DeepSeek and Beyond: A Comparative Analysis of AI Model Architectures

Introduction

In a significant shift from conventional language model development, DeepSeek has emerged with a groundbreaking approach that challenges traditional AI training methods. This technical analysis delves into DeepSeek's revolutionary architecture, which turns the standard fine-tuning pipeline on its head by prioritizing reinforcement learning before supervised training – a departure that promises to reshape our understanding of AI model development.

At the heart of this report lies a detailed examination of DeepSeek's performance across the AI landscape, benchmarking its capabilities against industry giants like Claude, GPT-4, and the emerging Kimi k1.5. From mastering complex mathematical reasoning to handling intricate coding challenges, the data reveals a nuanced picture of strengths and trade-offs among today's leading AI models.

Beyond theoretical frameworks, the analysis offers practical insights for organizations considering AI deployment, exploring options from lightweight 1.5B parameter models to robust 32B parameter versions. With comprehensive cost analysis and performance metrics spanning multiple providers, this report serves as a crucial resource for understanding the current state of AI capabilities and their practical implications in the rapidly evolving landscape of artificial intelligence.

DeepSeek Model Availability

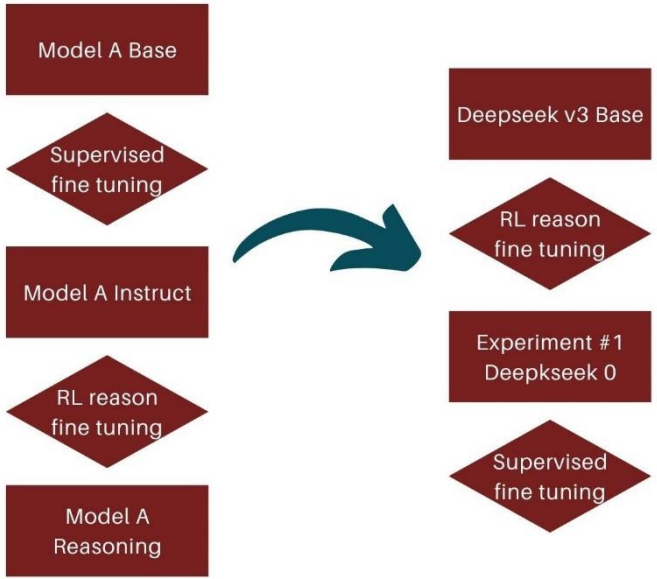
DeepSeek models are open-source and accessible through multiple platforms, allowing flexible deployment based on hardware and cost considerations.

Platforms Offering DeepSeek Models:

- **Groq**
  - High-speed inference.
  - Paid service with proprietary hardware acceleration.
- **AnythingLLM & LLM Studio**
  - Offer free versions for public use.
  - Require a GPU for inference—compute load is fully offloaded to the GPU.
- **Ollama *(Recommended for local deployment)***
  - Supports both CPU and GPU, making it ideal for users without a dedicated GPU.
  - Offers a free, offline-friendly environment for model execution.
- **Best Choice for Different Use Cases:**
  - For GPU users → AnythingLLM / LLM Studio (Leverage GPU acceleration).
  - For CPU users → Ollama (Efficient CPU inference, no GPU required).

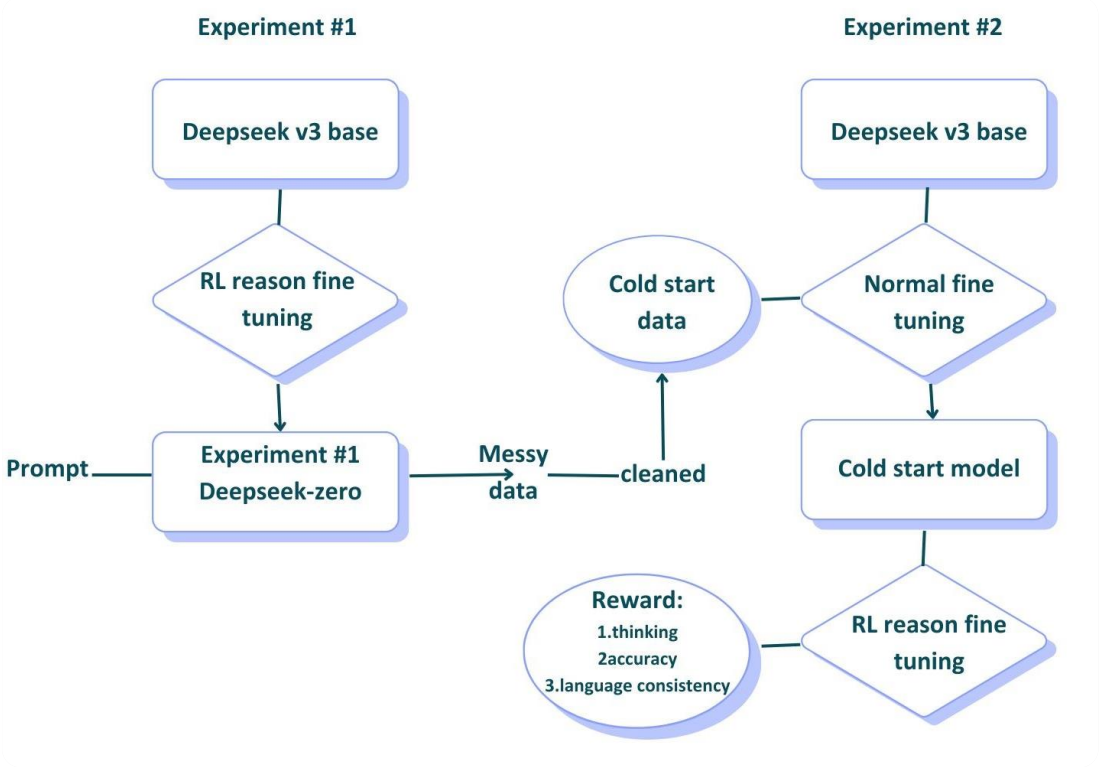
Traditional Approach vs. DeepSeek's Novel Approach

Traditional Approach ("The Usual")	
Base Model	The model starts with a pretrained base
Supervised Fine-Tuning (SFT)	The base model undergoes SFT to become an instruction-following model
RL Fine-Tuning	Reinforcement learning is applied to refine reasoning capabilities
Final Model	The model achieves improved reasoning and performance



DeepSeek Approach ("Experiment #1")	
Base Model (DeepSeek-V3)	DeepSeek-V3 is used as the foundation
RL-Based Reasoning Fine-Tuning	Unlike the traditional approach, DeepSeek first applies reinforcement learning for reasoning tasks
DeepSeek-Zero	The model obtained after RL fine-tuning is called DeepSeek-Zero
Supervised Fine-Tuning	DeepSeek-Zero then undergoes supervised fine-tuning to further refine its responses and alignment

DeepSeek Alternative Approach ("Experiment #2")

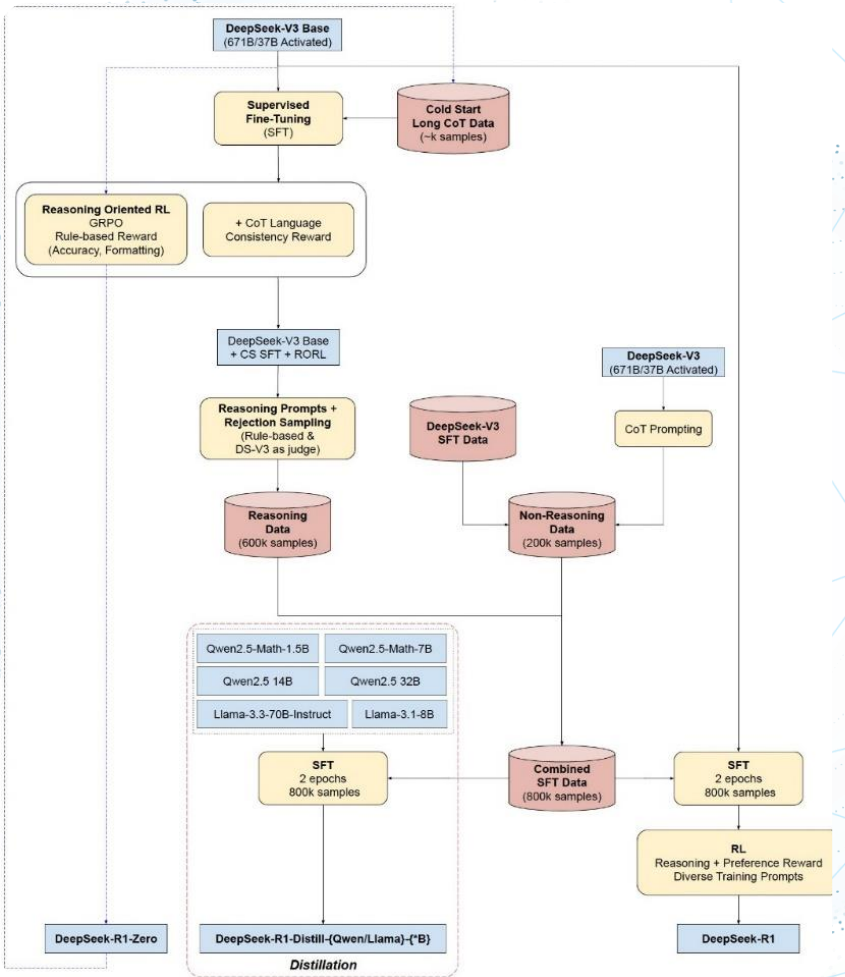


1. **Base Model (DeepSeek-V3)**: The same base model is used.
2. **Supervised Fine-Tuning First**: Instead of RL fine-tuning first, this approach applies supervised fine-tuning at the start.
3. **Cold Start Model**: A fine-tuned instruction-following model is obtained.
4. **RL-Based Reasoning Fine-Tuning**: After obtaining the cold start model, reinforcement learning is applied at the final stage

DeepSeek Approach ("Experiment #2")

The Role of Cold Start Data

Cold Start Data refers to a small dataset (~thousands of samples) of high-quality, manually curated reasoning examples. This dataset is introduced at the beginning of training to improve the model's performance in reasoning tasks and guide its early-stage learning.



Key Differences and Advantages

Earlier Reinforcement Learning: DeepSeek applies RL before supervised fine-tuning, allowing the model to develop reasoning capabilities before being refined with SFT.

Cold Start Data Utilization: The integration of Cold Start Data ensures that the model has a strong foundation in reasoning tasks before exposure to broader training data.

Reasoning-Oriented RL: Instead of applying RL purely for reward-based optimization later in the pipeline, DeepSeek prioritizes reasoning fine-tuning at an earlier stage.

Performance Scaling by Model Size

DeepSeek Models for Private or Local Use

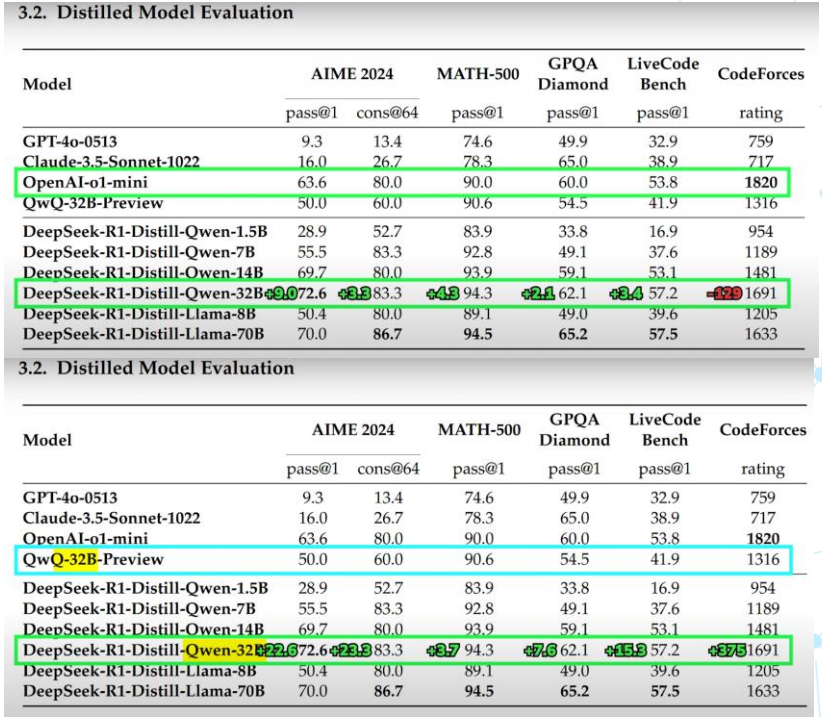
We have access to the DeepSeek-R1 model with 404GB, but this is not viable for most local deployments. Instead, we have distilled versions of the model, including Llama-based and Qwen-based models, available in various parameter sizes and quantization levels.

Smallest Viable Model: The lowest available model starts with 1.5B parameters, which can be quantized to 4-bit, making it approximately 1.5GB in size.

Performance Scaling: As model size increases, performance improves. Around 32B parameters, the models begin to match or surpass GPT-0I Mini in performance.

The attached image demonstrates that as the model size increases, particularly around 32B parameters, we see marked improvements in:

- AIME 2024 pass rates (+9.0 boost for 32B Qwen Distill)
- MATH-500 accuracy (+4.3 boost)
- GPQA Diamond benchmark (+2.1 boost)
- LiveCode Bench accuracy (+3.4 boost)



3.2. Distilled Model Evaluation

Model	AIME 2024		MATH-500	GPQA Diamond	LiveCode Bench	CodeForces
	pass@1	cons@64	pass@1	pass@1	pass@1	rating
GPT-4o-0513	9.3	13.4	74.6	49.9	32.9	759
Claude-3.5-Sonnet-1022	16.0	26.7	78.3	65.0	38.9	717
OpenAI-o1-mini	63.6	80.0	90.0	60.0	53.8	**1820**
QwQ-32B-Preview	50.0	60.0	90.6	54.5	41.9	1316
DeepSeek-R1-Distill-Qwen-1.5B	28.9	52.7	83.9	33.8	16.9	954
DeepSeek-R1-Distill-Qwen-7B	55.5	83.3	92.8	49.1	37.6	1189
DeepSeek-R1-Distill-Qwen-14B	69.7	80.0	93.9	59.1	53.1	1481
DeepSeek-R1-Distill-Qwen-32B	+9.0 72.6	+3.3 83.3	+4.3 94.3	+2.1 62.1	+3.4 57.2	-120 1691
DeepSeek-R1-Distill-Llama-8B	50.4	80.0	89.1	49.0	39.6	1205
DeepSeek-R1-Distill-Llama-70B	70.0	**86.7**	**94.5**	**65.2**	**57.5**	1633

3.2. Distilled Model Evaluation

Model	AIME 2024		MATH-500	GPQA Diamond	LiveCode Bench	CodeForces
	pass@1	cons@64	pass@1	pass@1	pass@1	rating
GPT-4o-0513	9.3	13.4	74.6	49.9	32.9	759
Claude-3.5-Sonnet-1022	16.0	26.7	78.3	65.0	38.9	717
OpenAI-o1-mini	63.6	80.0	90.0	60.0	53.8	**1820**
QwQ-32B-Preview	50.0	60.0	90.6	54.5	41.9	1316
DeepSeek-R1-Distill-Qwen-1.5B	28.9	52.7	83.9	33.8	16.9	954
DeepSeek-R1-Distill-Qwen-7B	55.5	83.3	92.8	49.1	37.6	1189
DeepSeek-R1-Distill-Qwen-14B	69.7	80.0	93.9	59.1	53.1	1481
DeepSeek-R1-Distill-Qwen-32B	+22.6 72.6	+23.3 83.3	+3.7 94.3	+7.6 62.1	+15.3 57.2	+375 1691
DeepSeek-R1-Distill-Llama-8B	50.4	80.0	89.1	49.0	39.6	1205
DeepSeek-R1-Distill-Llama-70B	70.0	**86.7**	**94.5**	**65.2**	**57.5**	1633

Benchmarking & Model Performance Analysis

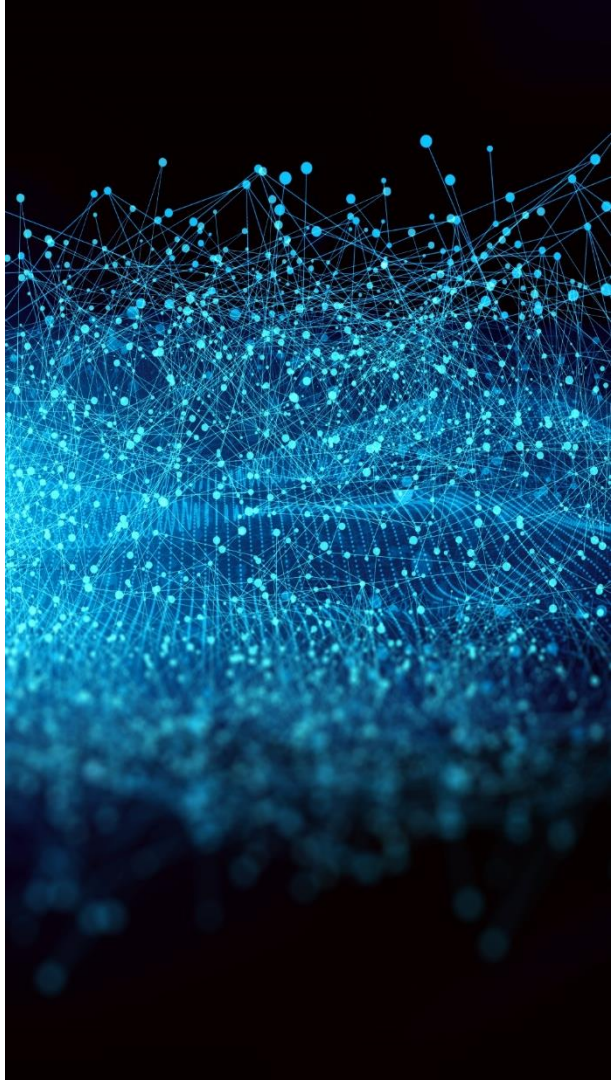
API Provider	Model	Context window	Latency(first chunk) (sec)	Input token price	Output token price	Output speed(Median) (token/sec)	Blended price (USD/1M tokens)	License
Deepseek	Deepseek R1	128k	27.52	$0.55	$2.19	62	$0.96	Open
Together.ai	Deepseek LLM 67B	4k	0.48	$0.90	$0.90	28	$0.90	Open
Deepseek	Deepseek V3	65.5k	1.13	$0.27	$1.1	50	$0.48	Open
Together.ai		128k	0.62	$1.25	$1.25	15	$1.25	Open
Azure	o1	200k	31.09	$15	$60	33	$26.25	Proprietary
Azure	o1-mini	128k	13.5	$3.3	$13.20	75	$5.78	Proprietary
Openai		128k	11.08	$3	$12	208	$5.25	Proprietary
Azure	Gpt-4o	128k	0.98	$2.50	$10	144	$4.38	Proprietary
Openai		128k	0.38	$2.50	$10	115	$4.38	Proprietary
Azure	Gpt-4o mini	128k	0.79	$0.15	$0.60	178	$0.26	Proprietary
Openai		128k	0.4	$0.15	$0.60	78	$0.26	Proprietary
Anthropic	Claude 3 haiku	200k	0.41	$0.25	$1.25	137	$0.50	Proprietary
Aws		200k	0.79	$0.25	$1.25	106	$0.50	Proprietary
Anthropic	Claude 3 sonnet	200k	0.84	$3	$15	84	$6	Proprietary
Aws		200k	0.86	$3	$15	43	$6	Proprietary
Anthropic	Claude 3 opus	200k	1.97	$15	$75	28	$30	Proprietary
Google vertex		200k	2.38	$15	$75	27	$30	Proprietary
Aws		200k	1.42	$15	$75	24	$30	Proprietary
Anthropic	Claude 3.5 haiku	200k	0.83	$0.8	$4	65	$1.6	Proprietary
Google vertex		200k	0.95	$0.8	$4	65	$1.6	Proprietary
Aws standard/Aws optimized		200k	0.85/0.58	$0.8/$1	$4/$5	54/100	$1.6/$2	Proprietary
Anthropic	Claude 3.5 sonnet	200k	1.18	$3	$15	84	$6	Proprietary
Google vertex		200k	0.85	$3	$15	73	$6	Proprietary
Aws		200k	1.03	$3	$15	44	$6	Proprietary
Aws standard/Aws optimized	Llama 3.1 405B	128k	1.95/0.81	$2.4/$3	$2.4/$3	30/65	$2.4/$3	Open
Aws	Llama 3.3 70B	128k	0.94	$0.71	$0.71	31	$0.71	Open
Google AI studio	Gemini 1.5 pro	2m	0.76	$1.25	$5	63	$2.19	Proprietary
Google Ai vertex		2m	0.41	$1.25	$5	59	$2.19	Proprietary

Model Benchmarks including Kimi

Benchmark (Metric)	Kimi k1.5	Claude-3.5-Sonnet-1022	GPT-4o 0513	DeepSeek V3	OpenAI o1-mini	OpenAI o1-1217	DeepSeek R1
Text							
MMLU (Pass@1)	87.4	88.3	87.2	88.5	85.2	91.8	90.8
MMLU-Redux (EM)	-	88.9	88	89.1	86.7	-	92.9
MMLU-Pro (EM)	-	78	72.6	75.9	80.3	-	84
DROP (3-shot F1)	-	88.3	83.7	91.6	83.9	90.2	92.2
IF-Eval (Prompt Strict)	**87.2**	86.5	84.3	86.1	84.8	-	83.3
GPQA Diamond (Pass@1)	-	65	49.9	59.1	60	75.7	71.5
SimpleQA (Correct)	-	28.4	38.2	24.9	7	47	30.1
FRAMES (Acc.)	-	72.5	80.5	73.3	76.9	-	**82.5**
AlpacaEval2.0 (LC-winrate)	-	52	51.1	70	57.8	-	**87.6**
ArenaHard (GPT-4-1106)	-	85.2	80.4	85.5	92	-	**92.3**
Code							
LiveCodeBench (Pass@1-CoT)	88.3	38.9	32.9	36.2	53.8	63.4	65.9
Codeforces (Percentile)	94	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	41.6	48.9	49.2
Aider-Polyglot (Acc.)	-	45.3	16	49.6	32.9	**61.7**	53.3
Math							
AIME 2024 (Pass@1)	**60.8**	16	9.3	39.2	63.6	79.2	**79.8**
MATH-500 (EM)	96.2	78.3	74.6	90.2	90	**96.4**	97.3
CNMO 2024 (Pass@1)	-	13.1	10.8	43.2	67.6	-	**78.8**
Vision							
MathVista-Test (Pass@1)	**74.9**	-	-	-	71.4	71	70.1
MMMU-Val (Pass@1)	**70**	-	-	-	70.3	**77.3**	68
MathVision-Full (Pass@1)	**38.6**	-	-	-	35.9	-	31
Chinese							
CLUEWSC (EM)	**91.7**	85.4	87.9	90.9	89.9	-	**92.8**
C-Eval (EM)	**88.3**	76.7	76	86.5	68.9	-	**91.8**
C-SimpleQA (Correct)	-	55.4	58.7	**68**	40.3	-	63.7

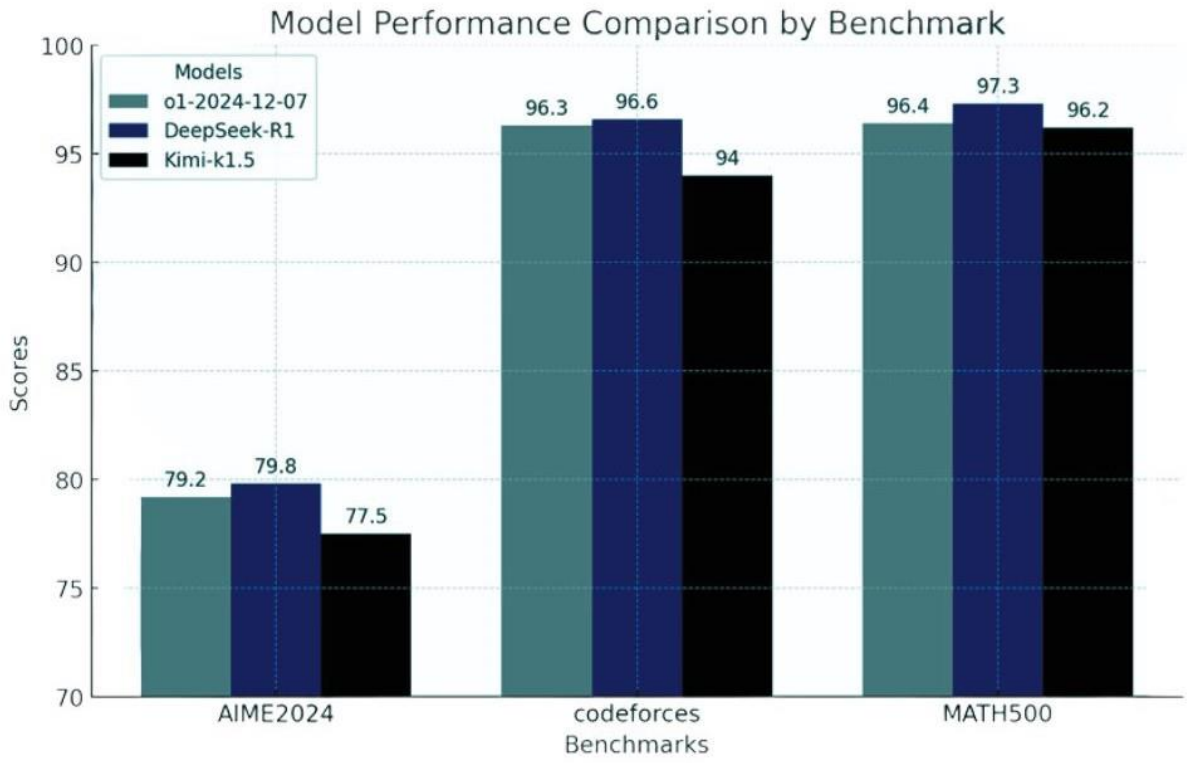
Where DeepSeek V3 is Better

1. **General Knowledge & Text Reasoning**
    - Highest on **MMLU (88.5)** and **MMLU-Redux (89.1)** (except DeepSeek R1).
    - Strong performance on **AlpacaEval 2.0 (70.0)**, indicating better alignment in instruction-following.
2. **Mathematical Reasoning**
    - **CNMO 2024 (43.2)** – Significantly better than Kimi (not reported).
3. **Reading Comprehension & Logical Reasoning**
    - **DROP (91.6 F1)** – Best in dataset assessing multi-step numerical and reading comprehension



Where Kimi K1.5 is Better

1. **Coding**
    - **LiveCodeBench (88.3)** – Significantly higher, suggesting strong real-world code generation.
    - **Codeforces Percentile (94)** – Second only to OpenAI's models.
2. **Mathematical Problem Solving**
    - **AIME 2024 (60.8)** – Higher than DeepSeek V3 (39.2).
    - **MATH-500 (96.2)** – Nearly top-tier, surpassing DeepSeek V3 (90.2)
3. **Vision-Based Math**
    - **MathVista-Test (74.9)** – Stronger ability in vision-based mathematical reasoning.
    - **MathVision-Full (38.6)** – Higher than DeepSeek V3 (not reported).



Conclusion

- **DeepSeek R1 and OpenAI o1-1217** dominate the table with the highest scores across multiple benchmarks.
- **Kimi K1.5 and DeepSeek V3** have specific strengths but are less consistent overall.
- **Claude 3.5-Sonnet and GPT-4o** remain strong contenders but do not lead in as many areas.

References

- https://arxiv.org/pdf/2501.12948
- https://arxiv.org/pdf/2501.12599v1
- https://x.com/SirrahChan/status/1881540279783887036
- https://artificialanalysis.ai/
- https://www.youtube.com/watch?v=CiS9gDfYZ-w&ab_channel=bycloud