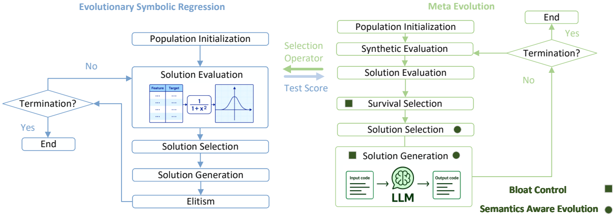
LLM-Meta-SR: In-Context Learning for Evolving Selection Operators in Symbolic Regression

Hengzhe Zhang

Aug 24, 2025

The Challenge in Symbolic Regression



Our Goal:

- Automate selection operator design
- Use LLMs to exceed expert-level design

Symbolic Regression (SR):

- Discover mathematical expressions from data
- Key component: Selection operators
- Currently: Manually designed by experts

Given dataset $\mathcal{D} = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^{N}$

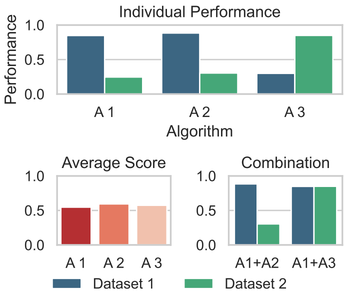
$$f^* = \arg\min_{f \in \mathcal{F}} \mathcal{L}(f(\mathbf{x}), y)$$

- $\mathcal{D}$: training dataset with N samples
- $\mathbf{x}^{(i)}$: input features of sample i
- $y^{(i)}$: target value of sample i
- $\mathcal{F}$: space of candidate symbolic functions
- $\mathcal{L}$: loss function (e.g., squared error)

Key Challenges in LLM-driven Algorithm Evolution

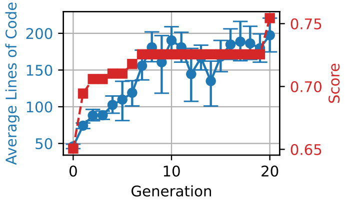
Challenge 1: Semantic Blindness

- Existing methods use only average scores
- Miss fine-grained behavioral differences



Challenge 2: Code Bloat

- LLMs generate unnecessarily complex code
- Reduces interpretability & efficiency



Our Solutions:

1. **Semantic-aware selection**: Use complementarity scores
2. **Bloat control**: Multi-objective survival + prompt constraints
3. **Domain knowledge**: Embed SR principles in prompts

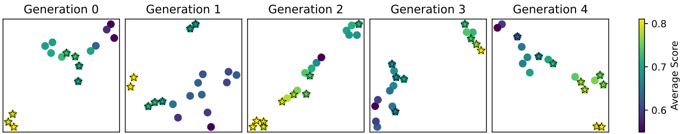
LLM-Meta-SR Framework

Meta-Evolution Workflow:

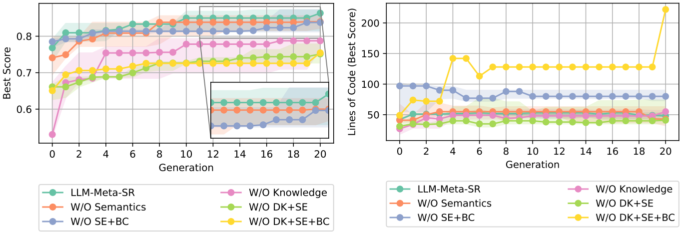
1. Initialize population with LLM
2. Evaluate on SR tasks
3. Semantic-aware parent selection
4. LLM crossover/mutation
5. Multi-objective survival selection

Key Innovations:

- Complementarity: $\mu_i = \frac{1}{d}\sum_{j=1}^{d} \max(s_{a,j}, s_{i,j})$
- Bloat control via Pareto dominance
- Domain knowledge integration

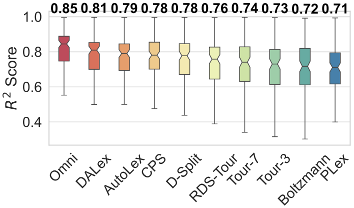


t-SNE visualization of operator semantics showing diversity

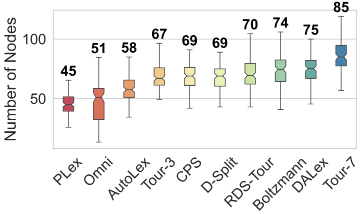


Evolution performance and code length control

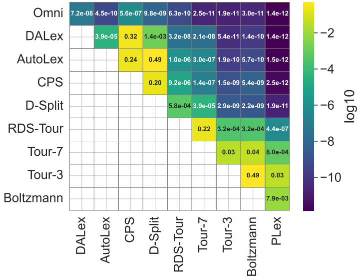
Experimental Results on 100+ Regression Datasets



Test R^2 scores



Model complexity



Statistical significance

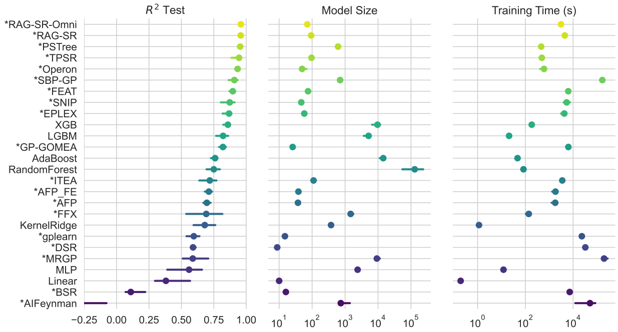
Key Findings:

- Outperforms 9 expert-designed baselines
- Smaller models with better interpretability
- Statistically significant improvements

Impact:

- LLMs exceed expert-level design
- Generalizes to unseen datasets

Advancing State-of-the-Art: RAG-SR-Omni



RAG-SR-Omni:

- Integrates evolved "Omni" operator into SOTA RAG-SR
- Achieves best performance across all metrics

Evolved Operator Properties:

- Balances specificity and complexity
- Adaptive stage-aware pressure
- Semantic complementarity
- Vectorized for efficiency

Conclusion: LLMs can design components surpassing human expertise

Thank You!

Questions?

Key Contributions:

- First LLM-driven meta-SR framework for automatic operator design
- Semantic-aware evolution and effective bloat control mechanisms
- LLMs can exceed expert-level algorithm design in symbolic regression

Paper available at: `https://arxiv.org/abs/2505.18602`