NEURIPS 2026 · EVALUATIONS & DATASETS TRACK

PINNBench

A Benchmark and Evaluation Study of Training Policy Selection in Hybrid PINN-Operator Solvers

Suan Lee* · **Namhyeon Kim** · **Dongmin Jin**
School of Computer Science, Semyung University *Corresponding author **suanlee@semyung.ac.kr**

SEMYUNG UNIVERSITY
DATA INTELLIGENCE LABORATORY

CODE · GITHUB

DATA · HUGGING FACE

3,545	13	8	3–5	10	77×
recorded runs	PDE configurations	equation families	policies per PDE	seeds per condition	accuracy–regret gap

01 · QUESTION

Which training policy?

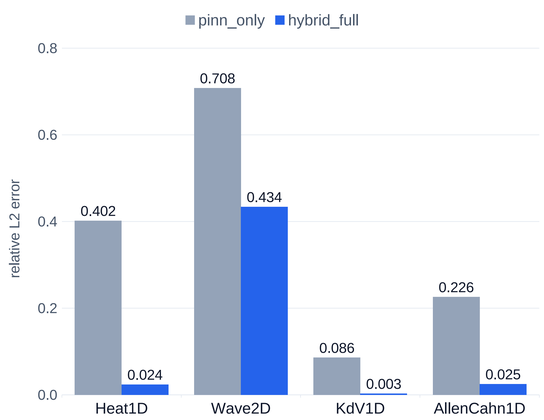
PINNBench fixes a hybrid PINN-operator solver family and asks **which training policy to commit to for a given PDE**, and how that choice should be scored.

STAGE 1 Pretraining 200 ep → STAGE 2 Physics refine 500 ep → STAGE 3 Causal calib. 200 ep

$u = g \cdot u_{OP} + (1 - g) \cdot u_{PINN}$ learned gate g

02 · IN DOMAIN

Hybridization helps



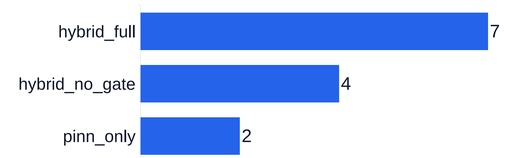
10 seeds, all BH-FDR significant. pinn_only gets the same reference data.

OPEN ARTIFACTS

- All 3,545 runs as JSON, with Croissant metadata
- PDEs, policies, pipeline and routers (MIT)
- make verify-claims recomputes every number

03 · FINDING

No policy dominates



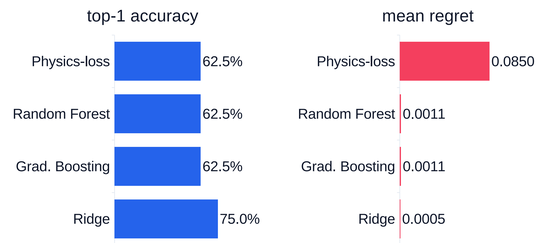
PDEs won by PDE residual, of 13 (10 seeds each).

04 · FINDING

Accuracy is not regret

TOP-1 ACCURACY picked the oracle-best policy?

REGRET how much worse is its pick?



Relative-L2 routing over the 8 PDEs with an exact solution.

77× **Same accuracy (5/8), regret 0.0850 vs 0.0011.**

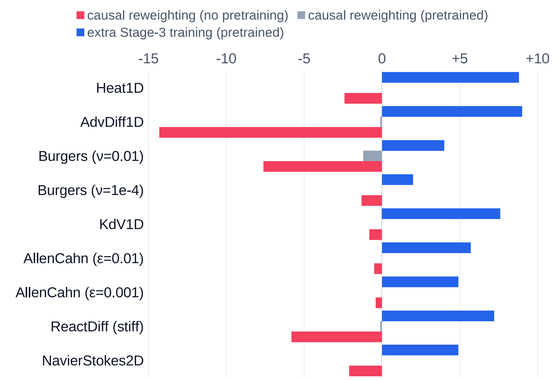
The gap tracks how heavy-tailed best-policy quality is: **206×** spread here, **26×** for tabular classifiers (gap ≤ 3.6×).

WORKED CASE · WAVE2D

The heuristic's one miss costs 0.310; the Random Forest's three misses cost 0.0037, 0.0038, 0.0013.

05 · FINDING

Causal gain = extra training



Δ PDE residual (%), positive = better

+6.0% uniform-weight Stage 3 after pretraining (causal: +5.9%)

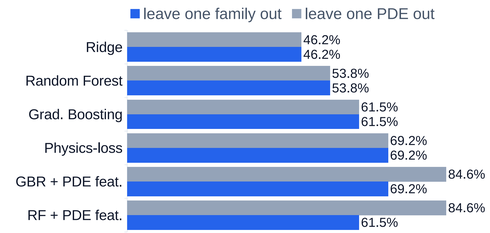
−3.9% causal vs uniform weights without pretraining, 9/9 worse

Epoch-matched controls, 10 seeds, BH-FDR over 54 tests.

06 · FAMILY HOLDOUT

Does PDE-aware routing memorize?

Retrain each router with the held-out PDE's **whole family** removed.



84.6% → 69.2% / 61.5%: at or below the training-free heuristic once the family is held out.

07 · ROBUSTNESS

Outside the idealized regime

Pretraining advantage (error without / with), 720 runs:

PDE	clean, dense	noisy, sparse
Heat1D	10.2×	8.6×
AdvDiff1D	3.7×	2.3×
Burgers1D	1.7×	2.1×
KdV1D	18.9×	**3.0×**
AllenCahn1D	2.5×	2.3×
ReactDiff1D	15.4×	8.7×
Heat3D (d = 3)	1.41×	—
Multiscale 2D	2.68×	—

Wins all 72 cells (noise σ ≤ 0.2 × ≥ 16 points); **no policy learns the fine scale** of the multi-scale PDE.

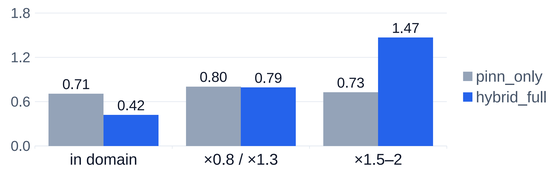
08 · PRACTICE

How to use these results

1. Score routers by regret, not accuracy alone.
2. Pretrain first; causal weights add nothing after it.
3. Distrust probe routing on shock problems.
4. Use a hybrid near its training regime.

09 · OUT OF DOMAIN

Wave2D under shift



Domain rescaled; rel. L2, 5 seeds: 2× worse at 1.5–2×.

Limits: analytical references (regret needs an oracle); 11 of 13 routing configurations are 1D.

Supported by MSIT and NRF of Korea grants (No. RS-2026-25498341 and No. RS-2025-24534935).