2021). As shown in Table 4, DeepSeekMath-Base 7B exhibits significant enhancements in performance on MMLU and BBH over its precursor, DeepSeek-Coder-Base-v1.5 (Guo et al., 2024), illustrating the positive impact of math training on language understanding and reasoning. Additionally, by including code tokens for continual training, DeepSeekMath-Base 7B effectively maintains the performance of DeepSeek-Coder-Base-v1.5 on the two coding benchmarks. Overall, DeepSeekMath-Base 7B significantly outperforms the general model Mistral 7B (Jiang et al., 2023) on the three reasoning and coding benchmarks.

## 3. Supervised Fine-Tuning

### 3.1. SFT Data Curation

We construct a mathematical instruction-tuning dataset covering English and Chinese problems from different mathematical fields and of varying complexity levels: problems are paired with solutions in chain-of-thought (CoT) (Wei et al., 2022), program-of-thought (PoT) (Chen et al., 2022; Gao et al., 2023), and tool-integrated reasoning format (Gou et al., 2023). The total number of training examples is 776K.

- **English mathematical datasets**: We annotate GSM8K and MATH problems with tool-integrated solutions, and adopt a subset of MathInstruct (Yue et al., 2023) along with the training set of Lila-OOD (Mishra et al., 2022) where problems are solved with CoT or PoT. Our English collection covers diverse fields of mathematics, e.g., algebra, probability, number theory, calculus, and geometry.
- **Chinese mathematical datasets**: We collect Chinese K-12 mathematical problems spanning 76 sub-topics such as linear equations, with solutions annotated in both CoT and tool-integrated reasoning format.

### 3.2. Training and Evaluating DeepSeekMath-Instruct 7B

In this section, we introduce DeepSeekMath-Instruct 7B which undergoes mathematical instruction tuning based on DeepSeekMath-Base. Training examples are randomly concatenated until reaching a maximum context length of 4K tokens. We train the model for 500 steps with a batch size of 256 and a constant learning rate of 5e-5.

We evaluate models' mathematical performance both without and with tool use, on 4 quantitative reasoning benchmarks in English and Chinese. We benchmark our model against the leading models of the time:

- **Closed-source models** include: (1) the GPT family among which GPT-4 (OpenAI, 2023) and GPT-4 Code Interpreter [2] are the most capable ones, (2) Gemini Ultra and Pro (Anil et al., 2023), (3) Inflection-2 (Inflection AI, 2023), (4) Grok-1 [3], as well as models recently released by Chinese companies including (5) Baichuan-3 [4], (6) the latest GLM-4 [5] from the GLM family (Du et al., 2022). These models are for general purposes, most of which have undergone a series of alignment procedures.
- **Open-source models** include: general models like (1) DeepSeek-LLM-Chat 67B (DeepSeek-AI, 2024), (2) Qwen 72B (Bai et al., 2023), (3) SeaLLM-v2 7B (Nguyen et al., 2023), and (4)

[2] https://openai.com/blog/chatgpt-plugins#code-interpreter
[3] https://x.ai/model-card
[4] https://www.baichuan-ai.com
[5] https://open.bigmodel.cn/dev/api#glm-4

ChatGLM3 6B (ChatGLM3 Team, 2023), as well as models with enhancements in mathematics including (5) InternLM2-Math 20B [6] which builds on InternLM2 and underwent math training followed by instruction tuning, (6) Math-Shepherd-Mistral 7B which applys PPO training (Schulman et al., 2017) to Mistral 7B (Jiang et al., 2023) with a process-supervised reward model, (7) the WizardMath series (Luo et al., 2023) which improves mathematical reasoning in Mistral 7B and Llama-2 70B (Touvron et al., 2023) using evolve-instruct (i.e., a version of instruction tuning that uses AI-evolved instructions) and PPO training with training problems primarily sourced from GSM8K and MATH, (8) MetaMath 70B (Yu et al., 2023) which is Llama-2 70B fine-tuned on an augmented version of GSM8K and MATH, (9) ToRA 34B Gou et al. (2023) which is CodeLlama 34B fine-tuned to do tool-integrated mathematical reasoning, (10) MAmmoTH 70B (Yue et al., 2023) which is Llama-2 70B instruction-tuned on MathInstruct.

As shown in Table 5, under the evaluation setting where tool use is disallowed, DeepSeekMath-Instruct 7B demonstrates strong performance of step-by-step reasoning. Notably, on the competition-level MATH dataset, our model surpasses all open-source models and the majority of proprietary models (e.g., Inflection-2 and Gemini Pro) by at least 9% absolute. This is true even for models that are substantially larger (e.g., Qwen 72B) or have been specifically enhanced through math-focused reinforcement learning (e.g., WizardMath-v1.1 7B). While DeepSeekMath-Instruct rivals the Chinese proprietary models GLM-4 and Baichuan-3 on MATH, it still underperforms GPT-4 and Gemini Ultra.

Under the evaluation setting where models are allowed to integrate natural language reasoning and program-based tool use for problem solving, DeepSeekMath-Instruct 7B approaches an accuracy of 60% on MATH, surpassing all existing open-source models. On the other benchmarks, our model is competitive with DeepSeek-LLM-Chat 67B, the prior state-of-the-art that is 10 times larger.

## 4. Reinforcement Learning

### 4.1. Group Relative Policy Optimization

Reinforcement learning (RL) has been proven to be effective in further improving the mathematical reasoning ability of LLMs after the Supervised Fine-Tuning (SFT) stage (Luo et al., 2023; Wang et al., 2023b). In this section, we introduce our efficient and effective RL algorithm, Group Relative Policy Optimization (GRPO).

#### *4.1.1. From PPO to GRPO*

Proximal Policy Optimization (PPO) (Schulman et al., 2017) is an actor-critic RL algorithm that is widely used in the RL fine-tuning stage of LLMs (Ouyang et al., 2022). In particular, it optimizes LLMs by maximizing the following surrogate objective:

$$\mathcal{J}_{PPO}(\theta) = \mathbb{E}[q \sim P(Q), o \sim \pi_{\theta_{old}}(O|q)] \frac{1}{|o|} \sum_{t=1}^{|o|} \min \left[ \frac{\pi_\theta(o_t|q,o_{<t})}{\pi_{\theta_{old}}(o_t|q,o_{<t})} A_t, \text{clip}\left( \frac{\pi_\theta(o_t|q,o_{<t})}{\pi_{\theta_{old}}(o_t|q,o_{<t})}, 1-\varepsilon, 1+\varepsilon \right) A_t \right], \quad (1)$$

where $\pi_\theta$ and $\pi_{\theta_{old}}$ are the current and old policy models, and $q, o$ are questions and outputs sampled from the question dataset and the old policy $\pi_{\theta_{old}}$, respectively. $\varepsilon$ is a clipping-related hyper-parameter introduced in PPO for stabilizing training. $A_t$ is the advantage, which is computed by applying Generalized Advantage Estimation (GAE) (Schulman et al., 2015), based

[6] https://github.com/InternLM/InternLM-Math

| Model | Size | English Benchmarks | | Chinese Benchmarks | |
|---|---|---|---|---|---|
| | | GSM8K | MATH | MGSM-zh | CMATH |
| **Chain-of-Thought Reasoning** | | | | | |
| Closed-Source Model | | | | | |
| Gemini Ultra | - | 94.4% | 53.2% | - | - |
| GPT-4 | - | 92.0% | 52.9% | - | 86.0% |
| Inflection-2 | - | 81.4% | 34.8% | - | - |
| GPT-3.5 | - | 80.8% | 34.1% | - | 73.8% |
| Gemini Pro | - | 86.5% | 32.6% | - | - |
| Grok-1 | - | 62.9% | 23.9% | - | - |
| Baichuan-3 | - | 88.2% | 49.2% | - | - |
| GLM-4 | - | 87.6% | 47.9% | - | - |
| Open-Source Model | | | | | |
| InternLM2-Math | 20B | 82.6% | 37.7% | - | - |
| Qwen | 72B | 78.9% | 35.2% | - | - |
| Math-Shepherd-Mistral | 7B | 84.1% | 33.0% | - | - |
| WizardMath-v1.1 | 7B | 83.2% | 33.0% | - | - |
| DeepSeek-LLM-Chat | 67B | 84.1% | 32.6% | 74.0% | 80.3% |
| MetaMath | 70B | 82.3% | 26.6% | 66.4% | 70.9% |
| SeaLLM-v2 | 7B | 78.2% | 27.5% | 64.8% | - |
| ChatGLM3 | 6B | 72.3% | 25.7% | - | - |
| WizardMath-v1.0 | 70B | 81.6% | 22.7% | 64.8% | 65.4% |
| **DeepSeekMath-Instruct** | 7B | 82.9% | 46.8% | 73.2% | 84.6% |
| **DeepSeekMath-RL** | 7B | **88.2%** | **51.7%** | **79.6%** | **88.8%** |
| **Tool-Integrated Reasoning** | | | | | |
| Closed-Source Model | | | | | |
| GPT-4 Code Interpreter | - | 97.0% | 69.7% | - | - |
| Open-Source Model | | | | | |
| InternLM2-Math | 20B | 80.7% | 54.3% | - | - |
| DeepSeek-LLM-Chat | 67B | 86.7% | 51.1% | 76.4% | 85.4% |
| ToRA | 34B | 80.7% | 50.8% | 41.2% | 53.4% |
| MAmmoTH | 70B | 76.9% | 41.8% | - | - |
| **DeepSeekMath-Instruct** | 7B | 83.7% | 57.4% | 72.0% | 84.3% |
| **DeepSeekMath-RL** | 7B | **86.7%** | **58.8%** | **78.4%** | **87.6%** |

Table 5 | Performance of Open- and Closed-Source models with both Chain-of-Thought and Tool-Integrated Reasoning on English and Chinese Benchmarks. Scores in gray denote majority votes with 32 candidates; The others are Top1 scores. DeepSeekMath-RL 7B beats all open-source models from 7B to 70B, as well as the majority of closed-source models. Although DeepSeekMath-RL 7B is only further trained on chain-of-thought-format instruction tuning data of GSM8K and MATH, it improves over DeepSeekMath-Instruct 7B on all benchmarks.