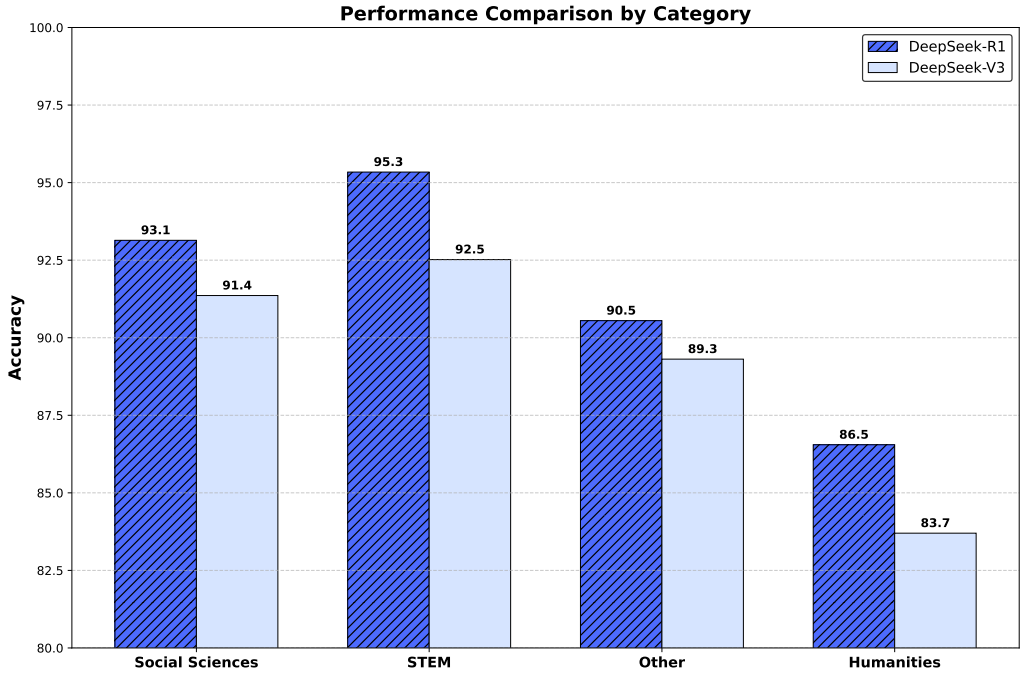

Figure 15 | The comparison of DeepSeek-V3 and DeepSeek-R1 across MMLU categories.

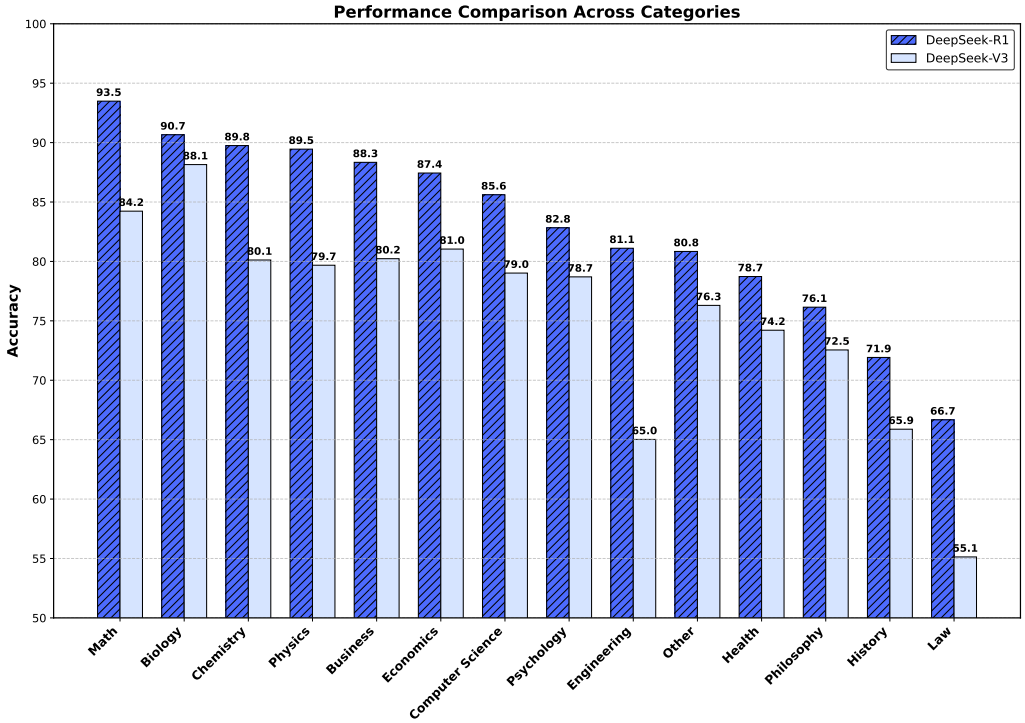

Figure 16 | The comparison of DeepSeek-V3 and DeepSeek-R1 across MMLU-Pro categories.

Table 12 | A Comparative Analysis of DeepSeek-V3 and DeepSeek-R1. DeepSeek-V3 is a non-reasoning model developed on top of DeepSeek-V3-Base, which also serves as the foundational base model for DeepSeek-R1. Numbers in bold denote the performance is statistically significant (t−test with $p < 0.01$).

	Benchmark (Metric)	**V3-Base**	**V3**	**R1-Zero**	**R1**
English	MMLU (EM)	87.1	88.5	88.8	**90.8**
	MMLU-Redux (EM)	86.2	89.1	85.6	**92.9**
	MMLU-Pro (EM)	64.4	75.9	68.9	**84.0**
	DROP (3-shot F1)	89.0	91.6	89.1	**92.2**
	IF-Eval (Prompt Strict)	58.6	**86.1**	46.6	83.3
	GPQA Diamond (Pass@1)	-	59.1	**75.8**	71.5
	SimpleQA (Correct)	20.1	24.9	30.3	30.1
	FRAMES (Acc.)	-	73.3	82.3	82.5
	AlpacaEval2.0 (LC-winrate)	-	70.0	24.7	**87.6**
	ArenaHard (GPT-4-1106)	-	85.5	53.6	**92.3**
Code	LiveCodeBench (Pass@1-COT)	-	36.2	50.0	**65.9**
	Codeforces (Percentile)	-	58.7	80.4	**96.3**
	Codeforces (Rating)	-	1134	1444	**2029**
	SWE Verified (Resolved)	-	42.0	43.2	**49.2**
	Aider-Polyglot (Acc.)	-	49.6	12.2	**53.3**
Math	AIME 2024 (Pass@1)	-	39.2	77.9	**79.8**
	MATH-500 (Pass@1)	-	90.2	95.9	**97.3**
	CNMO 2024 (Pass@1)	-	43.2	**88.1**	78.8
Chinese	CLUEWSC (EM)	82.7	90.9	93.1	92.8
	C-Eval (EM)	90.1	86.5	**92.8**	91.8
	C-SimpleQA (Correct)	-	**68.0**	66.4	63.7

benchmark. In contrast, DeepSeek-V3 shows a relative advantage in instruction-following capabilities, suggesting different optimization priorities between the two models.

To further elucidate the specific knowledge domains that benefit most from post-training, we conduct a fine-grained analysis of model performance across various subject categories within MMLU and MMLU-Pro. These categories, predefined during the construction of the test sets, allow for a more systematic assessment of domain-specific improvements.

As illustrated in Figure 16, performance improvements on MMLU-Pro are observed across all domains, with particularly notable gains in STEM-related categories such as mathematics and physics. Similarly, on MMLU, the largest improvements from DeepSeek-V3 to DeepSeek-R1 are also observed in STEM domains. However, unlike MMLU-Pro, gains in the STEM domain are smaller, suggesting differences in the impact of post-training between the two benchmarks.

Our hypothesis is that MMLU represents a relatively easier challenge compared to MMLU-Pro. In STEM tasks of MMLU, post-training on DeepSeek-V3 may have already achieved near-saturation performance, leaving minimal room for further improvement in DeepSeek-R1. It surprised us that the non-STEM tasks, such as social sciences and humanities, are improved with the long CoT, which might attribute to the better understanding of the question.

Table 13 | Performance on latest math competitions. Participants with their USAMO index (`AMC score` + 10 × `AIME score`) surpassing 251.5 are qualified for USAMO.

Average Score	**AMC 12 2024**	**AIME 2025**	**USAMO Index**
Human Participants	61.7	6.2/15	123.7
GPT-4o 0513	84.0	2.0/15	104.0
DeepSeek V3	98.3	3.3/15	131.3
OpenAI o1-1217	141.0	12.0/15	261.0
DeepSeek R1	143.7	11.3/15	256.7

E.2. Generalization to Real-World Competitions

Despite rigorous efforts to eliminate data contamination, variations of test set questions or discussions of related problems may still exist on websites that were included in the pre-training corpus. This raises an important question: can DeepSeek-R1 achieve comparable performance on test sets that were released after its training? To investigate this, we evaluate our model on AIME 2025, providing insights into its generalization capabilities on unseen data. As shown in Table 13, in AIME 2025 (`https://artofproblemsolving.com/wiki/index.php/2025_AIME_II_Problems`), DeepSeek-R1 achieves a 75% solve rate (Pass@1), approaching o1's performance of 80%. Most notably, the model attains a score of 143.7/150 in AMC 12 2024 (`https://artofproblemsolving.com/wiki/index.php/2024_AMC_12B_Problems`) - a performance that, when combined with its AIME results, yields a score exceeding the qualification threshold for attending the USAMO (United States of America Mathematical Olympiad `https://artofproblemsolving.com/wiki/index.php/AMC_historical_results?srsltid=AfmBOoqQ6pQic5NCan_NX1wYgr-aoHgJ33hsq7KSekF-rUwY8TBaBao` 1). This performance positions DeepSeek-R1 among the nation's top-tier high school students.

E.3. Mathematical Capabilities Breakdown by Categories

To assess DeepSeek-R1's mathematical reasoning capabilities comprehensively, we evaluated its performance across diverse categories of quantitative reasoning problems. Our test set comprised 366 problems drawn from 93 mathematics competitions held in 2024 (`https://artofproblemsolving.com/community/c3752401_2024_contests`), including mathematical olympiads and team selection tests. As shown in Figure 17, DeepSeek-R1 significantly outperforms the representative non-reasoning model GPT-4o 0513. DeepSeek-R1 demonstrates relatively strong proficiency in number theory and algebra, while exhibiting considerable room for improvement in geometry and combinatorics.

E.4. An Analysis on CoT Length

Adaptive CoT length: During training, DeepSeek-R1 was permitted to think for a long time (i.e., to generate a lengthy chain of thought) before arriving at a final solution. To maximize success on challenging reasoning tasks, the model learned to dynamically scale computation by generating more thinking tokens to verify or correct its reasoning steps, or to backtrack and explore alternative approaches when initial attempts proved unsuccessful. The complexity of a problem directly correlates with the number of thinking tokens required: more difficult problems typically demand more extensive computation. For extremely easy questions, like 1 + 1 =?, the model tends to use fewer tokens (< 100 tokens) to answer the question.