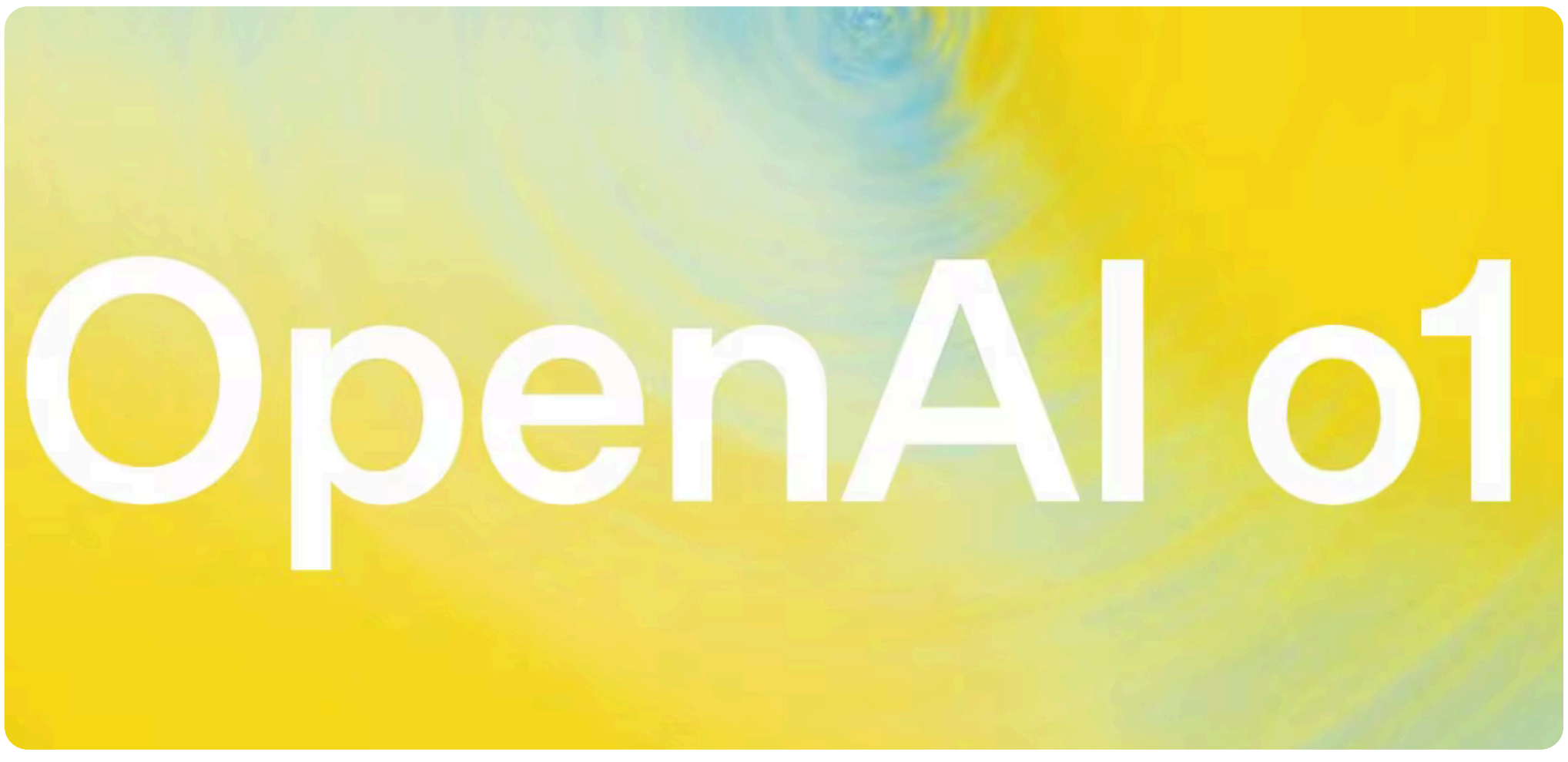
Analytics Vidhya

deepseek R1

VS



Which One is Faster, Cheaper and Smarter?

DeepSeek R1

The DeepSeek R1 model boasts a **671 billion parameters** architecture and has been trained on the DeepSeek V3 Base model. Its focus on Chain of Thought (CoT) reasoning makes it a strong contender for tasks requiring advanced comprehension and reasoning. Interestingly, despite its large parameter count, only 37 billion parameters are activated during most operations, similar to DeepSeek V3.

DeepSeek R1 isn't just a monolithic model; the ecosystem includes six distilled models fine-tuned on synthetic data derived from DeepSeek R1 itself. These smaller models vary in size and target specific use cases, offering solutions for developers who need lighter, faster models while maintaining impressive performance.

Distilled Model Lineup

Model	Base Model	Download
DeepSeek-R1-Distill-Qwen-1.5B	Qwen2.5-Math-1.5B	🤗 HuggingFace
DeepSeek-R1-Distill-Qwen-7B	Qwen2.5-Math-7B	🤗 HuggingFace
DeepSeek-R1-Distill-Llama-8B	Llama-3.1-8B	🤗 HuggingFace
DeepSeek-R1-Distill-Qwen-14B	Qwen2.5-14B	🤗 HuggingFace
DeepSeek-R1-Distill-Qwen-32B	Qwen2.5-32B	🤗 HuggingFace
DeepSeek-R1-Distill-Llama-70B	Llama-3.3-70B-Instruct	🤗 HuggingFace

These distilled models enable flexibility, catering to both local deployment and API usage. Notably, the Llama 33.7B model outperforms the o1 Mini in several benchmarks, underlining the strength of the distilled variants.

Model	#Total Params	#Activated Params	Context Length	Download
DeepSeek-R1-Zero	671B	37B	128K	🤗 HuggingFace
DeepSeek-R1	671B	37B	128K	🤗 HuggingFace

High Performance at Low Cost

How did they do it?

- **Reinforcement Learning First** – Instead of relying heavily on human-labeled data, DeepSeek R1 used **self-evolution** to boost reasoning, cutting annotation costs. A small supervised dataset provided a solid "cold start."

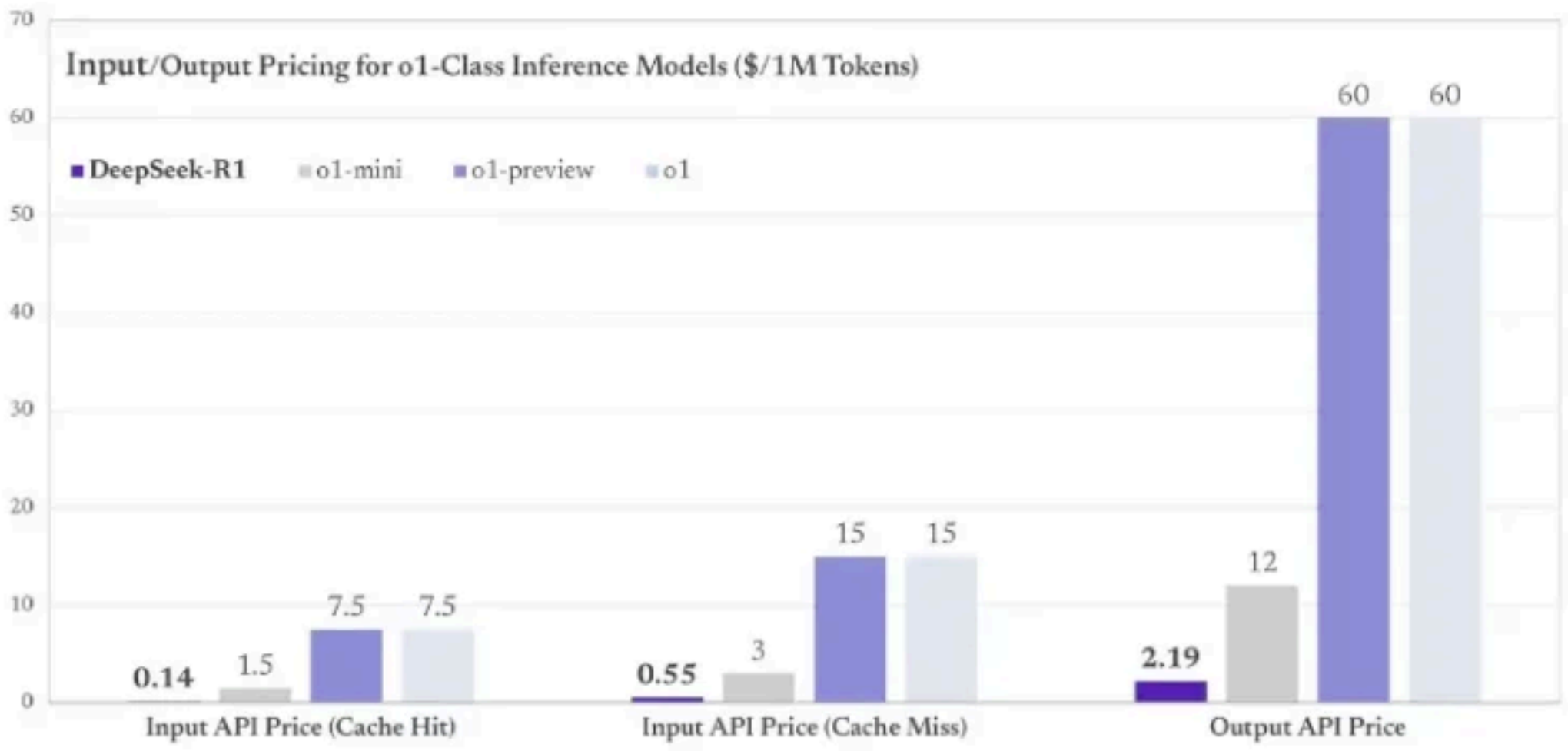
- **Smart Distillation** – They transferred powerful reasoning from large models into smaller, efficient ones (e.g., 14B), achieving top-tier performance with lower costs.

- **Benchmark Focus** – R1 shines in key areas like math & coding, ranking impressively on AIME (79.8%) and MMLU (90.8%).

- **Efficient Architecture** – Techniques like **Sparse Attention** and **Mixture of Experts (MoE)** keep it fast & cost-effective.

- **Strategic Design** – Prioritizing reasoning over general NLP ensures optimized performance where it matters most.

Price Comparison



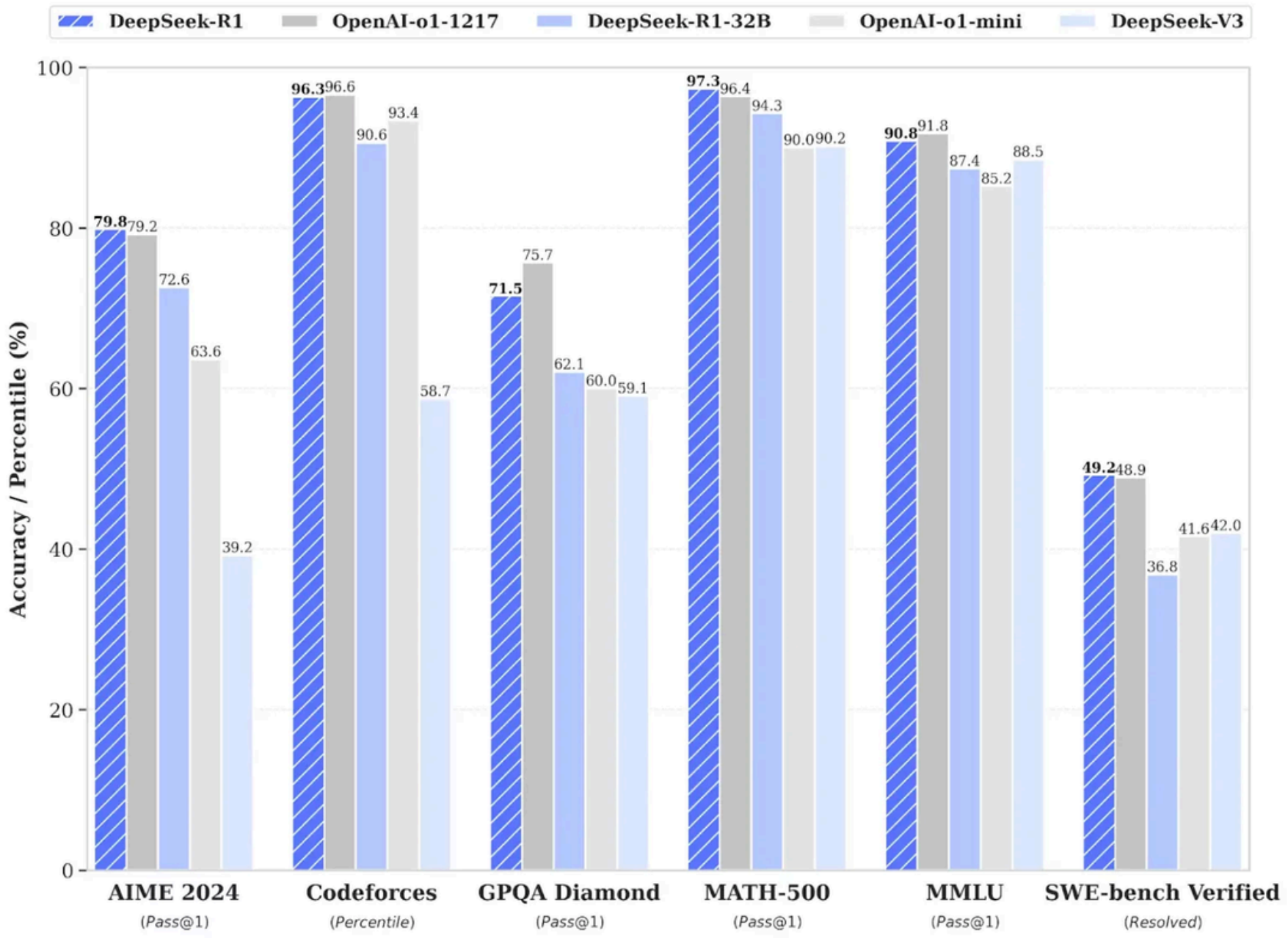
DeepSeek R1 scores comparably to OpenAI o1 in most evaluations and even outshines it in specific cases. This high level of performance is complemented by accessibility; DeepSeek R1 is free to use on the DeepSeek chat platform and offers affordable API pricing. Here's a cost comparison:

- DeepSeek R1 API: 55 Cents for input, $2.19 for output (1 million tokens)
- OpenAI o1 API: $15 for input, $60 for output (1 million tokens)

API is **96.4% cheaper than chatgpt**.

DeepSeek R1's lower costs and free chat platform access make it an attractive option for budget-conscious developers and enterprises looking for scalable AI solutions.

Comparison of Different Benchmarks



1. AIME 2024 (Pass@1)

- DeepSeek-R1: 79.8% accuracy
- OpenAI o1-1217: 79.2% accuracy
- **Explanation**:
    - This benchmark evaluates performance on the American Invitational Mathematics Examination (AIME), a challenging math contest.
    - DeepSeek-R1 slightly outperforms OpenAI-o1-1217 by 0.6%, meaning it’s marginally better at solving these types of math problems.

2. **Codeforces (Percentile)**

- DeepSeek-R1: 96.3%
- OpenAI o1-1217: 96.6%
- **Explanation**:
    - Codeforces is a popular competitive programming platform, and percentile ranking shows how well the models perform compared to others.
    - OpenAI-o1-1217 is slightly better (by 0.3%), meaning it may have a slight advantage in handling algorithmic and coding challenges.

3. **GPQA Diamond (Pass@1)**

- DeepSeek-R1: 71.5%
- OpenAI o1-1217: 75.7%
- **Explanation**:
    - GPQA Diamond assesses a model's ability to answer complex general-purpose questions.
    - OpenAI-o1-1217 performs better by 4.2%, indicating stronger general question-answering capabilities in this category.

4. **MATH-500 (Pass@1)**

- DeepSeek-R1: 97.3%
- OpenAI o1-1217: 96.4%
- **Explanation**:
    - This benchmark measures math problem-solving skills across a wide range of topics.
    - DeepSeek-R1 scores higher by 0.9%, showing it might have better precision and reasoning for advanced math problems.

5. **MMLU (Pass@1)**

- DeepSeek-R1: 90.8%
- OpenAI o1-1217: 91.8%
- **Explanation**:
    - MMLU (Massive Multitask Language Understanding) tests the model's general knowledge across subjects like history, science, and social studies.
    - OpenAI-o1-1217 is 1% better, meaning it might have a broader or deeper understanding of diverse topics.

6. **SWE-bench Verified (Resolved)**

- DeepSeek-R1: 49.2%
- OpenAI o1-1217: 48.9%
- **Explanation**:
    - This benchmark evaluates the model's performance in resolving software engineering tasks.
    - DeepSeek-R1 has a slight 0.3% advantage, indicating a similar level of coding proficiency with a small lead.