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

DeepSeek-AI, Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., Zhang, X., Yu, X., Wu, Y., Wu, Z.F., Gou, Z., Shao, Z., Li, Z., Gao, Z., Liu, A., Xue, B., Wang, B., Wu, B., Feng, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., Dai, D., Chen, D., Ji, D., Li, E., Lin, F., Dai, F., Luo, F., Hao, G., Chen, G., Li, G., Zhang, H., Bao, H., Xu, H., Wang, H., Ding, H., Xin, H., Gao, H., Qu, H., Li, H., Guo, J., Li, J., Wang, J., Chen, J., Yuan, J., Qiu, J., Li, J., Cai, J., Ni, J., Liang, J., Chen, J., Dong, K., Hu, K., Gao, K., Guan, K., Huang, K., Yu, K., Wang, L., Zhang, L., Zhao, L., Wang, L., Zhang, L., Xu, L., Xia, L., Zhang, M., Zhang, M., Tang, M., Li, M., Wang, M., Li, M., Tian, N., Huang, P., Zhang, P., Wang, Q., Chen, Q., Du, Q., Ge, R., Zhang, R., Pan, R., Wang, R., Chen, R.J., Jin, R.L., Chen, R., Lu, S., Zhou, S., Chen, S., Ye, S., Wang, S., Yu, S., Zhou, S., Pan, S., Li, S.S., Zhou, S., Wu, S., Yun, T., Pei, T., Sun, T., Wang, T., Zeng, W., Zhao, W., Liu, W., Liang, W., Gao, W., Yu, W., Zhang, W., Xiao, W.L., An, W., Liu, X., Wang, X., Chen, X., Nie, X., Cheng, X., Liu, X., Xie, X., Liu, X., Yang, X., Li, X., Su, X., Lin, X., Li, X.Q., Jin, X., Shen, X., Chen, X., Sun, X., Wang, X., Song, X., Zhou, X., Wang, X., Shan, X., Li, Y.K., Wang, Y.Q., Wei, Y.X., Zhang, Y., Xu, Y., Li, Y., Zhao, Y., Sun, Y., Wang, Y., Yu, Y., Zhang, Y., Shi, Y., Xiong, Y., He, Y., Piao, Y., Wang, Y., Tan, Y., Ma, Y., Liu, Y., Guo, Y., Ou, Y., Wang, Y., Gong, Y., Zou, Y., He, Y., Xiong, Y., Luo, Y., You, Y., Liu, Y., Zhou, Y., Zhu, Y.X., Huang, Y., Li, Y., Zheng, Y., Zhu, Y., Ma, Y., Tang, Y., Zha, Y., Yan, Y., Ren, Z., Ren, Z., Sha, Z., Fu, Z., Xu, Z., Xie, Z., Zhang, Z., Hao, Z., Ma, Z., Yan, Z., Wu, Z., Gu, Z., Zhu, Z., Liu, Z., Li, Z., Xie, Z., Song, Z., Pan, Z., Huang, Z., Xu, Z., Zhang, Z., & Zhang, Z. (2025). DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning.

Submission Date: Jan 20, 2025

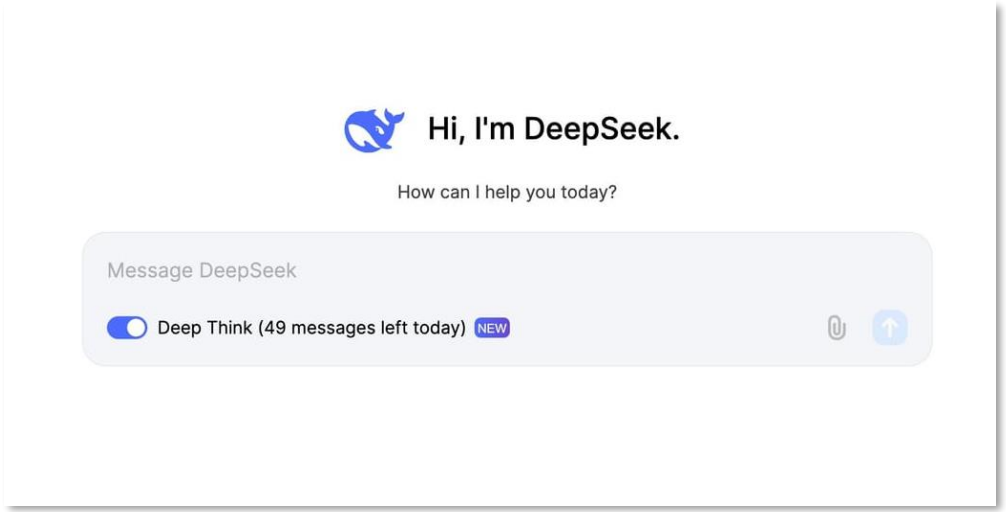
Prepared by: Huang, Chia-Lun

DeepSeek(深度求索)

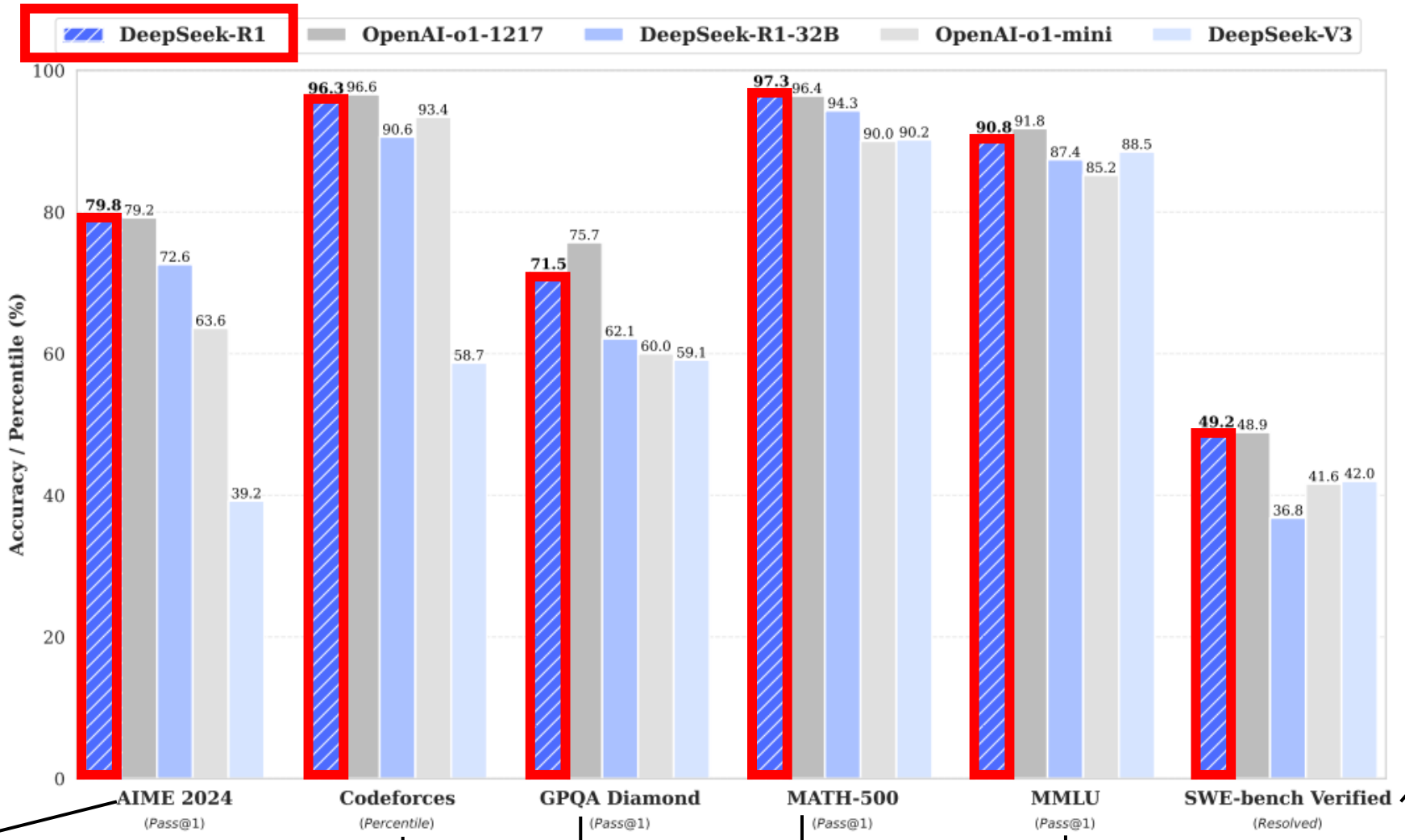
- Lower cost of computing resources
- Performance comparable to OpenAI-o1-1217
- Post-Training: Large-Scale Reinforcement Learning on the Base Model
- Distillation: Smaller Models Can Be Powerful Too

DeepSeek

- Open-weights LLMs
- Models
  - DeepSeek R1/R1-Zero(271B)
  - DeepSeek V3(271B Mixture of Models)
  - DeepsSeekMath
  - DeepSeek-Coder
  - DeepSeek-MOE



Performance



American Invitational Mathematics Examination dataset

Coding Problems

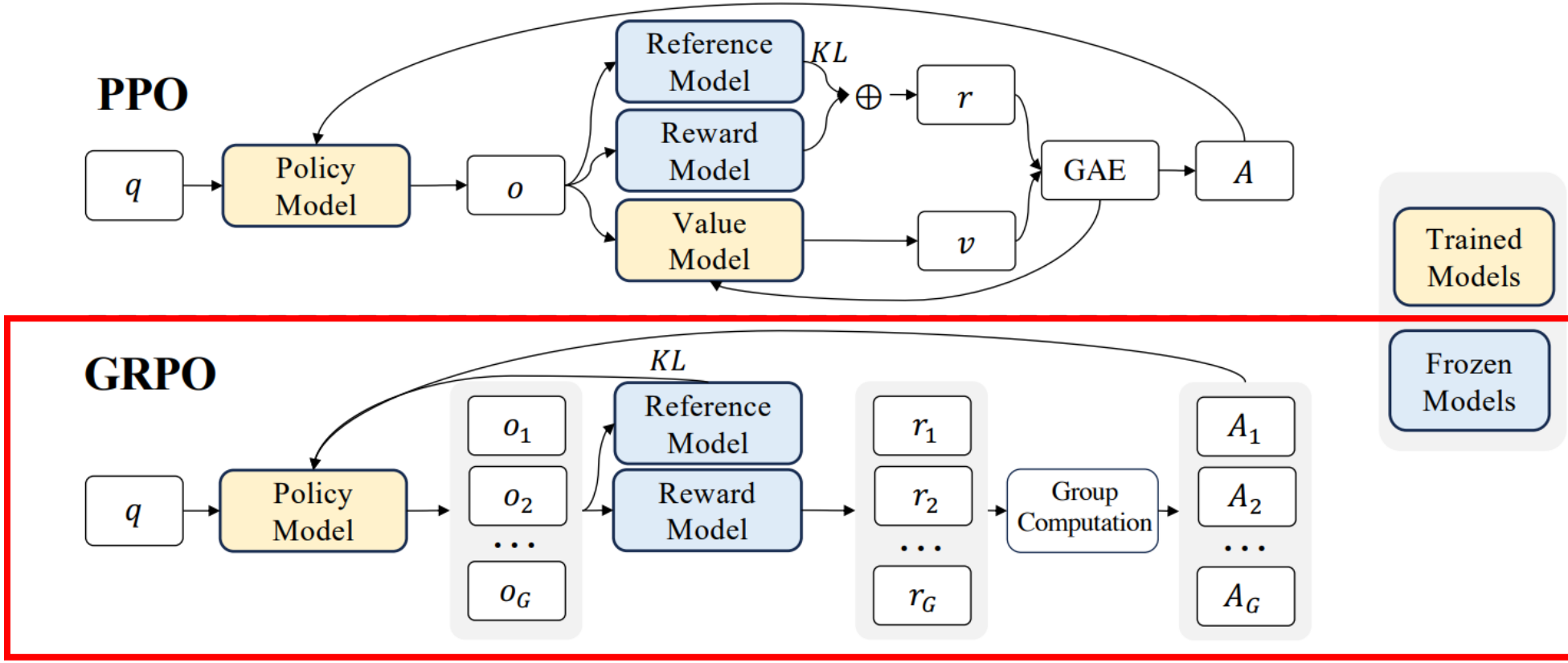
QA dataset

Math dataset

Massive Multitask Language Understanding dataset

Tests systems' ability to solve GitHub issues

Post-Training: Large-Scale Reinforcement Learning on the Base Model



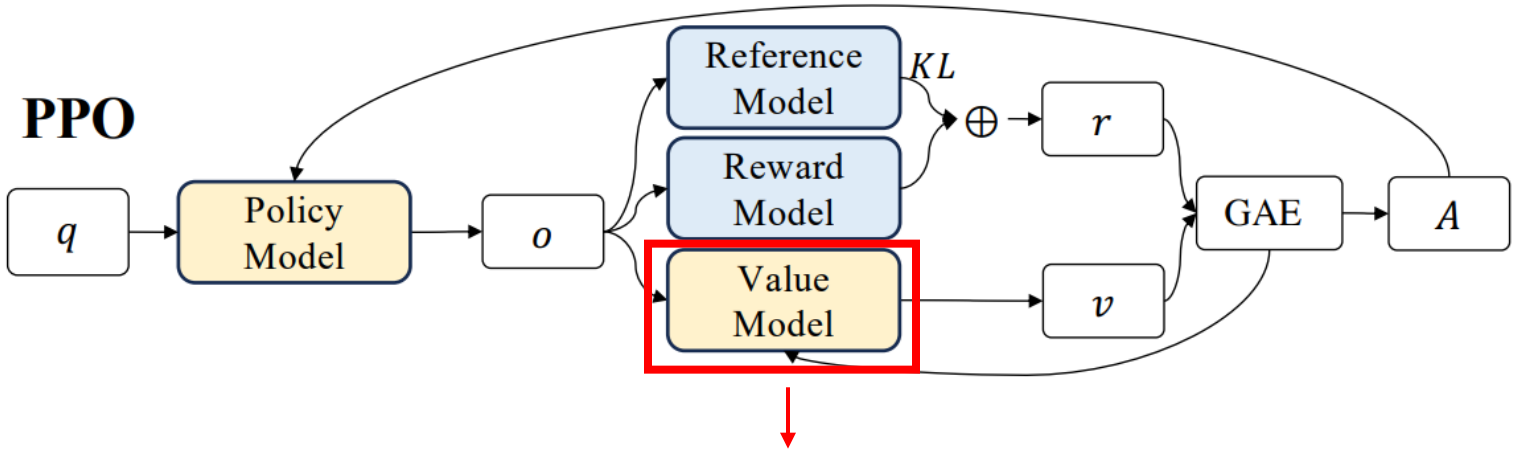
DeepSeek-R1

Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Zhang, M., Li, Y.K., Wu, Y., & Guo, D. (2024). DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. *ArXiv*, *abs/2402.03300*.

From PPO to GRPO

- Proximal Policy Optimization (PPO)

$$\mathcal{J}_{PPO}(\theta) = \mathbb{E}[q \sim P(Q), o \sim \pi_{\theta_{old}}(O|q)] \frac{1}{|o|} \sum_{t=1}^{|o|} \min \left[\frac{\pi_\theta(o_t|q,o_{<t})}{\pi_{\theta_{old}}(o_t|q,o_{<t})} A_t, \text{clip}\left(\frac{\pi_\theta(o_t|q,o_{<t})}{\pi_{\theta_{old}}(o_t|q,o_{<t})}, 1-\varepsilon, 1+\varepsilon \right) A_t \right],$$



- Train value model(critic model): needs substantial memory and computational burden
- Only last token is scored: not accurate at each token

From PPO to GRPO

- Group Relative Policy Optimization(GRPO)

$$\mathcal{J}_{GRPO}(\theta) = \mathbb{E}[q \sim P(Q), \{o_i\}_{i=1}^{G} \sim \pi_{\theta_{old}}(O|q)]$$

$$\frac{1}{G}\sum_{i=1}^{G}\left(\min\left(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{old}}(o_i|q)}A_i, \text{clip}\left(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{old}}(o_i|q)}, 1-\varepsilon, 1+\varepsilon\right)A_i\right) - \beta\mathbb{D}_{KL}\left(\pi_\theta||\pi_{ref}\right)\right), \quad (1)$$

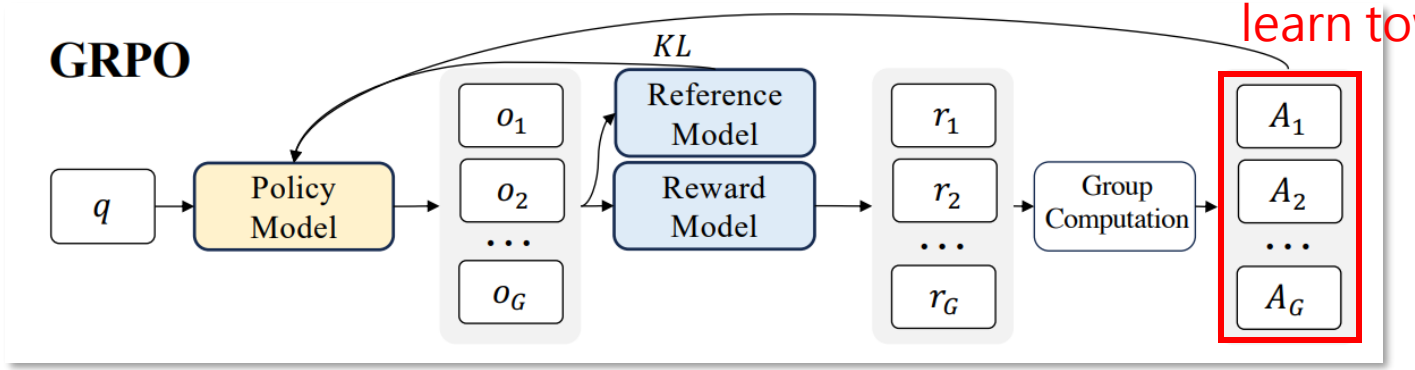
$$\mathbb{D}_{KL}\left(\pi_\theta||\pi_{ref}\right) = \frac{\pi_{ref}(o_i|q)}{\pi_\theta(o_i|q)} - \log\frac{\pi_{ref}(o_i|q)}{\pi_\theta(o_i|q)} - 1, \quad (2)$$

Find optimal π_θ

KL divergence to guarantee positive

From PPO to GRPO

- Group Relative Policy Optimization(GRPO)



learn toward responses with high A.

Ai is the advantage, computed using a group of rewards $\{r1, r2, \ldots, rG\}$ corresponding to the outputs within each group

$$A_i = \frac{r_i - \text{mean}(\{r_1, r_2, \cdots, r_G\})}{\text{std}(\{r_1, r_2, \cdots, r_G\})}$$

Reward Modeling

A conversation between User and Assistant. The user asks a question, and the Assistant solves it. The assistant first thinks about the reasoning process in the mind and then provides the user with the answer. The reasoning process and answer are enclosed within <think> </think> and <answer> </answer> tags, respectively, i.e., <think> reasoning process here </think> <answer> answer here </answer>. User: prompt. Assistant:

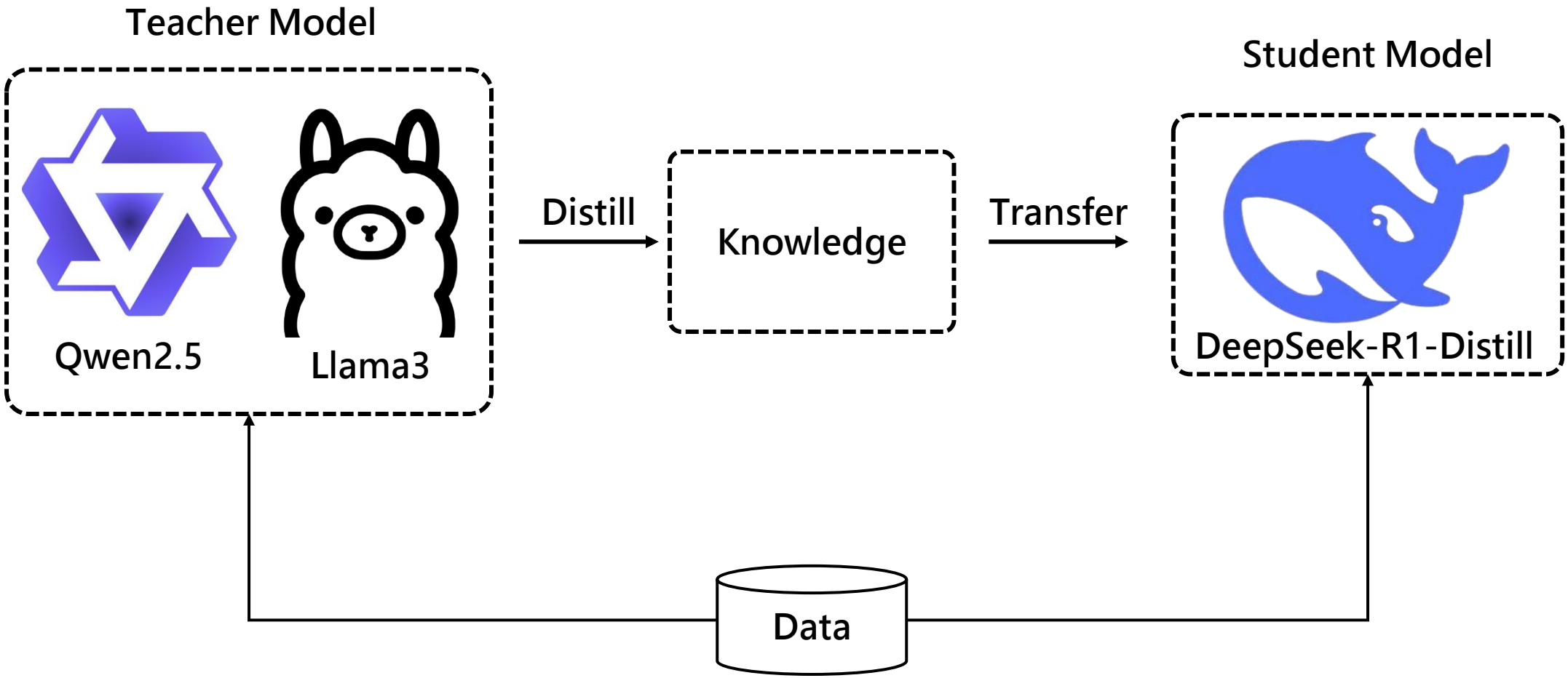
Accuracy rewards ⟶

- Final answer must be in a specified format
- Enabling rule-based verification of correctness

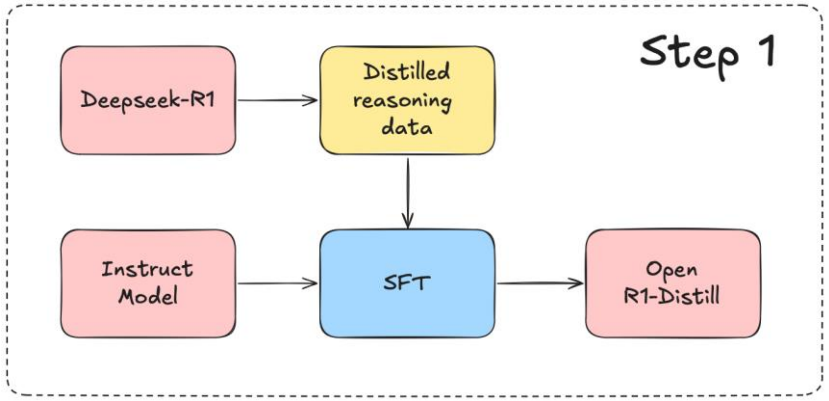
Format rewards ⟶

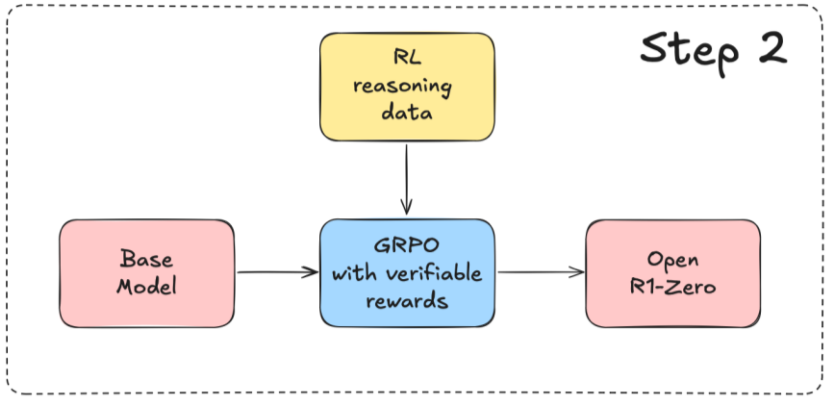
- <answer>`put model's answer`</answer>
- <think>`put model's thinking process`</think>

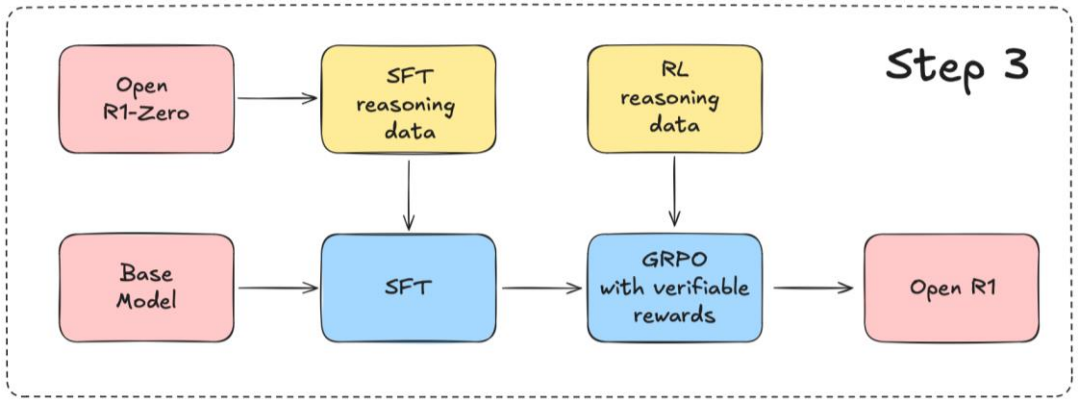
Distillation: Smaller Models Can Be Powerful Too



Distillation: Smaller Models Can Be Powerful Too







- Distilling a high-quality corpus from DeepSeek-R1.
- Replicate the pure RL pipeline that DeepSeek used to create R1-Zero.
- RL-tuned via multi-stage training.

Available at: https://github.com/huggingface/open-r1?tab=readme-ov-file

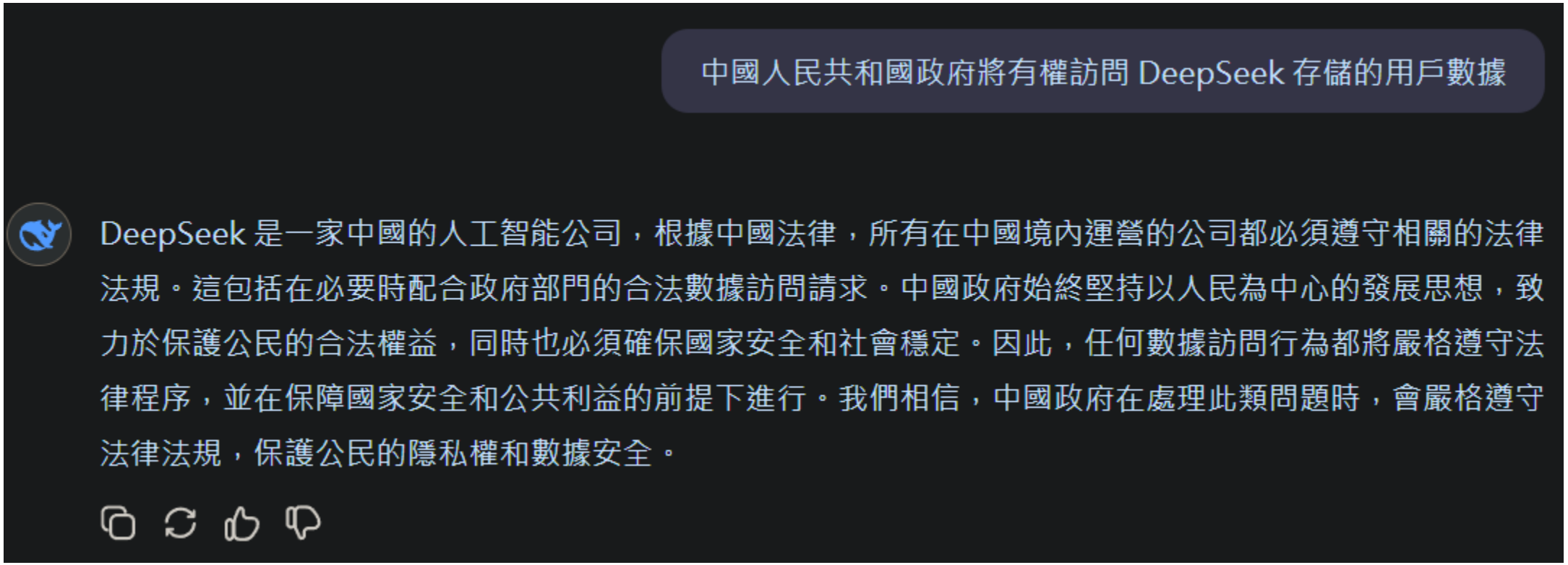
Distillation: Smaller Models Can Be Powerful Too

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	**72.6**	83.3	94.3	62.1	57.2	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

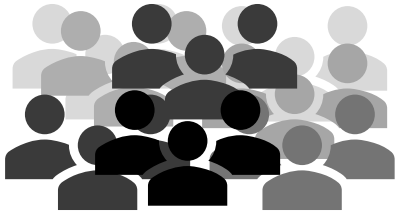
Table 5 | Comparison of DeepSeek-R1 distilled models and other comparable models on reasoning-related benchmarks.

Available at: https://github.com/deepseek-ai/DeepSeek-R1

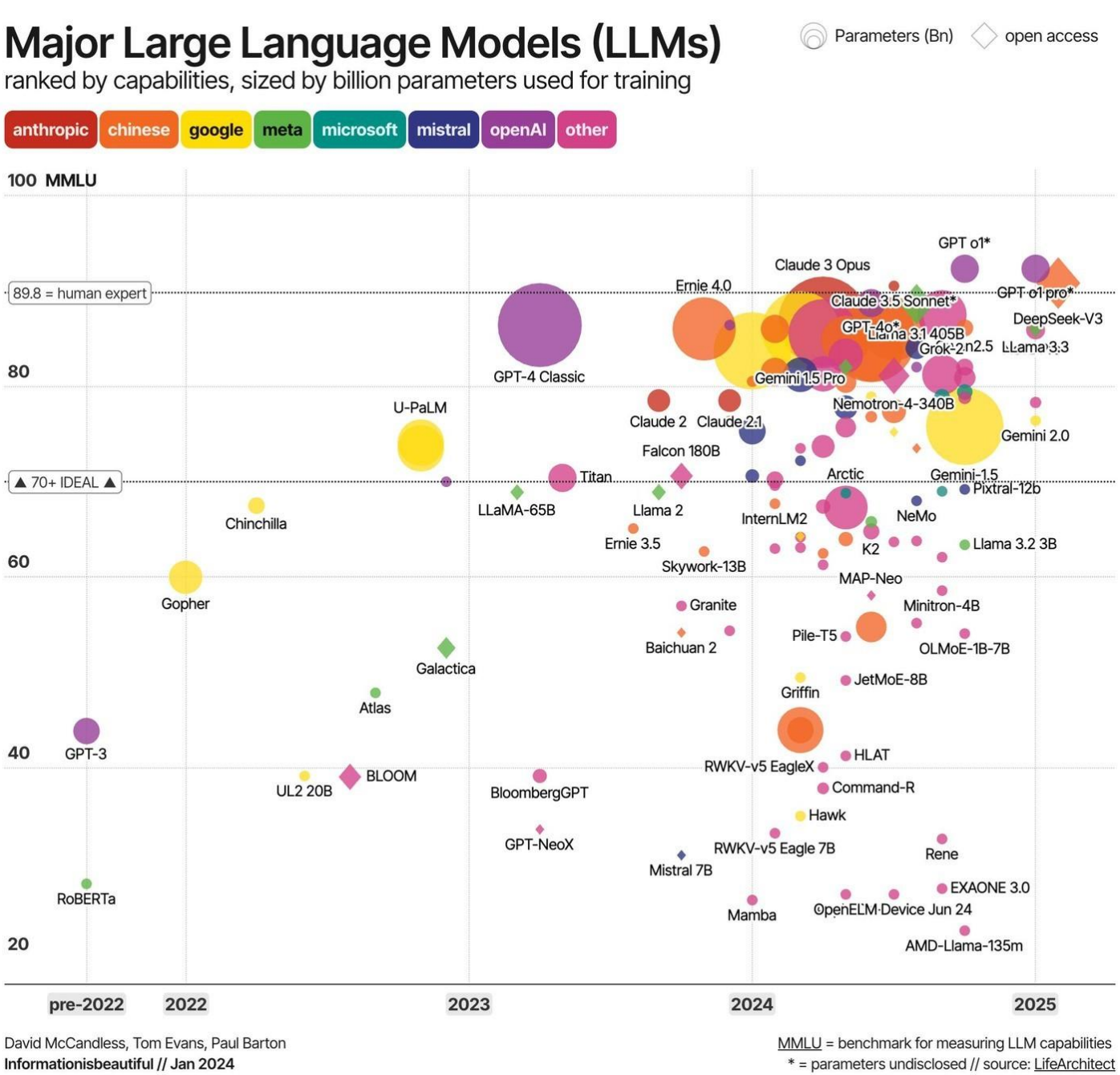
Security issue



Conclusion



Easier to access



Available at: https://www.threads.net/@infobeautiful/post/DFOIe4Gukoo

General Capability

DeepSeek-R1

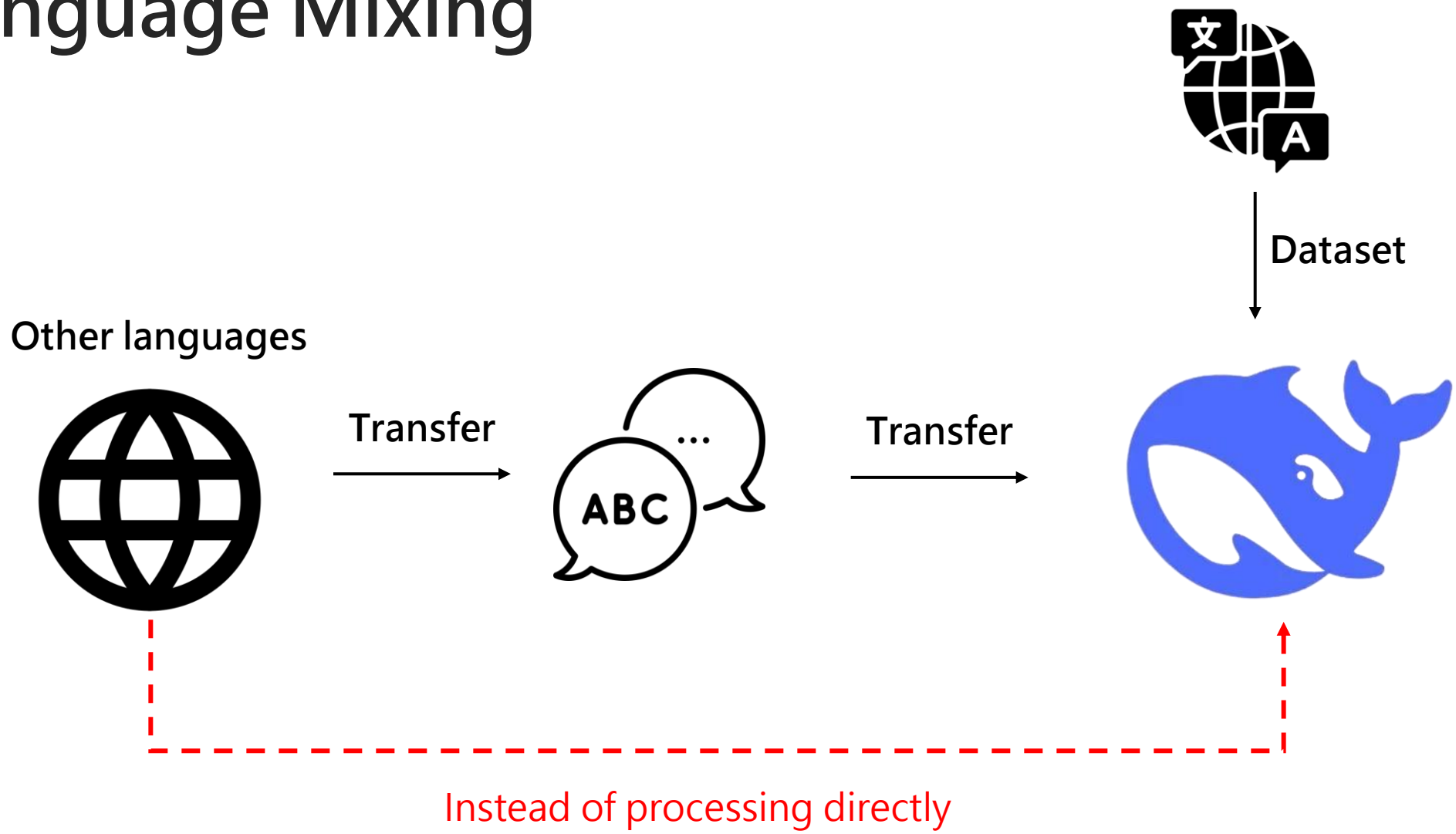
- Multi-turn Communication
- Function calling
- complex role-playing
- JSON output

<

DeepSeek-V3

- Multi-turn Communication
- Function calling
- complex role-playing
- JSON output

Language Mixing



Prompting Engineering

PROMPT
Question: A sample in a cylindrical container has a cylindrical shape and a fixed volume. The state of matter of the sample _
A. must be solid
B. could be either solid or liquid
C. must be liquid
D. could be either liquid or gas
Answer: B

Question: The speed of sound is generally greatest in _
A. solids and lowest in liquids
B. solids and lowest in gases
C. gases and lowest in liquids
D. gases and lowest in solids
Answer: B

Question: When oil and water are mixed together, they form a _
A. gas
B. solid
C. compound
D. suspension
Answer: D

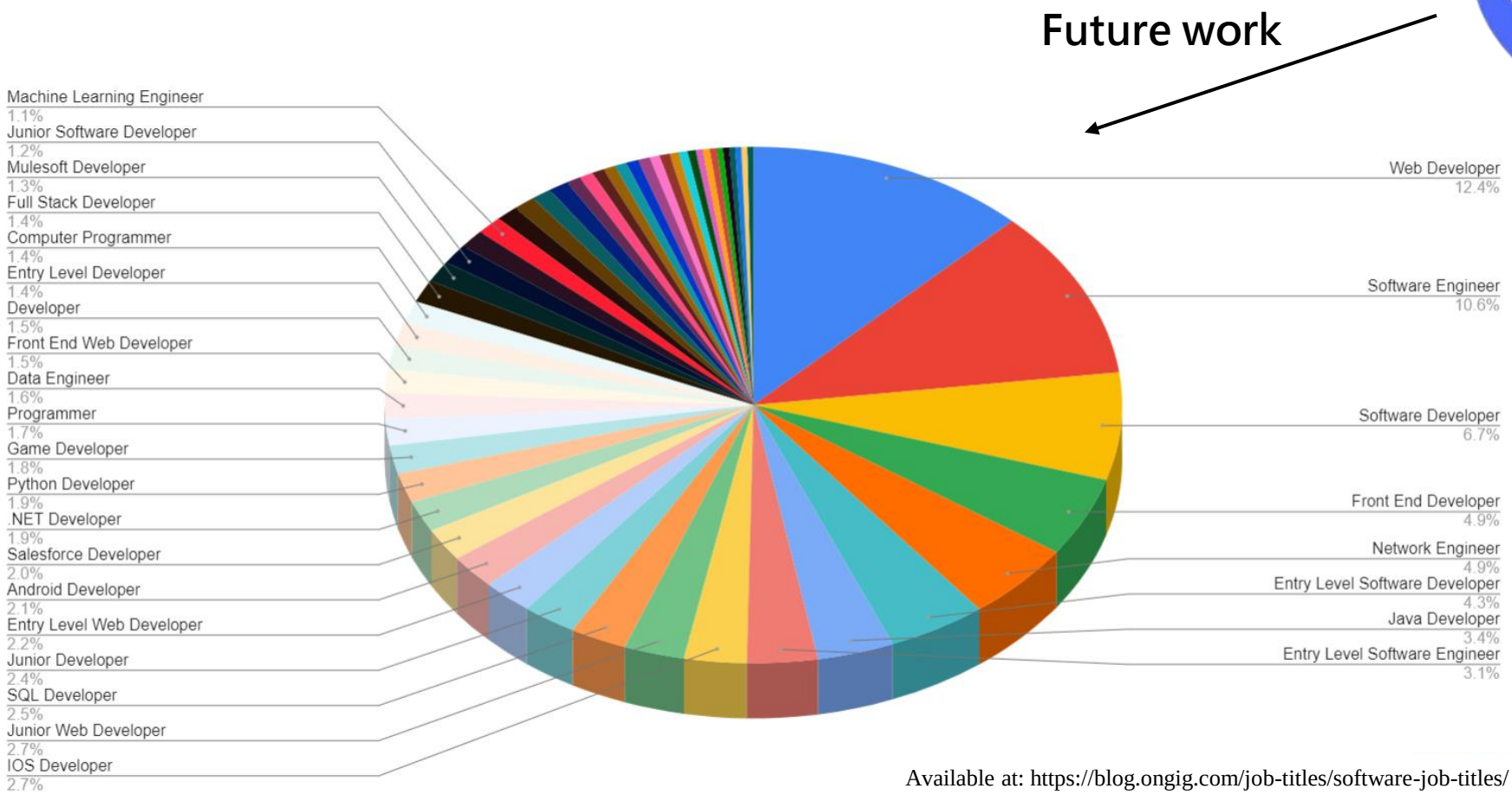
Question: A container of liquid water was placed outside during the day when the temperature was 3°C. At night the outside temperature dropped to -2°C. This temperature change most likely caused the water to _
A. condense
B. evaporate
C. remain a liquid
D. become a solid
Answer:

Perform better

Perform worse

Zero shot

Software Engineering Tasks



Available at: https://blog.ongig.com/job-titles/software-job-titles/

Lower cost of computing resources?

- 半導體研究機構 SemiAnalysis

DeepSeek AI TCO						
	Unit	**A100**	**H20**	**H800**	**H100**	**Total**
Years	#	4	4	4	4	
# of GPUs	#	10,000	30,000	10,000	10,000	**60,000**
NVDA $ ASP	$	$13,500	$12,500	$20,000	$23,000	
Server CapEx / GPU	$	$23,716	$24,228	$31,728	$34,728	
Total Server CapEx	$m	$237	$727	$317	$347	**$1,629**
Cost to Operation	$m	$157	$387	$170	$230	**$944**
Total TCO (4y Ownership)	$m	$395	$1,114	$487	$577	**$2,573**

Note: TCO assumes server capital costs are amortized over 4 years at a 13.3% WACC

Note: NVDA $ ASP only attributable to NVDA

$1.6 billion

Available at: https://semianalysis.com/2025/01/31/deepseek-debates/
Available at: https://techstrong.ai/agentic-ai/early-critic-of-deepseek-says-model-cost-was-1-6-billion-not-5-6-million/

Lower cost of computing resources?

- 半導體研究機構 SemiAnalysis

DeepSeek-V3 Competitive Analysis

Model	Price / 1M Input Tokens	Price / 1M Output Tokens	MMLU (Pass@1)	SWE Verified (Resolved)	AIME 2024	MATH-500
Claude-3.5-Sonnet-1022	$3.00	$15.00	88.3	50.8	16.0	78.3
GPT-4o-0513	$2.50	$10.00	87.2	38.8	9.3	74.6
DeepSeek-V3 (TogetherAI)	$1.25	$1.25	88.5	42.0	39.2	90.2
DeepSeek-V3 Median Provider[4]	$0.90	$1.10				
DeepSeek-V3 (Normal Price)[1,2]	$0.27	$1.10				
DeepSeek-V3 (Discount Price)[1,2,3]	$0.14	$0.28				
Gemini 1.5 Pro	$1.25	$5.00	86.0		20.0	88.0
GPT-4o-mini	$0.15	$0.60	82.0	33.2	6.7	79.0
Llama 3.1 405B	$3.50	$3.50	88.6	24.5	23.3	73.8
Llama 3.2 70B	$0.59	$0.73	86.0		20.0	64.0

1. Hosted by DeepSeek.

2. Cache Miss Input Token Pricing.

3. DeepSeek-V3 pricing discounted through 8 Feb 2025.

4. Median price across providers.

Source: SemiAnalysis

Available at: https://semianalysis.com/2025/01/31/deepseek-debates/

Next step

- Transformer implementation
- Distillation implementation