Decontamination To prevent benchmark contamination, we implemented comprehensive decontamination procedures for both pre-training and post-training data. DeepSeek-V3 base has a knowledge cutoff date of July 2024, predating evaluation benchmarks like CNMO 2024, and we filtered out any text segments (including web pages and GitHub files) that contained matching 10-gram sequences from evaluation questions or reference solutions. As one example of our decontamination efforts, in the mathematics domain alone, our decontamination process identified and removed approximately six million potential pre-training texts. For post-training, mathematical SFT data and RL training prompts were sourced exclusively from pre-2023 competitions and underwent the same n-gram filtering protocol used in pre-training, ensuring no overlap between training and evaluation data. These measures ensure our model evaluation results reflect genuine problem-solving capabilities rather than memorization of test data.

However, we acknowledge that the n-gram based decontamination method cannot prevent the paraphrase of testset. Therefore, it is possible that benchmarks released before 2024 may suffer from contamination issues.

Evaluation Prompts Following the setup in DeepSeek-V3, standard benchmarks such as MMLU, DROP, GPQA Diamond, and SimpleQA are evaluated using prompts from the simple-evals framework. For MMLU-Redux, we adopt the Zero-Eval prompt format (Lin, 2024) in a zero-shot setting. In terms of MMLU-Pro, C-Eval and CLUE-WSC, since the original prompts are few-shot, we slightly modify the prompt to the zero-shot setting. The CoT in few-shot may hurt the performance of DeepSeek-R1. Other datasets follow their original evaluation protocols with default prompts provided by their creators. For code and math benchmarks, the HumanEval-Mul dataset covers eight mainstream programming languages (Python, Java, C++, C#, JavaScript, TypeScript, PHP, and Bash). Model performance on LiveCodeBench is evaluated using CoT format, with data collected between August 2024 and January 2025. The Codeforces dataset is evaluated using problems from 10 Div.2 contests, along with expert-crafted test cases, after which the expected ratings and percentages of competitors are calculated. SWE-Bench verified results are obtained via the agentless framework (Xia et al., 2024). AIDER-related benchmarks are measured using a "diff" format. DeepSeek-R1 outputs are capped at a maximum of 32,768 tokens for each benchmark.

Table 18 to Table 32 present examples of our evaluation formats on different benchmarks. We also detail the specific capabilities of large language models assessed by each benchmark in the corresponding table captions.

Baselines We conduct comprehensive evaluations against several strong baselines, including DeepSeek-V3, Claude-Sonnet-3.5-1022, GPT-4o-0513, OpenAI-o1-mini, and OpenAI-o1-1217. Since accessing the OpenAI-o1-1217 API is challenging in mainland China, we report its performance based on official reports. For distilled models, we also compare the open-source model QwQ-32B-Preview (Qwen, 2024a).

We set the maximum generation length to 32,768 tokens for the models. We found that using greedy decoding to evaluate long-output reasoning models results in higher repetition rates and significant variability across different checkpoints. Therefore, we default to pass@k evaluation (Chen et al., 2021) and report pass@1 using a non-zero temperature. Specifically, we use a sampling temperature of 0.6 and a top-p value of 0.95 to generate k responses (typically between 4 and 64, depending on the test set size) for each question. Sepcifically, we use $k = 64$ for AIME and GPQA, $k = 16$ for MATH and CodeForces, and $k = 8$ for LCB. Pass@1 is then calculated as

$$\text{pass@1} = \frac{1}{k}\sum_{i=1}^{k} p_i,$$

where p_i denotes the correctness of the i-th response. This method provides more reliable performance estimates. For AIME 2024, we also report consensus (majority vote) results using 64 samples, denoted as cons@64.

D.2. Main Results

Table 8 | Comparison between DeepSeek-R1 and other representative models. Numbers in bold denote the performance is statistically significant (t−test with $p < 0.01$).

	Benchmark (Metric)	**Claude-3.5-Sonnet-1022**	**GPT-4o 0513**	**DeepSeek V3**	**OpenAI o1-mini**	**OpenAI o1-1217**	**DeepSeek R1**
	Architecture	-	-	MoE	-	-	MoE
	# Activated Params	-	-	37B	-	-	37B
	# Total Params	-	-	671B	-	-	671B
English	MMLU (EM)	88.3	87.2	88.5	85.2	**91.8**	90.8
	MMLU-Redux (EM)	88.9	88.0	89.1	86.7	-	**92.9**
	MMLU-Pro (EM)	78.0	72.6	75.9	80.3	-	**84.0**
	DROP (3-shot F1)	88.3	83.7	91.6	83.9	90.2	**92.2**
	IF-Eval (Prompt Strict)	**86.5**	84.3	86.1	84.8	-	83.3
	GPQA Diamond (Pass@1)	65.0	49.9	59.1	60.0	**75.7**	71.5
	SimpleQA (Correct)	28.4	38.2	24.9	7.0	**47.0**	30.1
	FRAMES (Acc.)	72.5	80.5	73.3	76.9	-	**82.5**
	AlpacaEval2.0 (LC-winrate)	52.0	51.1	70.0	57.8	-	**87.6**
	ArenaHard (GPT-4-1106)	85.2	80.4	85.5	92.0	-	92.3
Code	LiveCodeBench (Pass@1-COT)	38.9	32.9	36.2	53.8	63.4	**65.9**
	Codeforces (Percentile)	20.3	23.6	58.7	93.4	96.6	96.3
	Codeforces (Rating)	717	759	1134	1820	2061	2029
	SWE Verified (Resolved)	**50.8**	38.8	42.0	41.6	48.9	49.2
	Aider-Polyglot (Acc.)	45.3	16.0	49.6	32.9	**61.7**	53.3
Math	AIME 2024 (Pass@1)	16.0	9.3	39.2	63.6	79.2	79.8
	MATH-500 (Pass@1)	78.3	74.6	90.2	90.0	96.4	97.3
	CNMO 2024 (Pass@1)	13.1	10.8	43.2	67.6	-	**78.8**
Chinese	CLUEWSC (EM)	85.4	87.9	90.9	89.9	-	**92.8**
	C-Eval (EM)	76.7	76.0	86.5	68.9	-	**91.8**
	C-SimpleQA (Correct)	55.4	58.7	**68.0**	40.3	-	63.7

Standard Benchmark We evaluate DeepSeek-R1 on multiple benchmarks. For education-oriented knowledge benchmarks such as MMLU, MMLU-Pro, and GPQA Diamond, DeepSeek-R1 demonstrates superior performance compared to DeepSeek-V3. This improvement is primarily attributed to enhanced accuracy in STEM-related questions, where significant gains are achieved through large-scale reinforcement learning. Additionally, DeepSeek-R1 excels on FRAMES, a long-context-dependent QA task, showcasing its strong document analysis capabilities. This highlights the potential of reasoning models in AI-driven search and data analysis tasks.

DeepSeek-R1 also delivers impressive results on IF-Eval, a benchmark designed to assess a model's ability to follow format instructions. These improvements can be linked to the inclusion of instruction-following data during the final stages of SFT and RL training. Furthermore, remarkable performance is observed on AlpacaEval2.0 and ArenaHard, indicating DeepSeek-R1's strengths in writing tasks and open-domain question answering.

On math tasks, DeepSeek-R1 demonstrates performance on par with OpenAI-o1-1217, surpassing other models by a large margin. A similar trend is observed on coding algorithm tasks, such as LiveCodeBench and Codeforces, where reasoning-focused models dominate these benchmarks. On engineering-oriented coding tasks, OpenAI-o1-1217 outperforms DeepSeek-R1 on Aider but achieves comparable performance on SWE Verified. We believe the engineering performance of DeepSeek-R1 will improve in the next version, as the amount of related RL training data currently remains very limited.

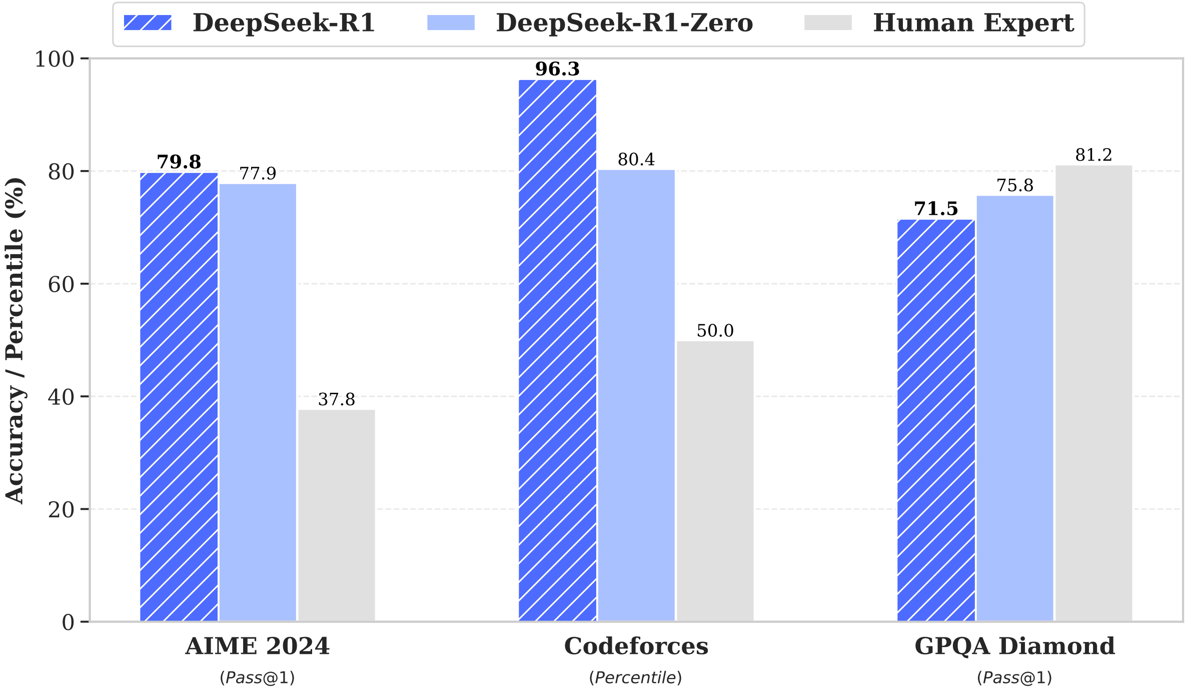

Figure 10 | The benchmark performance of DeepSeek-R1 and DeepSeek-R1-Zero is compared with human scores across different datasets. For AIME and Codeforces, the human scores represent the average performance of all human competitors. In the case of GPQA, the human score corresponds to Ph.D.-level individuals who had access to the web for answering the questions.

Figure 10 presents a comparative analysis of the performance of DeepSeek-R1-Zero, DeepSeek-R1, and human participants across several benchmark competitions. Notably, the AIME is a mathematics competition designed for high school students, and DeepSeek-R1 demonstrates performance that surpasses the mean score achieved by human competitors in this event. On the Codeforces platform, DeepSeek-R1 outperforms 96.3% of human participants, underscoring its advanced problem-solving capabilities. In the case of GPQA, where human experts—typically individuals with Ph.D.-level qualifications and access to web resources—participate, human performance remains superior to that of DeepSeek-R1. However, we anticipate that enabling web access for DeepSeek-R1 could substantially enhance its performance on GPQA, potentially narrowing or closing the observed gap.