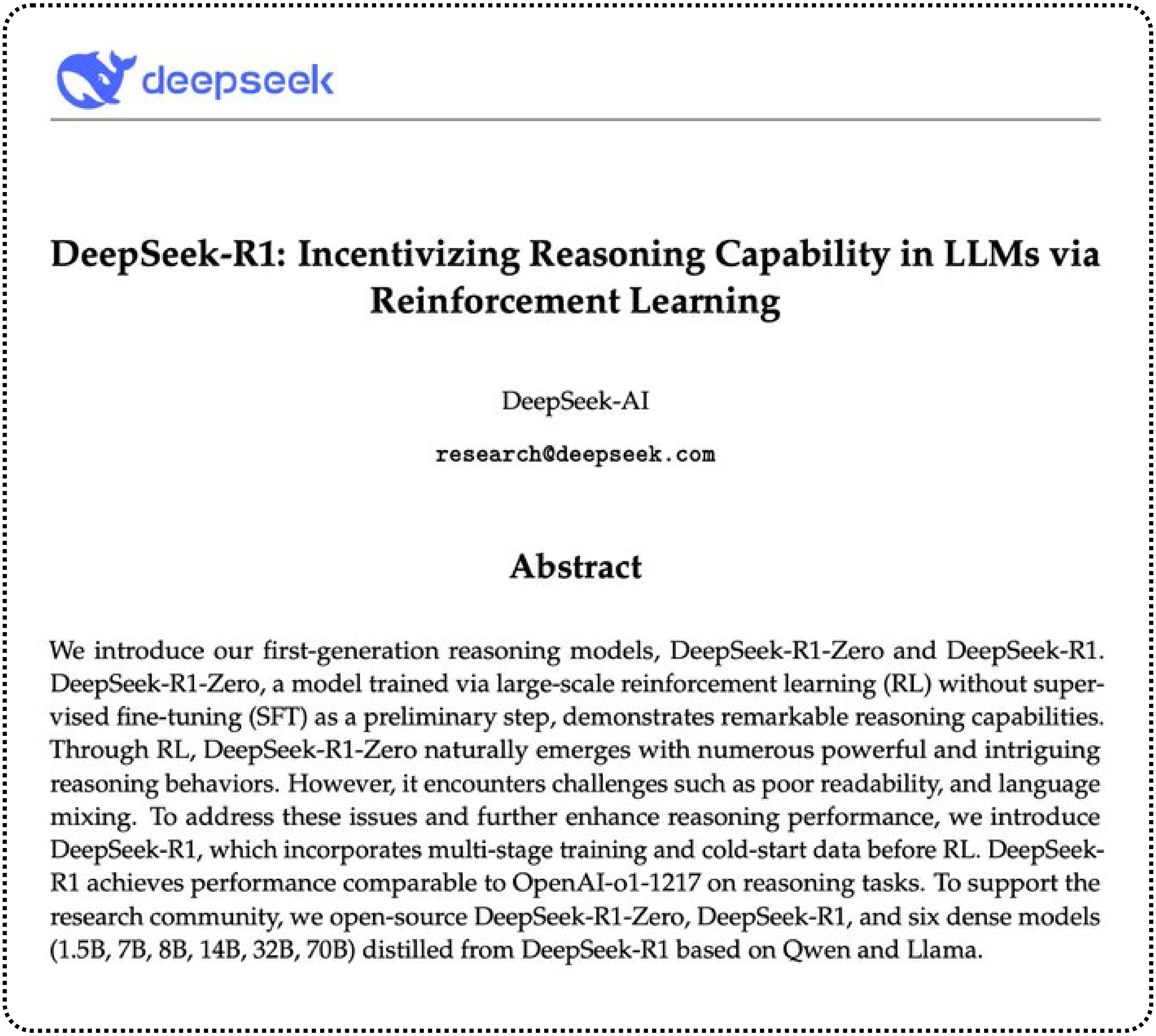
Analytics Vidhya

A Perfect Guide to DeepSeek R1



deepseek

DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning

DeepSeek-AI

research@deepseek.com

Abstract

We introduce our first-generation reasoning models, DeepSeek-R1-Zero and DeepSeek-R1. DeepSeek-R1-Zero, a model trained via large-scale reinforcement learning (RL) without supervised fine-tuning (SFT) as a preliminary step, demonstrates remarkable reasoning capabilities. Through RL, DeepSeek-R1-Zero naturally emerges with numerous powerful and intriguing reasoning behaviors. However, it encounters challenges such as poor readability, and language mixing. To address these issues and further enhance reasoning performance, we introduce DeepSeek-R1, which incorporates multi-stage training and cold-start data before RL. DeepSeek-R1 achieves performance comparable to OpenAI-o1-1217 on reasoning tasks. To support the research community, we open-source DeepSeek-R1-Zero, DeepSeek-R1, and six dense models (1.5B, 7B, 8B, 14B, 32B, 70B) distilled from DeepSeek-R1 based on Qwen and Llama.

Free Course in the description

deepseek

What’s DeepSeek R1?

- DeepSeek-R1 is a reasoning-focused large language model (LLM) developed to enhance reasoning capabilities in Generative AI systems through the method of advanced reinforcement learning (RL) techniques.

- It represents a significant step toward improving reasoning in LLMs, particularly without relying heavily on supervised fine-tuning (SFT) as a preliminary step.

- Essentially, DeepSeek-R1 addresses a key challenge in AI: enhancing reasoning without relying heavily on supervised fine-tuning (SFT).

- Innovative training methodologies power the models to tackle complex tasks like mathematics, coding, and logic.

DeepSeek-R1: Training

Reinforcement Learning

- DeepSeek-R1-Zero is trained exclusively using reinforcement learning (RL) without any SFT. This unique approach incentivizes the model to autonomously develop advanced reasoning capabilities like self-verification, reflection, and CoT (Chain-of-Thought) reasoning.

Reward Design

- The system assigns rewards for reasoning accuracy based on task-specific benchmarks.
- It also gives secondary rewards for structured, readable, and coherent reasoning outputs.

Rejection Sampling

- During RL, multiple reasoning trajectories are generated, and the best-performing ones are selected to guide the training process further.

Cold-Start Initialization with Human-Annotated Data

- For DeepSeek-R1, human-annotated examples of long CoT reasoning are used to initialize the training pipeline. This ensures better readability and alignment with user expectations.

- This step bridges the gap between pure RL training (which can lead to fragmented or ambiguous outputs) and high-quality reasoning outputs.

Multi-Stage Training Pipeline

- Stage 1: Cold-Start Data Pretraining: A curated dataset of human annotations primes the model with basic reasoning structures.

- Stage 2: Reinforcement Learning: The model tackles RL tasks, earning rewards for accuracy, coherence, and alignment.

- Stage 3: Fine-Tuning with Rejection Sampling: The system fine-tunes outputs from RL and reinforces the best reasoning patterns.

Distillation

- Larger models trained with this pipeline are distilled into smaller versions, maintaining reasoning performance while drastically reducing computational costs.

- Distilled models inherit the capabilities of larger counterparts, such as DeepSeek-R1, without significant performance degradation.

DeepSeek R1: Models

DeepSeek R1 comes with **two core** and **six distilled models.**

Core Models

DeepSeek-R1-Zero

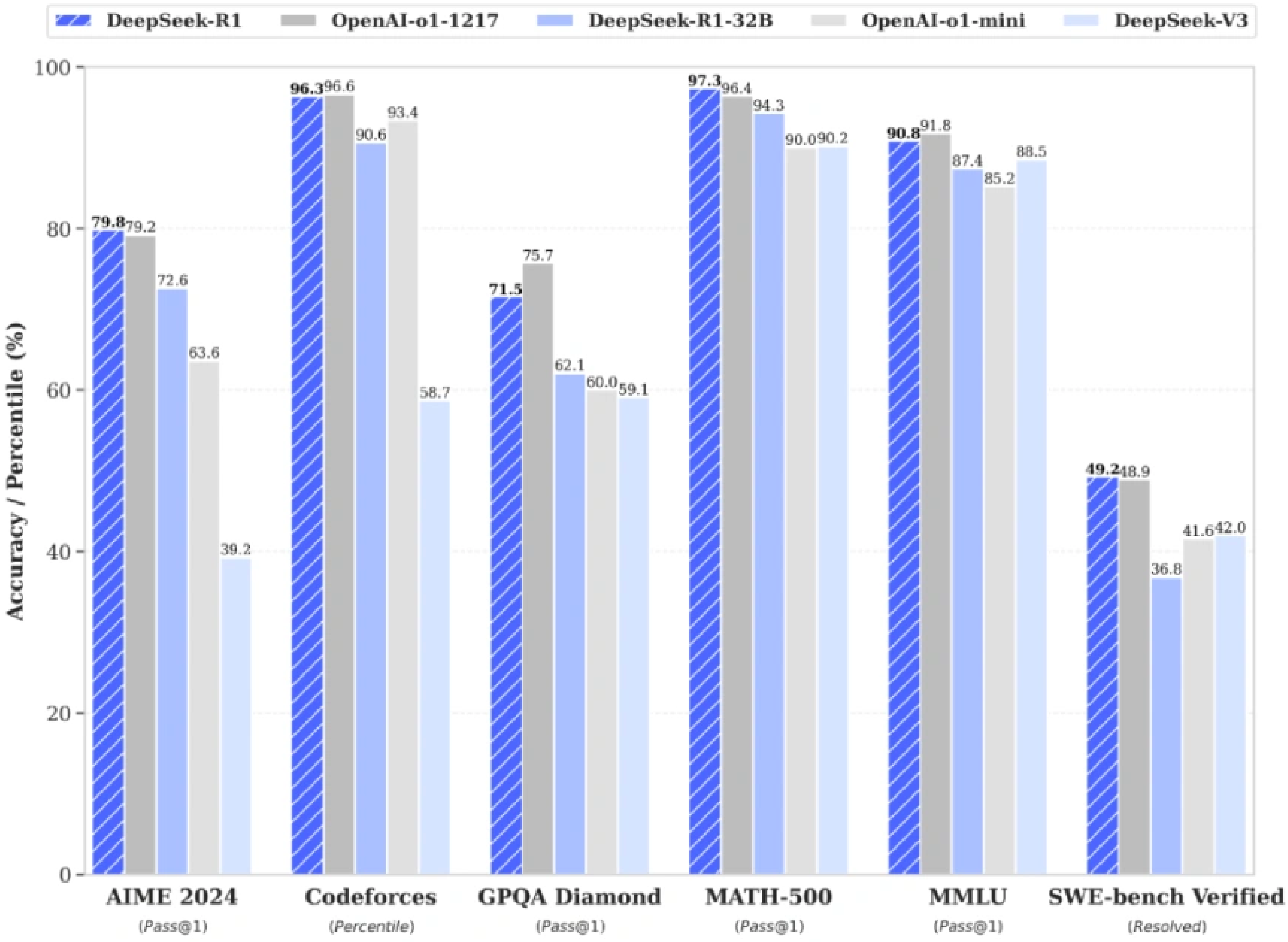
Trained exclusively through reinforcement learning (RL) on a base model, without any supervised fine-tuning.Demonstrates advanced reasoning behaviors like self-verification and reflection, achieving notable results on benchmarks such as:

- AIME 2024
- Codeforces

Challenges: Struggles with readability and language mixing due to a lack of cold-start data and structured fine-tuning.

DeepSeek-R1

Builds upon DeepSeek-R1-Zero by incorporating cold-start data (human-annotated long chain-of-thought (CoT) examples) for enhanced initialization.Introduces multi-stage training, including reasoning-oriented RL and rejection sampling for better alignment with human preferences.



Competes directly with OpenAI’s o1-1217, achieving:

- AIME 2024: Pass@1 score of 79.8%, marginally outperforming o1-1217.
- MATH-500: Pass@1 score of 97.3%, on par with o1-1217.

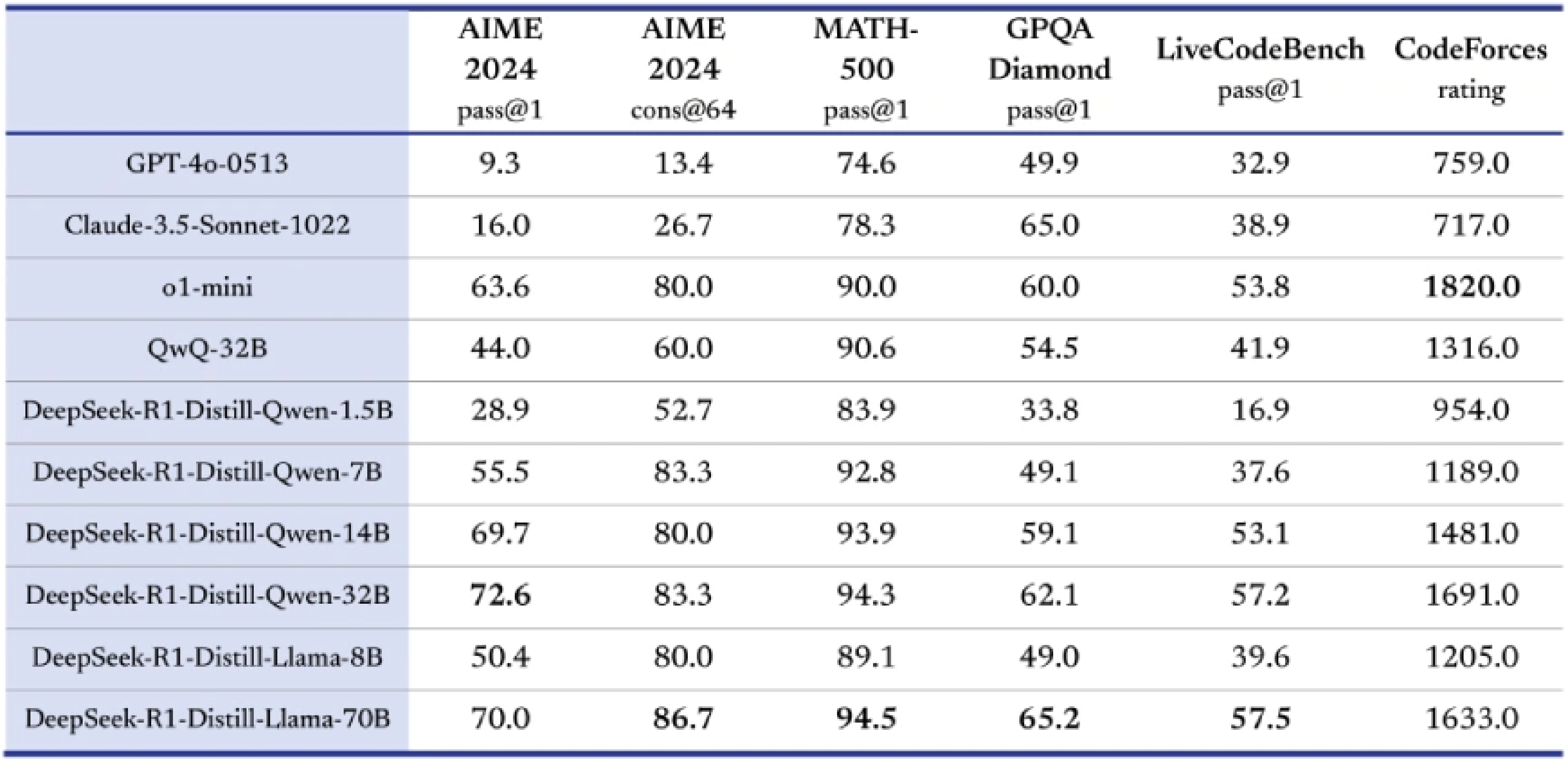
Excels in knowledge-intensive and STEM-related tasks, as well as coding challenges.

Distilled Models

In a groundbreaking move, DeepSeek-AI has also released distilled versions of the R1 model, ensuring that smaller, computationally efficient models inherit the reasoning prowess of their larger counterparts. These distilled models include:

- Qwen Series
- Llama Series

These smaller models outperform open-source competitors like QwQ-32B-Preview while competing effectively with proprietary models like OpenAI's o1-mini.



	AIME 2024 pass@1	AIME 2024 cons@64	MATH-500 pass@1	GPQA Diamond pass@1	LiveCodeBench pass@1	CodeForces rating
GPT-4o-0513	9.3	13.4	74.6	49.9	32.9	759.0
Claude-3.5-Sonnet-1022	16.0	26.7	78.3	65.0	38.9	717.0
o1-mini	63.6	80.0	90.0	60.0	53.8	**1820.0**
QwQ-32B	44.0	60.0	90.6	54.5	41.9	1316.0
DeepSeek-R1-Distill-Qwen-1.5B	28.9	52.7	83.9	33.8	16.9	954.0
DeepSeek-R1-Distill-Qwen-7B	55.5	83.3	92.8	49.1	37.6	1189.0
DeepSeek-R1-Distill-Qwen-14B	69.7	80.0	93.9	59.1	53.1	1481.0
DeepSeek-R1-Distill-Qwen-32B	**72.6**	83.3	94.3	62.1	57.2	1691.0
DeepSeek-R1-Distill-Llama-8B	50.4	80.0	89.1	49.0	39.6	1205.0
DeepSeek-R1-Distill-Llama-70B	70.0	**86.7**	**94.5**	**65.2**	**57.5**	1633.0