PDEBench-Lang: Does notation formation shape neural reasoning about PDEs?

Token Efforts: Raghav Anand Peilin Li Yit Xiaang Ztang Krivan Semlani

Motivation

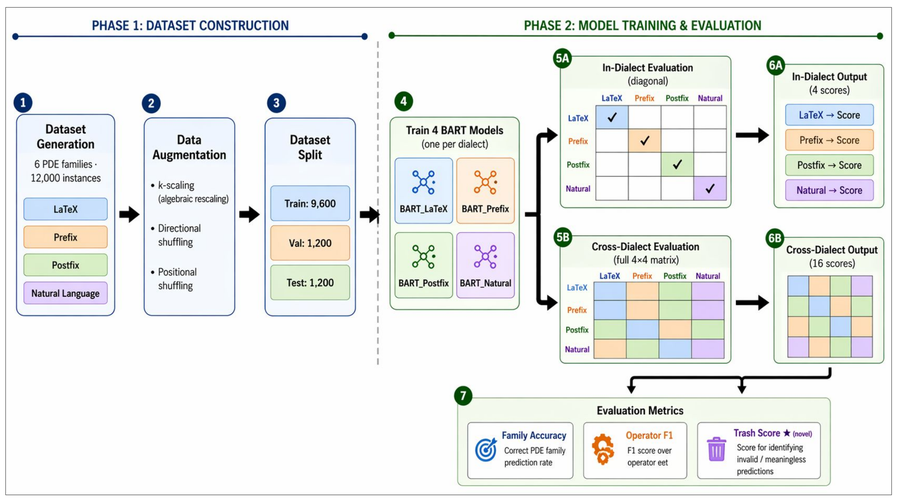
Bhatnagar et al. (2025) achieve a 4–6× speedup in PDE solving by using LLMs to predict solution operators, but hardcode Postfix as the input format purely for computational convenience but never validating whethern it is the right choice for reasoning quality. Since LLMs are pretrained on LaTeX and natural language, this representational mismatch may silently degrade performance, yet nobody has tested it.

We ask: does the symbolic representation format of a PDE affect how well a model reasons about it?

Approach

We generate 12,000 instances across 6 PDE families, each encoded in 4 dialects (LaTeX, Prefix, Postfix, Natural Language). Three augmentations — k-scaling, directional shuffling, and positional shuffling prevent trivial pattern matching. The data is split 9,600 / 1,200 / 1,200 for train / val / test.

We fine-tune four BART-base models, one per dialect. Each model is evaluated in two settings: **in-dialect** (tested on its own format, 4 scores) and **cross-dialect** (tested on all 4 formats, producing a full 4×4 matrix of 16 scores). All predictions are scored on three metrics: Family Accuracy, Operator F1, and the novel Trash Score.



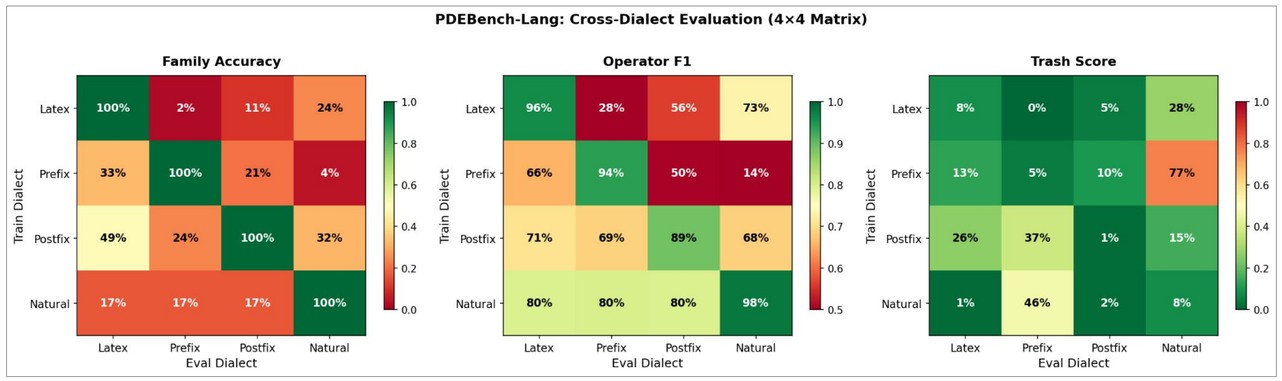
Results/Contributions

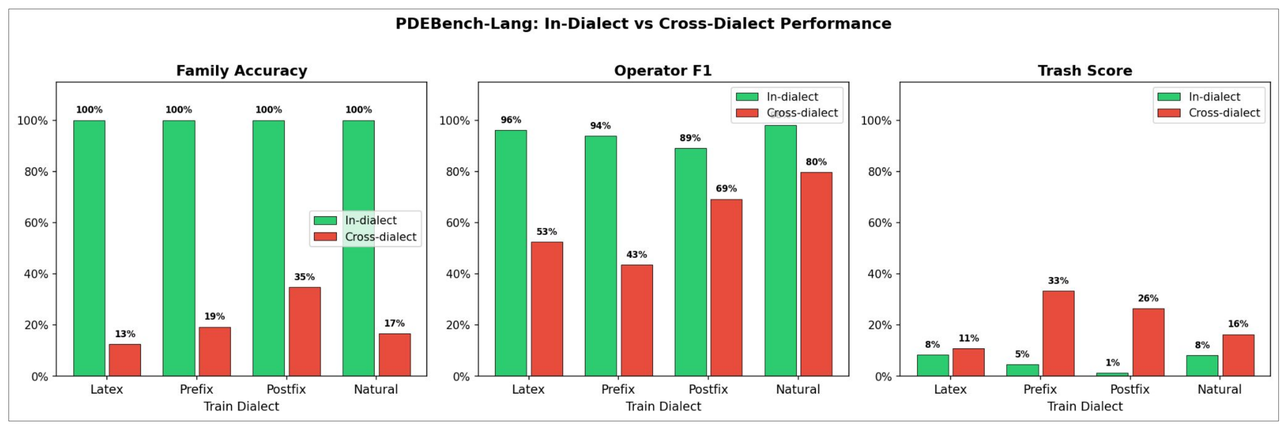
All four models hit 100% family accuracy in-dialect, but cross-dialect generalization collapses to 13–35%.

- In-dialect: 100% family accuracy; cross-dialect collapses to 13–35%
- Postfix generalizes best — stable family accuracy and operator F1 across dialects
- Natural Language transfers operator structure (~80% F1) even when family prediction fails
- Mismatched dialects: trash scores hit 46–77% — output breaks down entirely

Our contributions:

- A novel synthetic PDE dataset (12,000 instances, 6 families, 4 dialects) with k-scaling, directional and positional shuffling to prevent trivial pattern matching
- The first systematic evaluation of symbolic representation format on LLM reasoning quality establishing that **Postfix is the most transferable dialect**.





Key Findings

- Format matters: in-dialect works, cross-dialect collapses
- Postfix is the most transferable symbolic dialect
- Natural Language captures operator structure across formats
- Mismatched dialects produce meaningless output, not just wrong answers
- Postfix places operators immediately after operands — the model never needs to track brackets or precedence
- Every expression follows the same bottom-up pattern regardless of complexity — learned rules transfer directly
- Simpler tokenization means fewer out-of-vocabulary fragments compared to LaTeX's symbol-heavy syntax

References

1. Bhatnagar, A. et al. (2025). *From equations to insights.*
2. d'Ascoli, S. et al. (2024). *ODEFormer: Symbolic regression of dynamical systems.* ICLR 2024.
3. Lample, G. & Charton, F. (2020). *Deep learning for symbolic mathematics.* ICLR 2020.
4. Sun, H. et al. (2024). *PROSE-PDE: Multimodal learning of governing equations.*

Baseline

Dialect	Family Accuracy
Latex	16.67%
Postfix	16.67%
Prefix	16.67%
Natural L	16.67%

Zero-shot BART (no fine-tuning) achieves 16.7% family accuracy across all dialects — identical to random chance (1/6), with no variation by format. This confirms that without fine-tuning, representation format is irrelevant and the task is unsolved, establishing clear room for improvement

Future Work/Plan for Final Report

- Evaluate fine-tuned T5 across all 16 dialect combinations and compare against BART *(done!)*
- Curriculum learning: pre-train on a smaller/simpler PDE subset, then fine-tune on the full dataset *(done!)*
- Larger models: scale to BART-large or T5-large to test whether cross-dialect collapse is a model capacity issue
- Harder dataset: increase coefficient complexity, add more PDE families, introduce noise and edge cases to stress-test
- Real PDE instances: replace synthetic data with PDEs from actual physics literature to test real-world applicability