LLM-Driven Symbolic Regression: A Research Synthesis

Prompt: symbolic regression methods based on a LLM
Lanternfish here presents the top 5 papers.

Recent advancements extensively explore leveraging large language models (LLMs) for symbolic regression (SR), shifting from traditional methods like genetic programming. Approaches like SymbolicGPT and ChatSR frame SR as a language modeling or captioning task, enabling improved accuracy, data efficiency, and even conversational control via natural language instructions for specifying desired expression characteristics. LASR induces reusable concepts using LLM queries to accelerate SR, while other methods utilize LLMs for generating program scaffolds or learning policies to guide the search process. Furthermore, LLM-Meta-SR introduces a learning-to-evolve framework, utilizing LLMs to automatically design and optimize selection operators for evolutionary SR algorithms, achieving state-of-the-art performance and surpassing human-level algorithm design in certain benchmarks; these methods collectively demonstrate a move towards LLMs not only assisting but fundamentally driving advancements in SR techniques.

Symbolic Regression and Sequence Modelling with Conditional and Dynamic Language Models

Year: 2024 *Journal:* NA *Authors:* M Valipour Total Score: **8.5**/10 Quality Score: **8.0**/10 Relevance Score: **9.0**/10

Summary

SymbolicGPT introduces a novel transformer-based language model for symbolic regression, framing the problem as a captioning task where the model learns to generate equation skeletons. The approach leverages strengths of probabilistic language models like GPT for improved accuracy and data efficiency. A new scalable data generation method is also proposed to create synthetic equations for pre-training. Experiments demonstrate that SymbolicGPT outperforms competing methods in terms of accuracy, running time, and data efficiency. The model aims to generate equations that fit given data by treating symbolic regression as a language modeling task, exploiting the grammar inherent in mathematical expressions.

ChatSR: Conversational Symbolic Regression

Year: NA *Journal:* NA *Authors:* Y Li, W Li, L Yu, M Wu, J Liu, S Wei, Y Deng Total Score: **8.5**/10 Quality Score: **8.0**/10 Relevance Score: **9.0**/10

Summary

ChatSR, a novel symbolic regression paradigm, leverages multi-modal large language models to generate expressions from data based on natural language instructions. It trains a model to generate expressions by aligning data features with the LLM's word embeddings, enabling the specification of requirements – such as periodicity or symmetry – through natural language.

Experiments demonstrate that ChatSR achieves comparable performance to state-of-the-art methods while generating more concise expressions and exhibits strong zero-shot capability for satisfying user-defined constraints not present in the training data. The approach allows users to specify desired expression characteristics through natural language, simplifying the symbolic regression process and improving flexibility.

Symbolicgpt: A generative transformer model for symbolic regression

Year: 2021 *Journal:* arXiv preprint arXiv:2106.14131 *Authors:* M Valipour, B You, M Panju, A Ghodsi Total Score: **8.5**/10 Quality Score: **8.0**/10 Relevance Score: **9.0**/10

Summary

SymbolicGPT, a transformer-based language model, addresses symbolic regression by framing it as a language modeling task. Unlike traditional methods that train a new model for each dataset, SymbolicGPT is trained once and then rapidly solves regression problems as captioning tasks. It employs a T-net to obtain an order-invariant embedding of the input dataset, effectively representing the data in a way that doesn't depend on the number or order of points. The model then leverages a GPT architecture to predict the mathematical expression, separating constant optimization as a post-processing step. This approach exhibits significant speed improvements and data efficiency compared to methods like genetic programming or deep learning-based methods requiring per-instance training, and demonstrates robustness to varying input dataset sizes. The authors highlight that the model benefits from ongoing advancements in language model technology, offering a path towards improved performance without requiring architectural changes.

Symbolic regression with a learned concept library

Year: 2024 *Journal:* Advances in … *Authors:* A Grayeli, A Sehgal, O Costilla Reyes Total Score: **8.0**/10 Quality Score: **8.0**/10 Relevance Score: **8.0**/10

Summary

Recent work leverages large language models (LLMs) to enhance symbolic regression (SR). LASR, presented in this paper, induces abstract, reusable concepts using zero-shot LLM queries to accelerate SR. Other approaches use LLMs to either generate program scaffolds or learn a neural policy to accelerate search. Funsearch uses an LLM to mutate programs, while LLM-SR leverages LLMs for generating program sketches. A key difference in LASR is the dynamic discovery of concepts during the search process, instead of relying on static specifications. Furthermore, methods like Lilo, focus on learning interpretable libraries through LLM-based compression and documentation of code. Recent advancements also explore in-context symbolic regression, directly leveraging LLMs for function discovery. These methods all aim to leverage the reasoning capabilities of LLMs to improve the efficiency and effectiveness of SR.

LLM-Meta-SR: Learning to Evolve Selection Operators for Symbolic Regression

Year: 2025 *Journal:* arXiv preprint arXiv:2505.18602 *Authors:* H Zhang, Q Chen, B Xue, M

Zhang Total Score: **8.0**/10 Quality Score: **8.0**/10 Relevance Score: **8.0**/10

Summary

The paper proposes a learning-to-evolve framework (LLM-Meta-SR) that utilizes large language models (LLMs) to automatically design selection operators for evolutionary symbolic regression algorithms. It addresses key challenges in LLM-driven algorithm evolution – code bloat and a lack of semantic guidance – through bloat control strategies and semantic-aware selection mechanisms. The framework employs an LLM to generate selection operators, guided by domain knowledge and a feedback loop that incorporates semantic information about task performance. Experimental results demonstrate that the LLM-designed selection operators outperform nine expert-designed baselines on symbolic regression benchmarks, achieving state-of-the-art performance. This showcases the potential of LLMs to not only assist but exceed human-level algorithm design in symbolic regression.