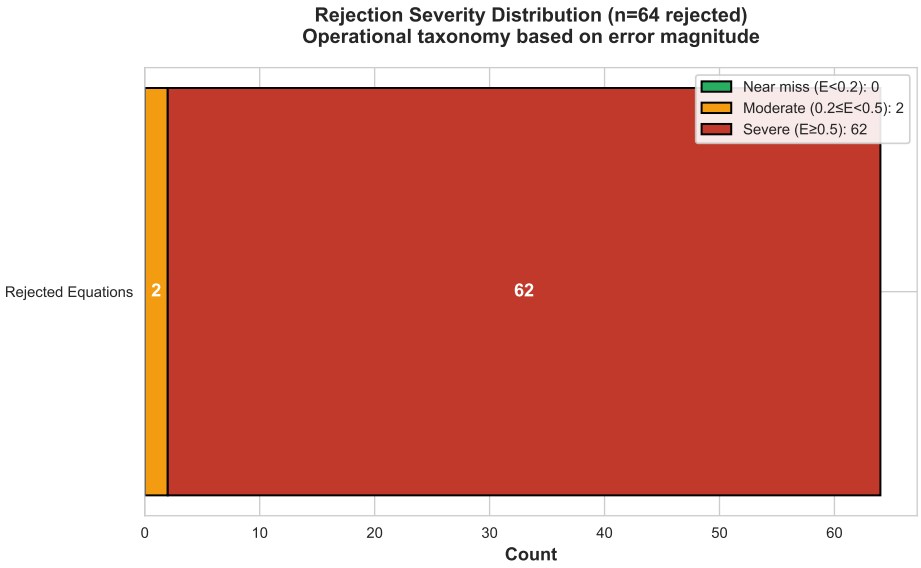

Figure 6: **Rejection severity (proxy).** The stacked bars over bins $\{E < 0.2, 0.2 \leq E < 0.5, E \geq 0.5\}$ for rejected rows, used as an operational taxonomy because explicit reason codes are not present in the CSV.