WizardMath: Empowering Mathematical Reasoning for Large Language Models via *Reinforced Evol-Instruct*

Haipeng Luo[2*] **Qingfeng Sun**[1*] **Can Xu**[1†] **Pu Zhao**[1] **Jianguang Lou**[1]
Chongyang Tao[1] **Xiubo Geng**[1] **Qingwei Lin**[1] **Shifeng Chen**[2†] **Dongmei Zhang**[1]
[1]Microsoft
[2]Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences
{caxu,qins,puzhao,jlou,chotao,xigeng,qlin,dongmeiz}@microsoft.com
{hp.luo,shifeng.chen}@siat.ac.cn

Abstract

Large language models (LLMs), such as GPT-4, have shown remarkable performance in natural language processing (NLP) tasks, including challenging mathematical reasoning. However, most existing open-source models are only pre-trained on large-scale internet data and without math-related optimization. In this paper, we present *WizardMath*, which enhances the mathematical reasoning abilities of Llama-2, by applying our proposed *Reinforced Evol-Instruct* method to the domain of math. Through extensive experiments on two mathematical reasoning benchmarks, namely GSM8k and MATH, we reveal the extraordinary capabilities of our model. *WizardMath* surpasses all other open-source LLMs by a substantial margin. Furthermore, our model even outperforms ChatGPT-3.5, Claude Instant-1, PaLM-2 and Minerva on GSM8k, simultaneously surpasses Text-davinci-002, PaLM-1 and GPT-3 on MATH. More details and model weights are public at https://github.com/nlpxucan/WizardLM[3] and https://huggingface.co/WizardLM.

1 Introduction

Recently, Large-scale language models (LLMs) have garnered significant attention and become the go-to approach for numerous natural language processing (NLP) tasks, including open domain conversation [1–4], coding [5–13] and math [14–19]. A conspicuous example is ChatGPT, developed by OpenAI. This model uses extensive pre-training on large-scale internet data and further fine-tuning with specific instruction data and methods. As a result, it achieves state-of-the-art zero-shot performance on various benchmarks. Subsequently, Anthropic, Google, and Meta also launched their competitive products one after another. Notably, Meta's series of Llama [4, 20] models have sparked an open-source revolution and quickly narrowed the gap with those closed-source LLMs. This trend also gradually stimulates the releases of MPT[8], Falcon [21], StarCoder [12], Alpaca [22], Vicuna [23], and WizardLM [24], etc. However, these open models still struggles with the scenarios which require complex multi-step quantitative reasoning, such as solving mathematical and science challenges [25–35].

[*] Equal contribution. Work done during the internship of Luo at Microsoft Research.

[†] Corresponding author: caxu@microsoft.com and shifeng.chen@siat.ac.cn

[3] We are working with our legal team to review and publicly release the code and data in accordance with our policy.

Preprint. Under review.

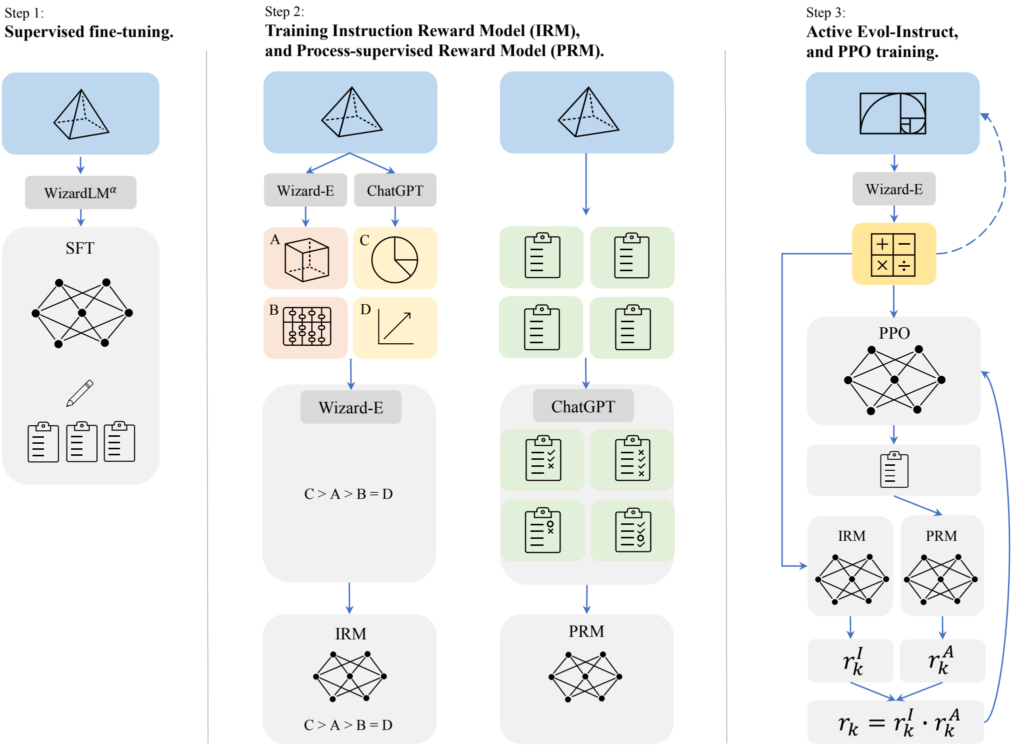

Figure 1: A diagram illustrating the three steps of our method: (1) supervised fine-tuning (SFT), (2) Instruction Reward Model (IRM) training and Process-supervised Reward Model (PRM) training, and (3) Active Evol-Instruct and reinforcement learning via proximal policy optimization (PPO).

Chain-of-thought (CoT) [31] proposes to design better prompts to generate step-by-step solutions, which can lead to improved performance. Self-Consistency [34] also achieves remarkable performance on many reasoning benchmarks, which generates several possible answers from the model and selects the correct one based on majority vote [35]. In recent, [36] finds that process supervision with reinforcement learning significantly outperforms outcome supervision for solving challenging MATH problems.

Inspired by *Evol-Instruct* and Process-supervised Reinforcement Learning, this work aims to enhance the mathematical reasoning abilities of the SOTA open-source LLM, Llama-2 [20]. As shown in the Figure 1, we propose a new method named *Reinforced Evol-Instruct*, which could firstly generate diverse math instructions data by math-specific *Evol-Instruct*, then we train an instruction reward model (IRM) and a process-supervised reward model (PRM) [16, 36–41], the former indicates the quality of the evolved instruction and the later receives feedback for each step in the solution. The brand-new *Evol-Instruct* method includes two downward evolution and upward evolution progress to produce the grade school math and challenging math respectively. Initially, we re-generate, filter and finetune the original math instruction data from GSM8k [42] and MATH [43]. Immediately, we train the Llama-2 models to obtain the reward models and our *WizardMath*.

We perform experiments on two mathematical reasoning benchmarks, namely GSM8k [42] and MATH [43], the results demonstrate that our *WizardMath* outperforms all other open-source LLMs, achieving state-of-the-art performance. Specifically, *WizardMath* observe a substantial improvement in pass@1 with an increase of +24.8 (81.6. vs. 56.8) on GSM8k, and +9.2 (22.7 vs. 13.5) on MATH. Notably, our model even also significantly surpasses OpenAI's ChatGPT-3.5[5], Anthropic's Claude Instant-1 [39], and Google's PaLM-2 [44] in terms of pass@1 on GSM8k.

The main contributions of this work are as following:

- We introduce *WizardMath* model, which enhances the mathematical reasoning abilities for open-source pretrained large language model Llama-2 [20].

- We propose a new method, *Reinforced Evol-Instruct*, alongside *Evol-Instruct* and Reinforcement Learning, for improving LLM reasoning performance.
- *WizardMath* surpasses all other open-source LLMs by a substantial margin in terms of mathematical reasoning, including Llama-2 70B [20], Llama-1 65B [4], Falcon-40B [21], MPT-30B[8], Baichuan-13B Chat[9] and ChatGLM2 12B [45] on both GSM8k [42] and MATH [43].
- *WizardMath* significantly outperforms various main closed-source LLMs, such as ChatGPT[5], GPT-3.5, Claude Instant [39], PaLM-2 [44], PaLM-1 [7] and Minerva[15] on GSM8k.

2 Method

In this section, we elaborate on the details of our *WizardMath*. Following WizardLM and PRMs[36], we propose *Reinforced Evol-Instruct*, which integrates the *Evol-Instruct* and reinforced process supervision method to evolve GSM8k and MATH, and fine-tune the pre-trained Llama-2 with the evolved data and reward models.

As shown in the Figure [1], our methods apply three steps:

1. Supervised fine-tuning.
2. Training instruction reward model, and process-supervised reward model.
3. Active Evol-Instruct, and PPO training.

2.1 Supervised fine-tuning

Following InstructGPT[2], we also firstly fine tune the base with supervised instruction-response pairs, which contains:

1. To make the parsing of each step easier, we few-shot re-generate 15k answers for GSM8k and MATH with an Alpha version of WizardLM 70B model to produce solutions in a step-by-step format, then find out those with a correct answer, and use this data to finetune base Llama model.
2. To enhance the model's ability to adhere to the neural and diverse instructions, we also sample 1.5k open-domain conversations from WizardLM's training data, then merge it with above math corpus as the final SFT training data.

2.2 *Evol-Instruct* principles for math

Motivated by the Evol-Instruct [24] method proposed by WiazrdLM and its effective application on WizardCoder [13], this work attempts to make math instructions with various complexities and diversity to enhance the pre-trained LLMs. Specifically, we adapt Evol-Instruct to a new paradigm including two evolution lines:

1. Downward evolution: It enhances instructions by making the questions easier. For example i): revising high difficulty questions to lower difficulty, or ii) producing a new and easier question with another different topic.
2. Upward evolution: Derived from original Evol-Instruct method, it deepens and generates new and harder questions by i) adding more constraints, ii) concretizing, iii) increasing reasoning.

2.3 *Reinforced Evol-Instruct*

Inspired by InstructGPT[2] and PRMs[36], we train two reward models to predict the quality of the instructions and the correctness of each step in the answer respectively:

1. Instruction Reward Model (IRM): This model aims to judge the quality of the evolved instructions on three aspects: i) Definition, ii) Precision, and iii) Integrity. To produce the ranking list training data of IRM, for each instruction, we firstly use ChatGPT and