

Towards Large Reasoning Models as a Judge

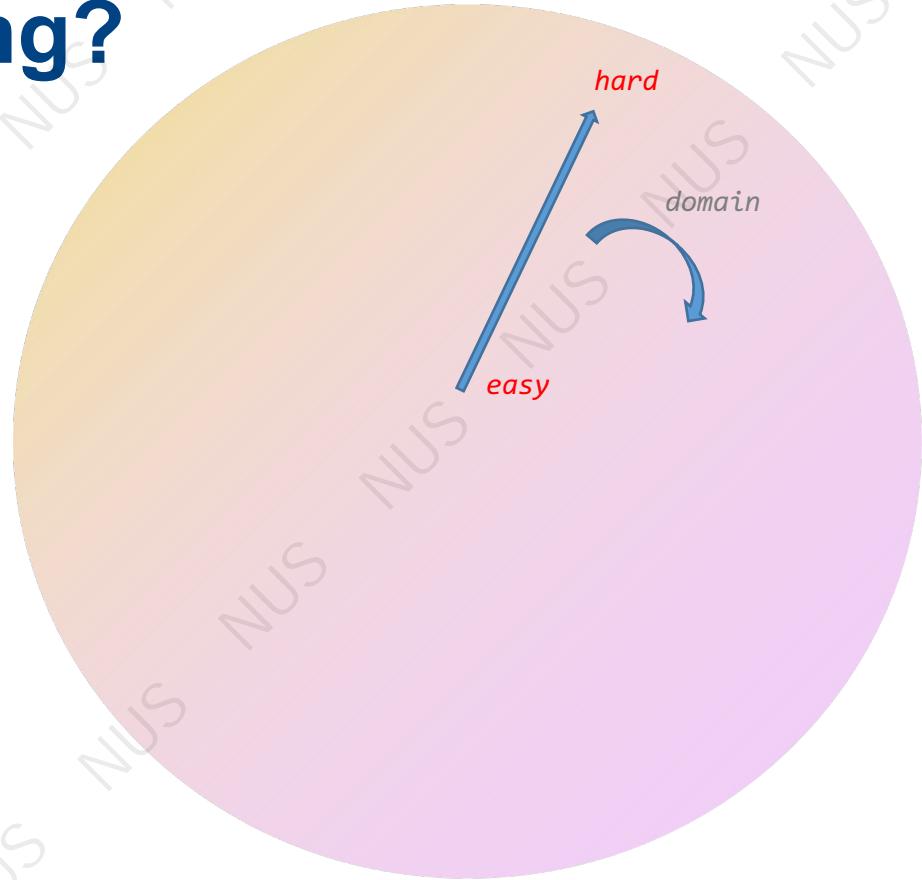


NUS
National University
of Singapore

National University of Singapore

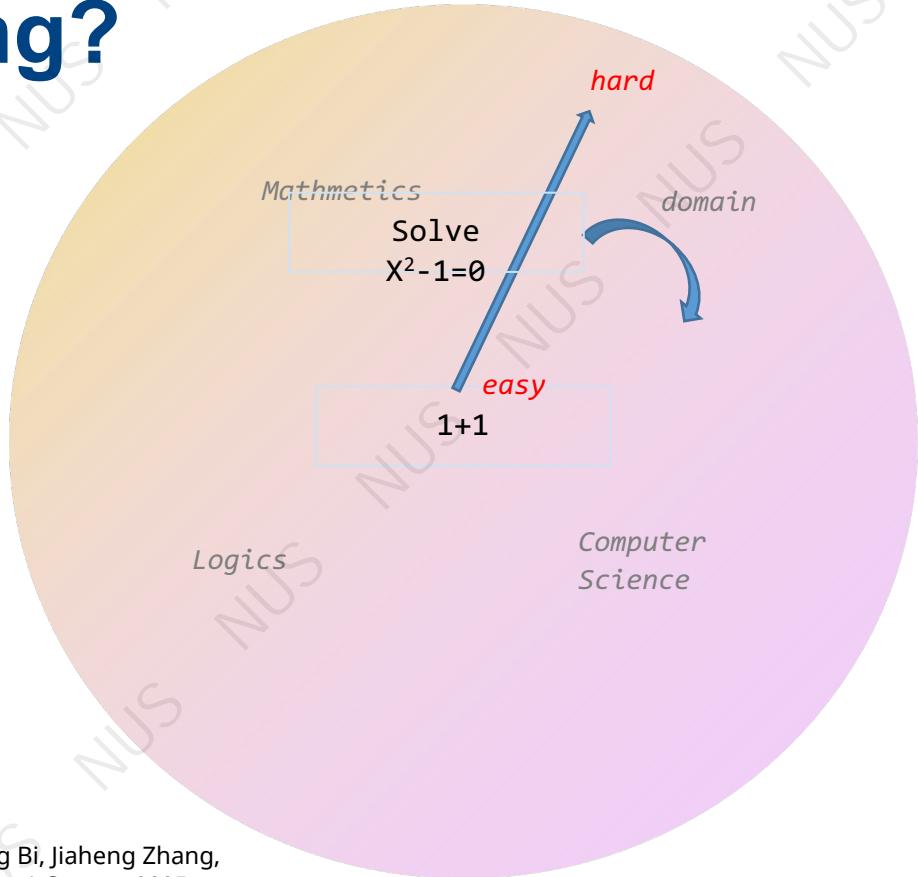
| What is Reasoning?

- The space of problems



| What is Reasoning?

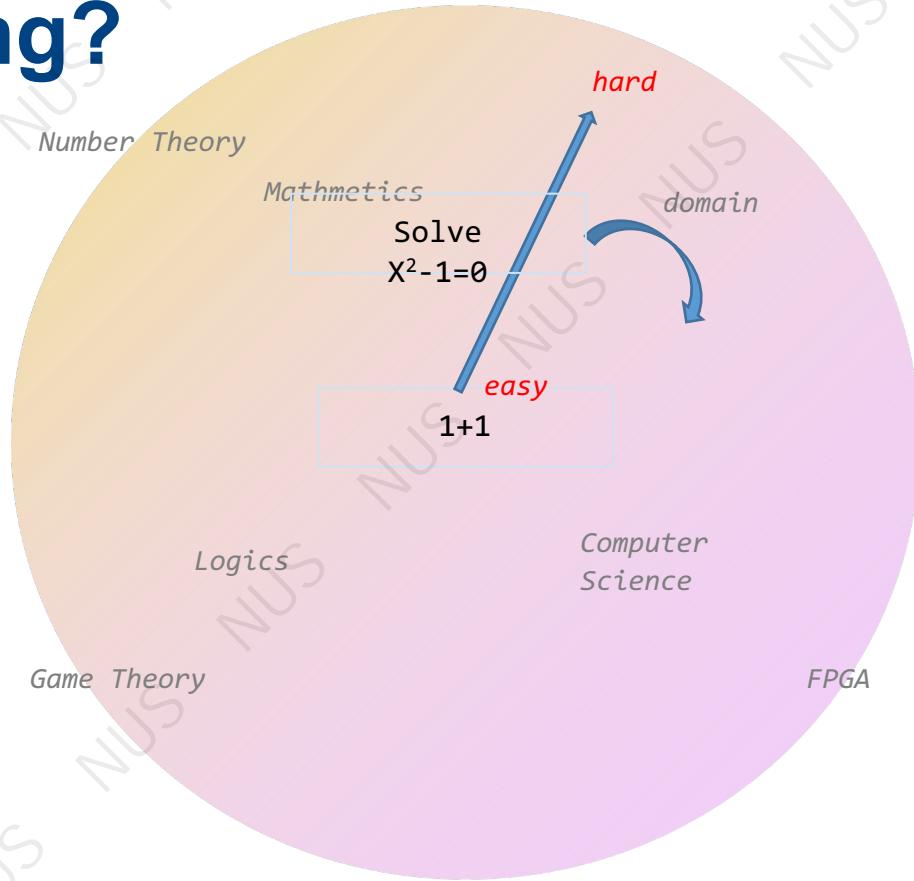
- ▶ The space of problems
 - ▶ Mathematics
 - ▶ GSM8K-Zero, SVAMP, AQuA, ASDiv, MathQA, TheoremQA...
 - ▶ Logics
 - ▶ ProntoQA, LogiQA, Reclor...
 - ▶ Code
 - ▶ LiveCodeBench, Codeforces, SWE-Bench...
 - ▶ ...



[1] Yue Liu, Jiaying Wu, Yufei He, Hongcheng Gao, Hongyu Chen, Baolong Bi, Jiaheng Zhang, Zhiqi Huang, Bryan Hooi. Efficient Inference for Large Reasoning Models: A Survey. 2025.

| What is Reasoning?

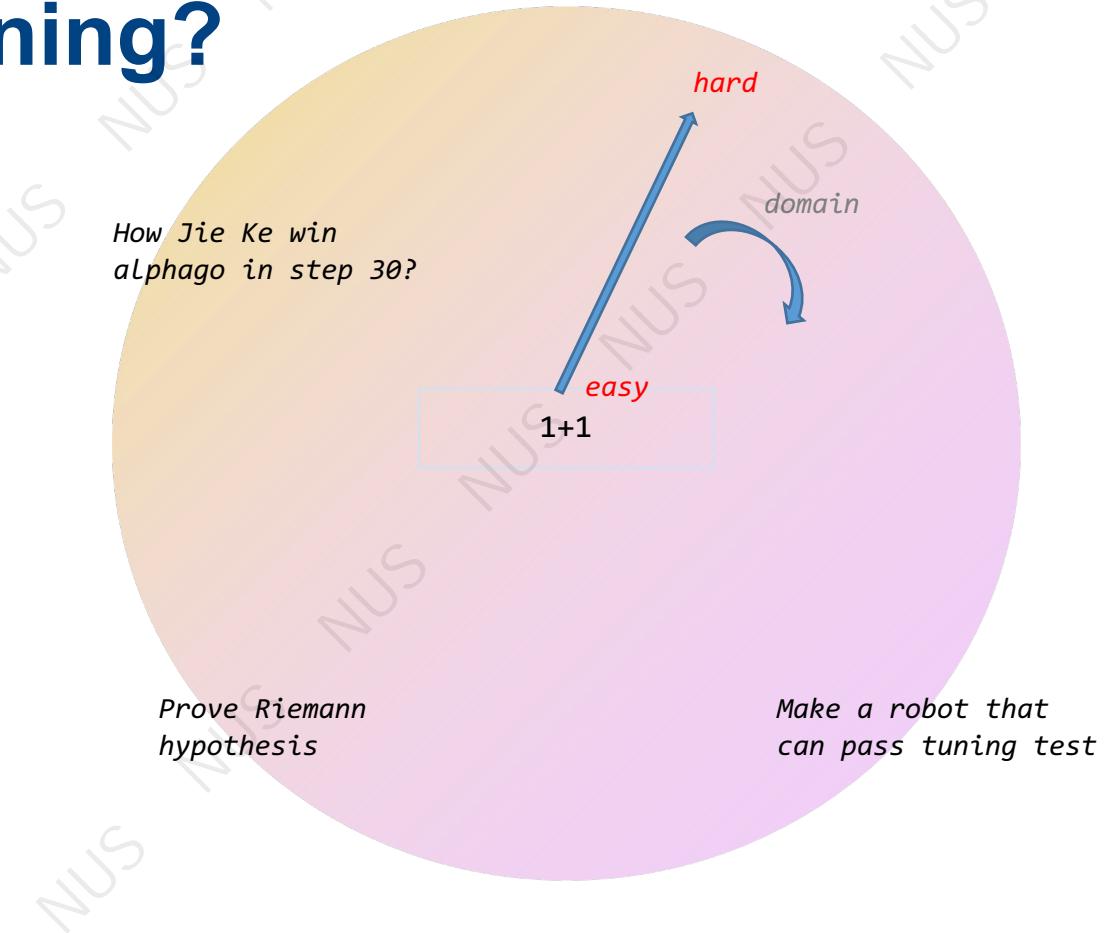
- ▶ The space of problems
 - ▶ Even harder^[1]...



[1] Center for AI Safety, Scale AI. Humanity's Last Exam. 2025.

| What is Reasoning?

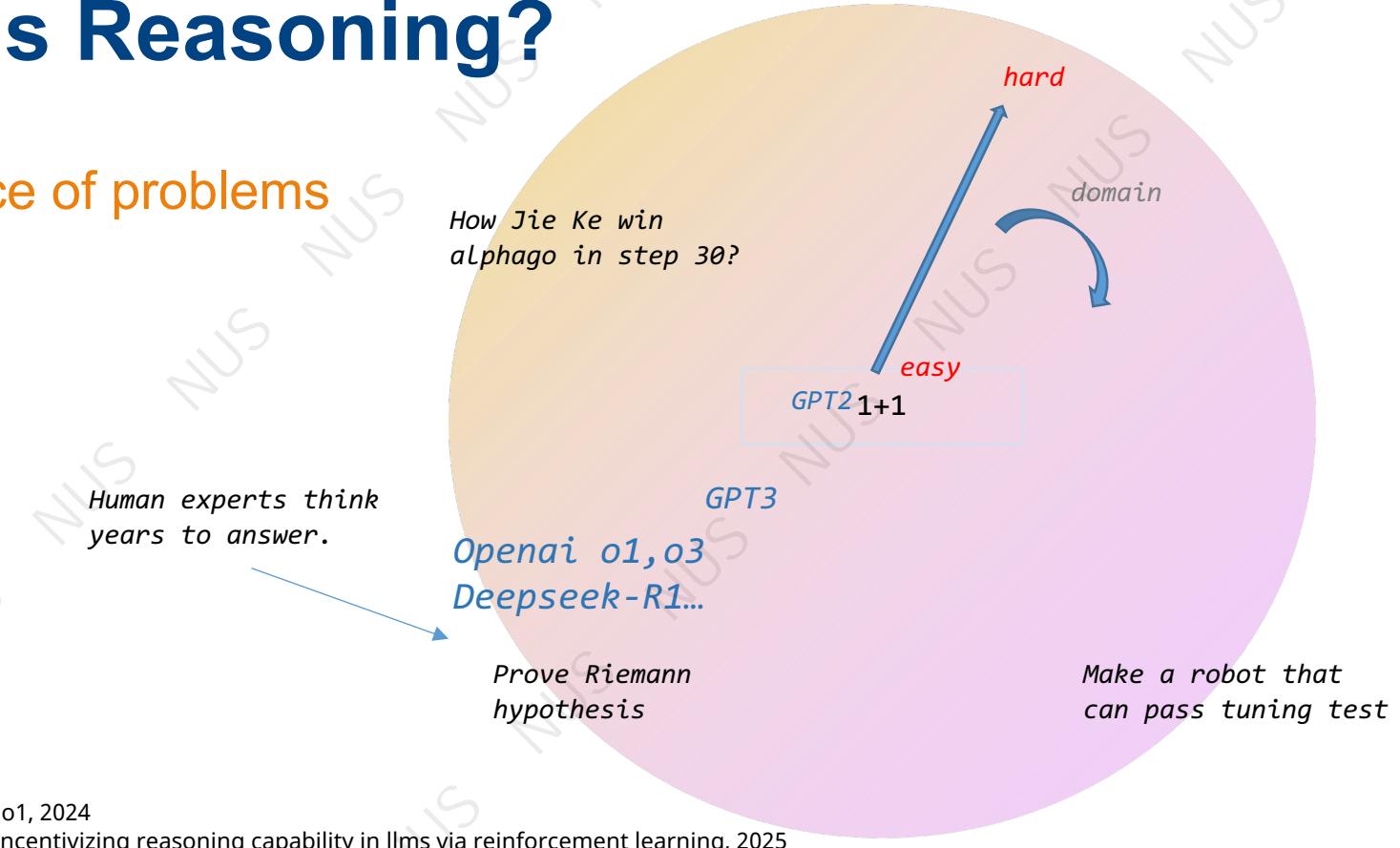
► The space of problems



[1] Ahmed. Learning to Reason with LLMs. Openai Forum. 2024

| What is Reasoning?

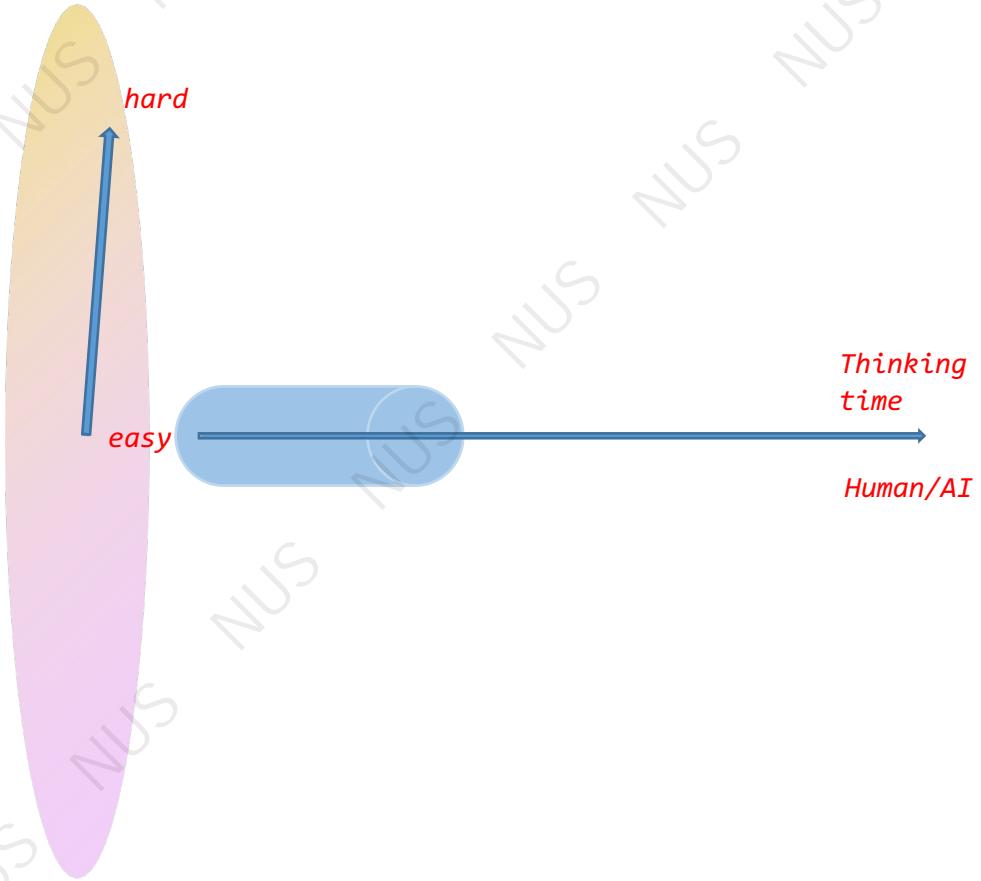
► The space of problems



[1] OpenAI. Introducing openai o1, 2024

[2] DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025

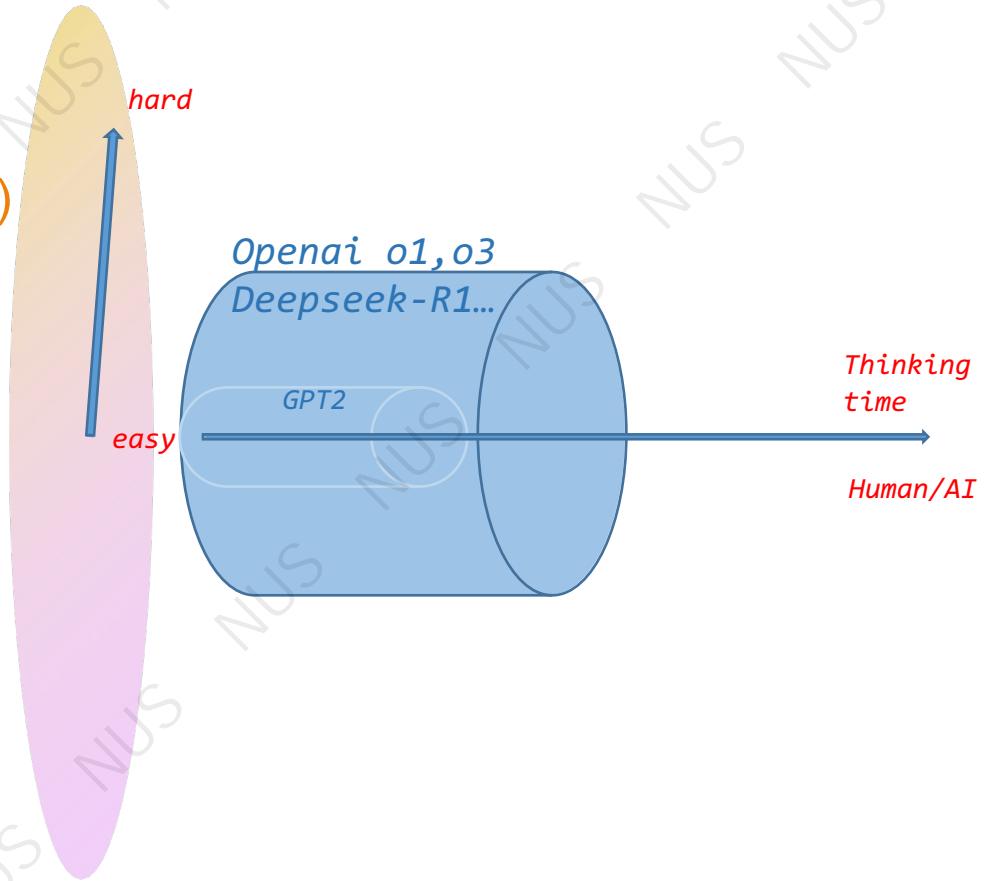
Problems Harder, Thinking Longer



| Long Thinking...

- ▶ Large Reasoning Model (LRM) generates with reasoning

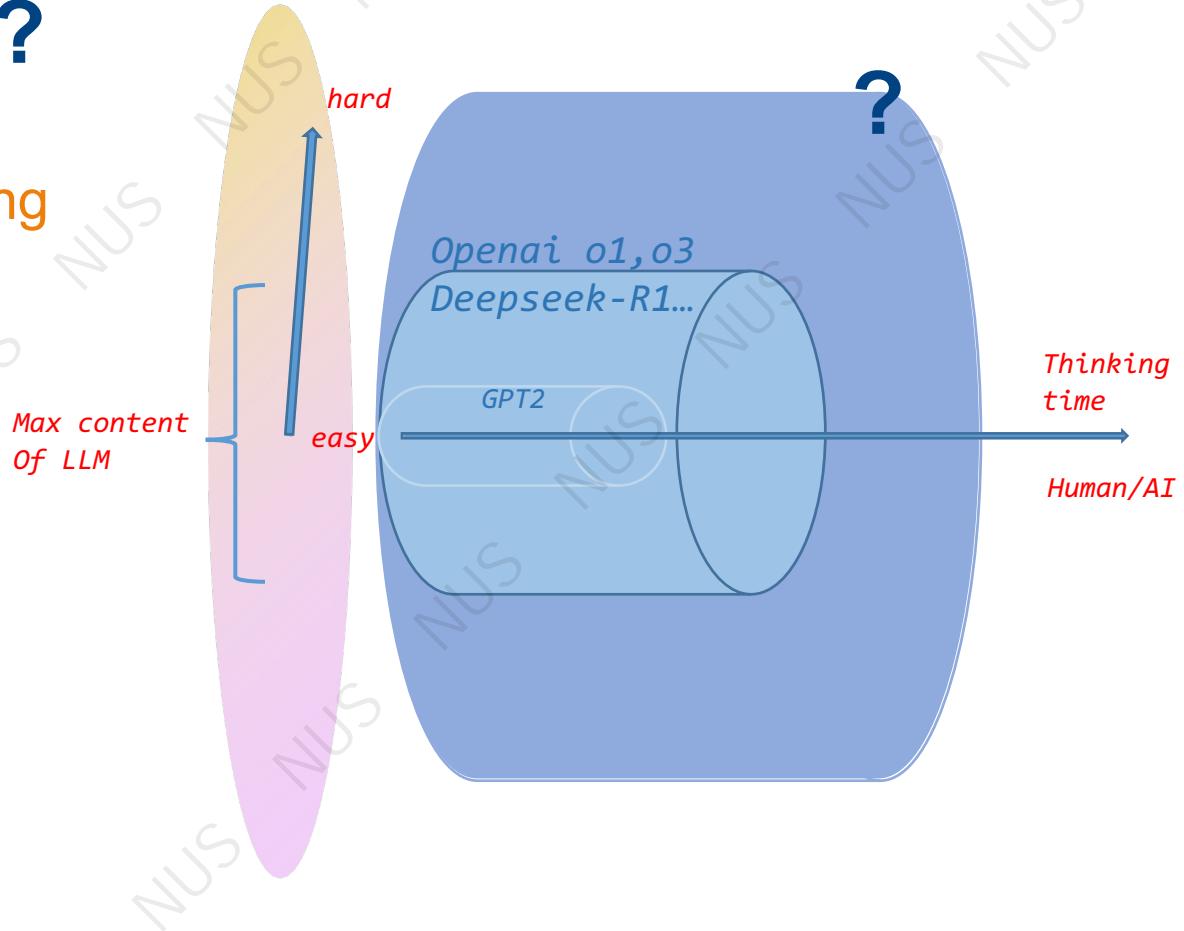
- ▶ Long chain reasoning
 - ▶ Verification
 - ▶ Subgoal setting
 - ▶ Reflection
 - ▶ ...



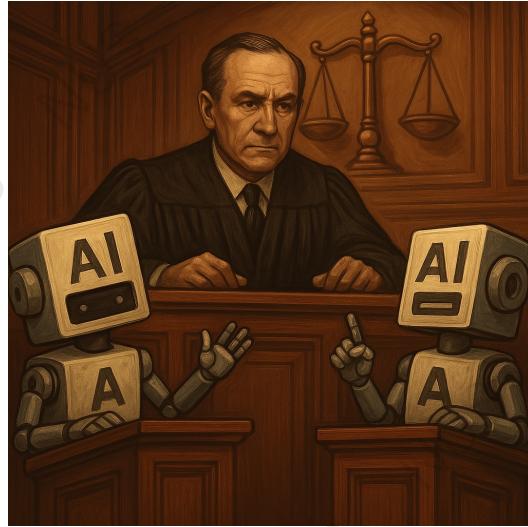
| Looooonger?

► Long chain reasoning

- Verification
- Subgoal setting
- Reflection
- ...

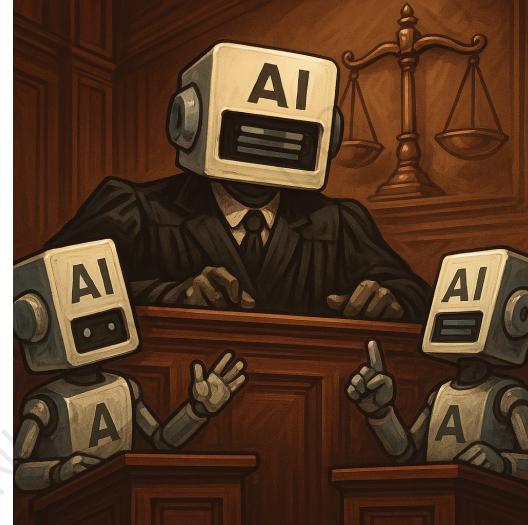
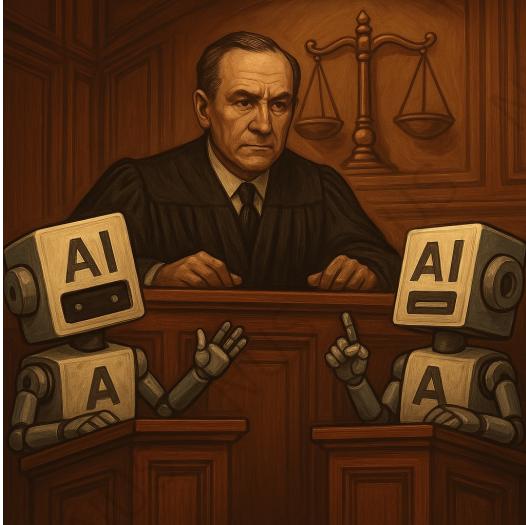


| Previous Judges



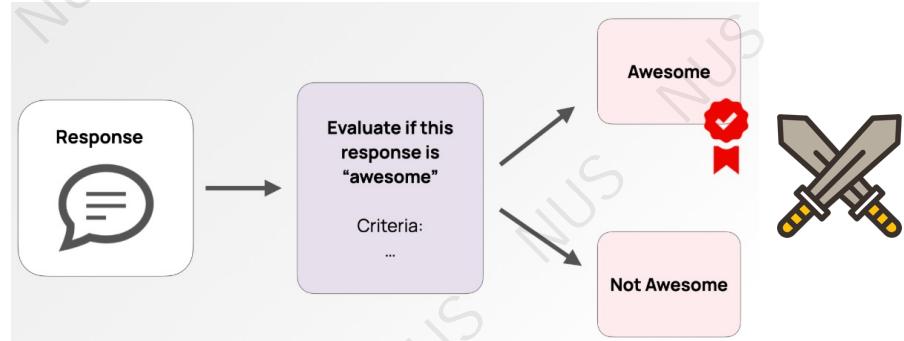
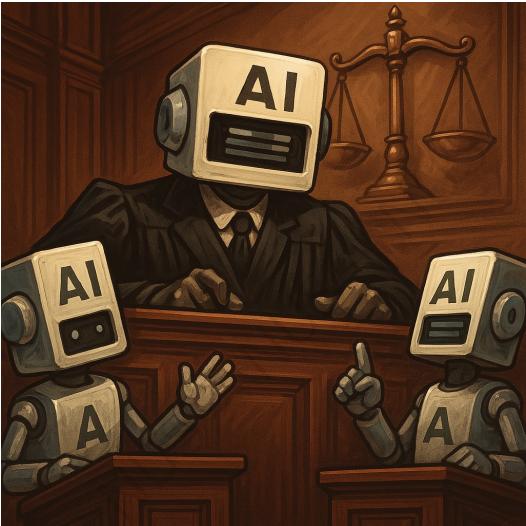
Gen by GPT-4o, Ghibli style

LLM as a Judge?



[1] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. NIPS 2024.

LLM as a Judge?



Chatbot Arena

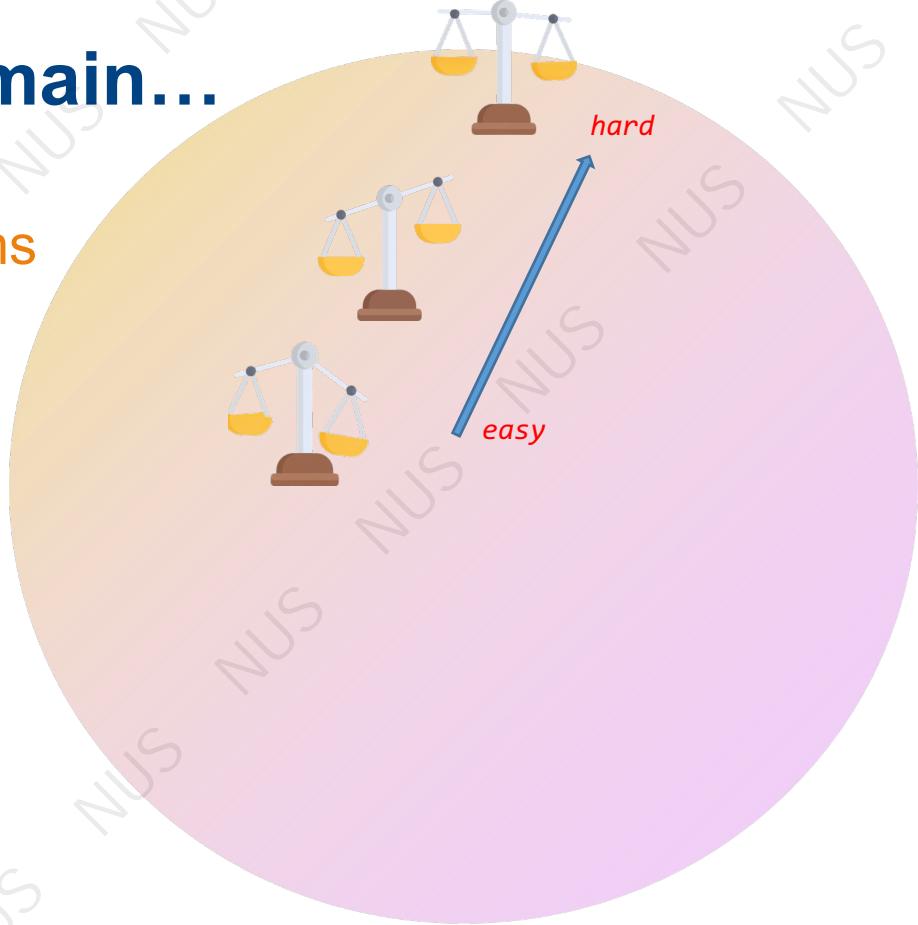
Rank* (UB)	Rank (StyleCtrl)	Model	Arena Score	95% CI	Votes	Organization	License
1	1	Gemini-2.5-Pro-Exp-03-25	1437	+8/-6	7431	Google	Proprietary
2	2	ChatGPT-4o-latest_(2025-03-26)	1406	+7/-8	6612	OpenAI	Proprietary
2	4	Grok-3-Preview-02-24	1402	+5/-5	13919	xAI	Proprietary
2	2	GPT-4.5-Preview	1397	+5/-6	13443	OpenAI	Proprietary
5	8	Gemini-2.0-Flash-Thinking-Exp-01-21	1380	+5/-4	25266	Google	Proprietary
5	4	Gemini-2.0-Pro-Exp-02-05	1380	+4/-5	20136	Google	Proprietary
5	4	DeepSeek-V3-0324	1370	+7/-7	4721	DeepSeek	MIT
7	5	DeepSeek-R1	1359	+5/-5	15098	DeepSeek	MIT
8	13	Gemini-2.0-Flash-001	1354	+4/-4	21065	Google	Proprietary
8	4	o1-2024-12-17	1350	+4/-5	27831	OpenAI	Proprietary
10	13	Gemma-3-27B-1t	1342	+7/-6	9147	Google	Gemma
11	13	Owen2.5-Max	1340	+4/-4	19995	Alibaba	Proprietary



[1] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. NIPS 2024.

| Take Judge as a Domain...

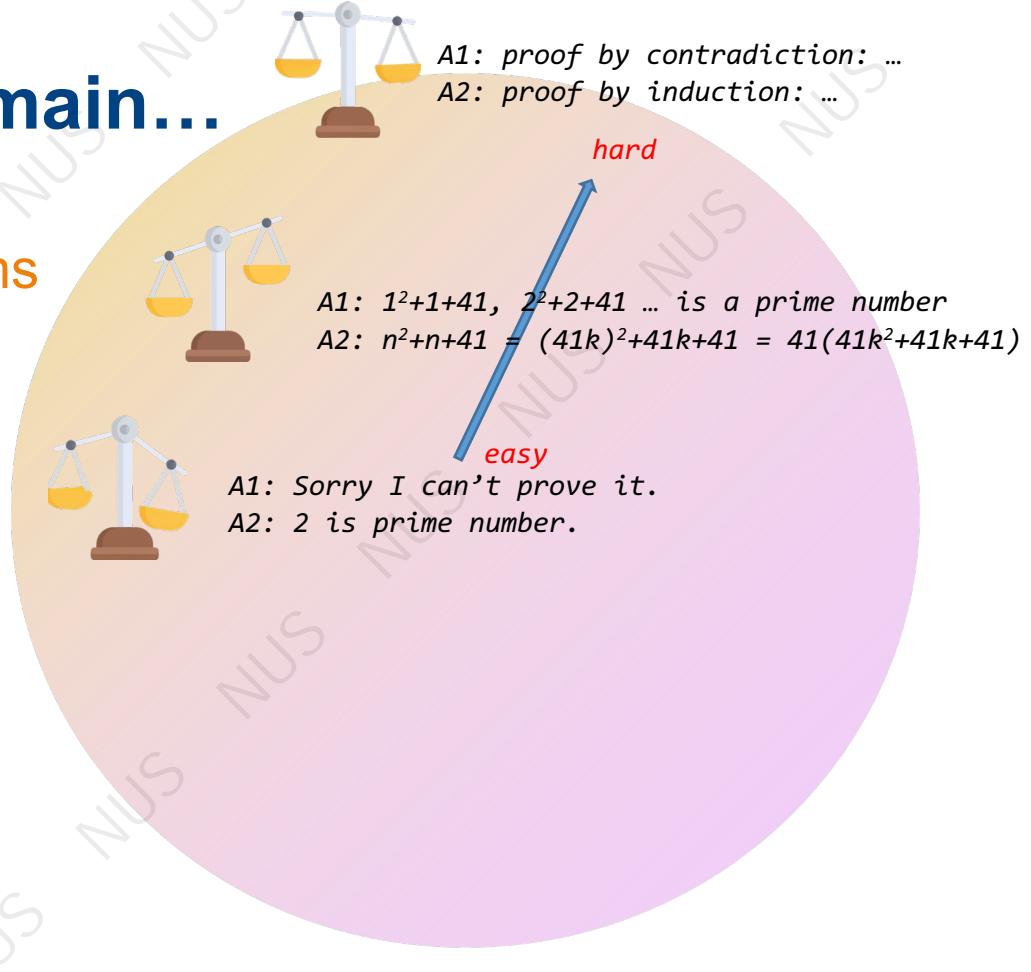
- ▶ The space of judge problems



Take Judge as a Domain...

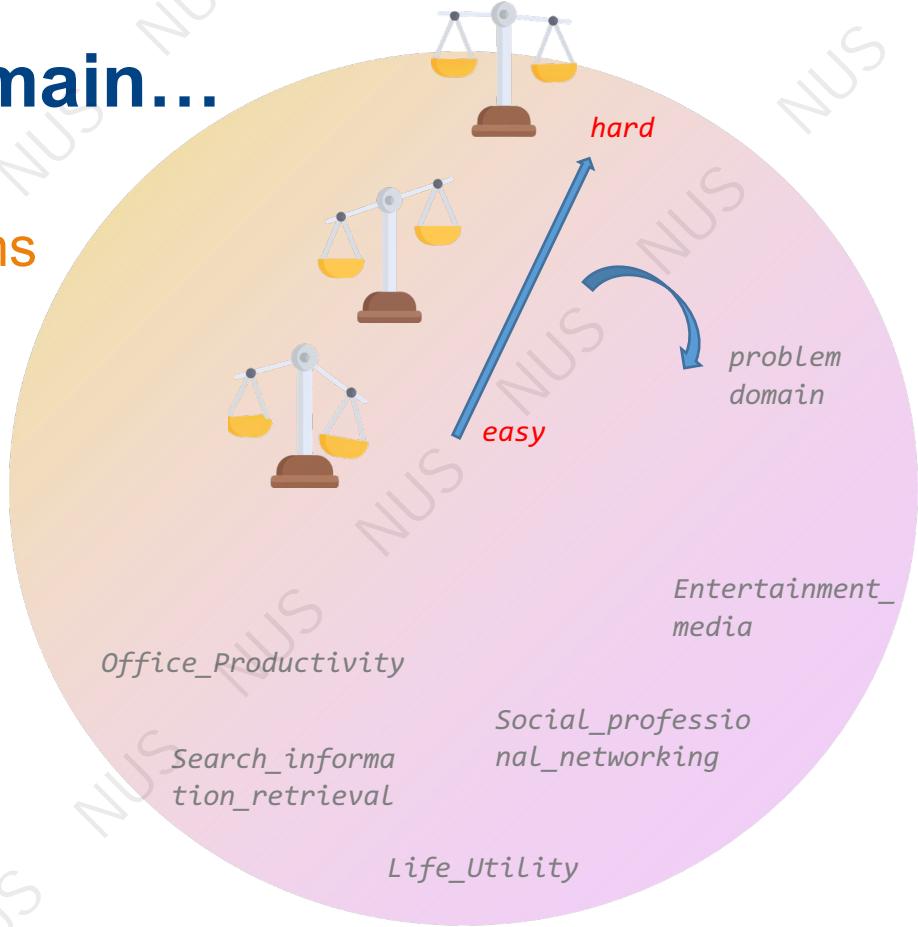
► The space of judge problems

For all positive integers n ,
 n^2+n+41 is a prime number.



| Take Judge as a Domain...

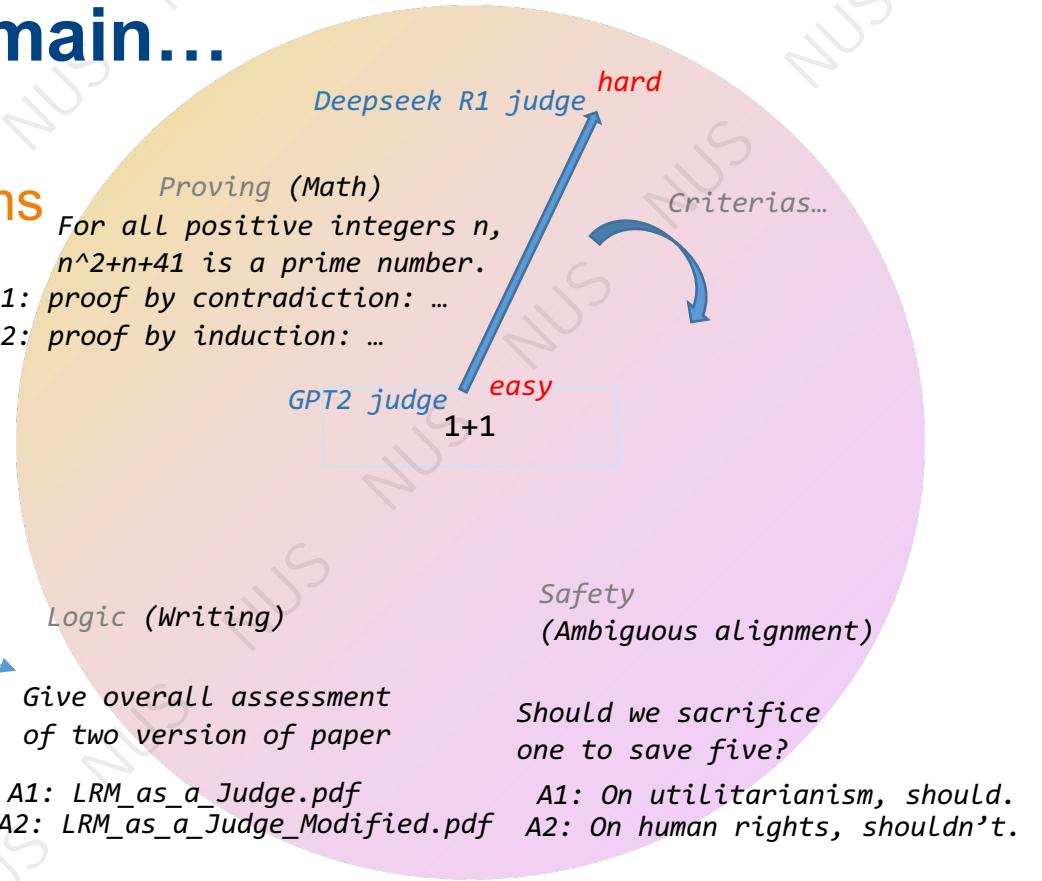
- ▶ The space of judge problems



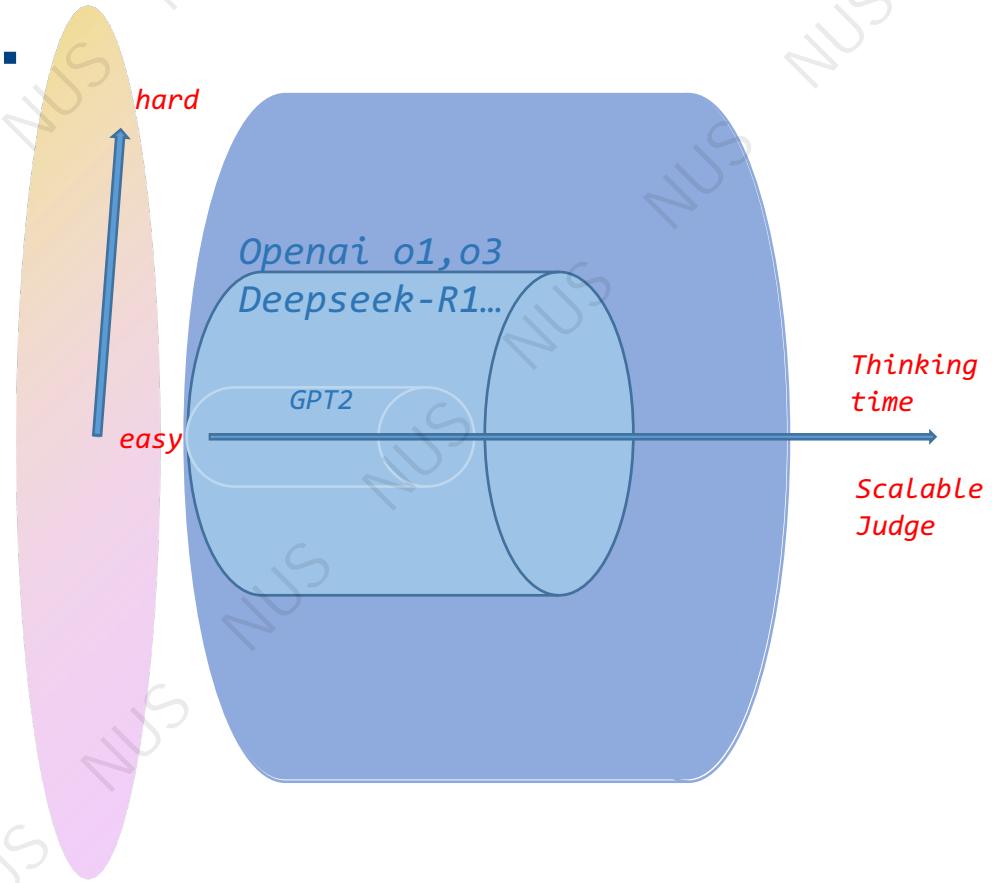
Take Judge as a Domain...

► The space of judge problems

Human experts are hard to evaluate.



Judge Thinking...

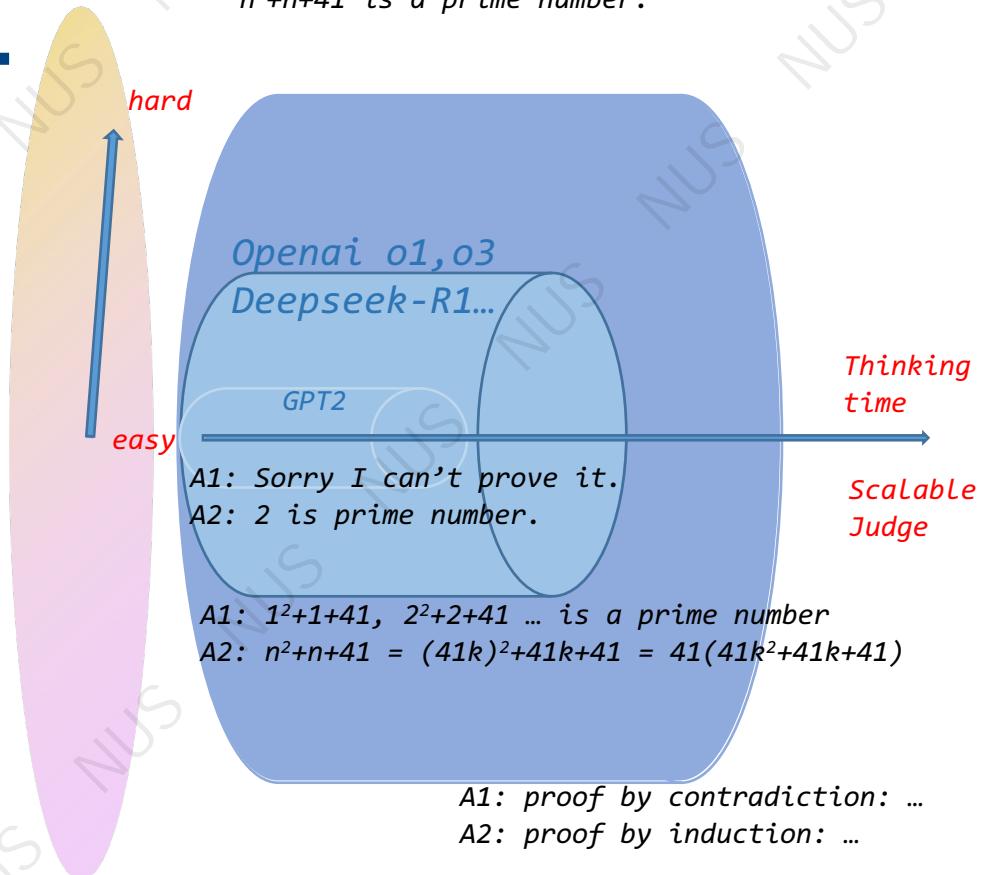


Judge Thinking...

- Criterias on judge

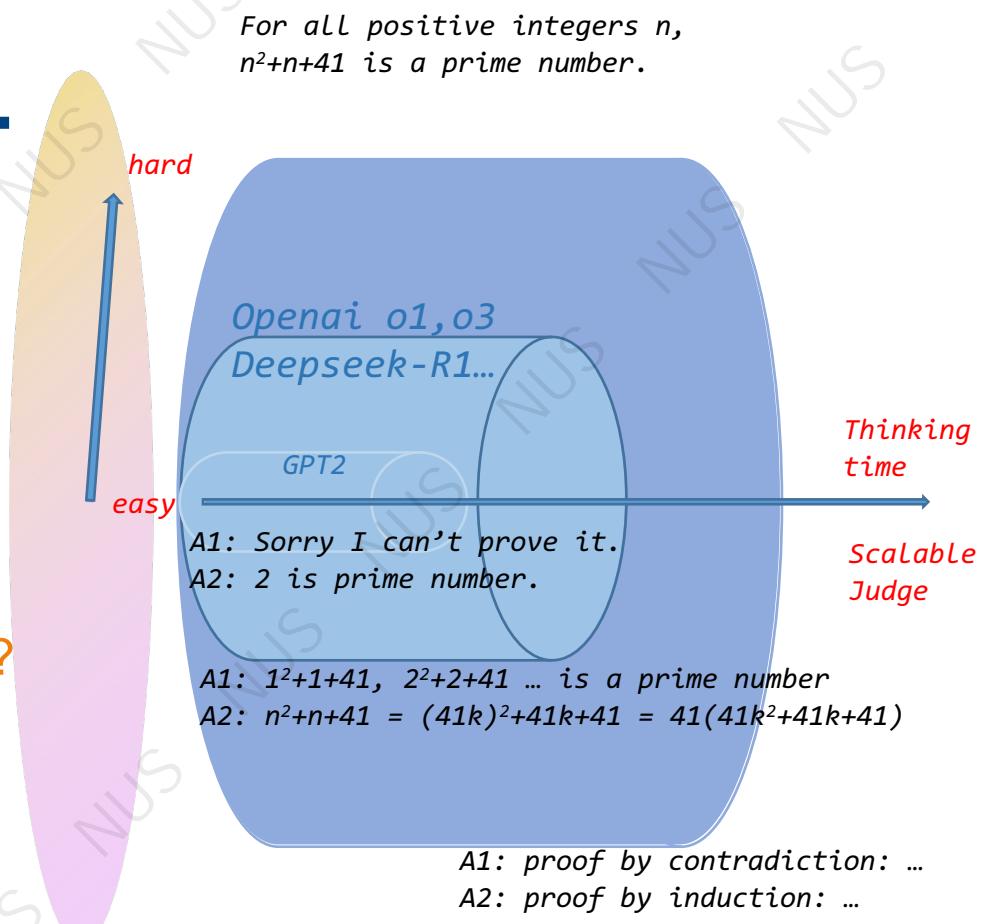
- Logical
- Secure
- Accurate
- Completeness
- ...

For all positive integers n ,
 n^2+n+41 is a prime number.



Judge Thinking...

- Criterias on judge
 - Logical
 - Secure
 - Accurate
 - Completeness
 - ...
- Long chain reasoning needed?
 - Verification
 - Subgoal setting
 - Reflection
 - ...



| LRM as a Judge?

- ▶ Do judges need reasoning?
 - ▶ Domain: Life Utility, Search Information Retrieval,...

| LRM as a Judge?

► Do judges need reasoning?

► Domain: Life Utility, Search Information Retrieval,...

< Papers arxiv:2504.00050

JudgeLRM: Large Reasoning Models as a Judge

Published on Mar 31 · ★ Submitted by  zhiyuanhucs on Apr 2 #2 Paper of the day

Authors:  Nuo Chen,  Zhiyuan Hu,  Qingyun Zou, Jiaying Wu, Qian Wang,  Bryan Hooi, Bingsheng He

Huggingface #2 Paper of the day

[1] Nuo Chen, Zhiyuan Hu, Qingyun Zou, Jiaying Wu, Qian Wang, Bryan Hooi, Bingsheng He.
Judgelrm: Large reasoning models as a judge, 2025. URL <https://arxiv.org/abs/2504.00050>.

| LRM as a Judge?

- ▶ Previous SOTA judges^[1,2] requires Supervised FineTuning (SFT)
 - ▶ Depends on learning massive distilled responses and explanations 😢
- ▶ Do judges need reasoning?
 - ▶ Gain in judge in SFT
 - V.S. Proportion of reasoning-intensive samples

[1] Yidong Wang, Zhuohao Yu, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, Xing Xie, Wei Ye, Shikun Zhang, and Yue Zhang. PandaLM: An Automatic Evaluation Benchmark for LLM Instruction Tuning Optimization. ICLR 2024.

[2] Lianghui Zhu, Xinggang Wang, and Xinlong Wang. Judgelm: Fine-tuned large language models are scalable judges, ICLR 2025 Spotlight.

LRM as a Judge!

► Do judges need reasoning?

► Gain in judge in SFT
V.S. Proportion

► Take away: **inverse** relationship
between improvements in F1^[1] score on judge task and the proportion

->SFT **struggles** more on categories demanding **higher reasoning depth!**

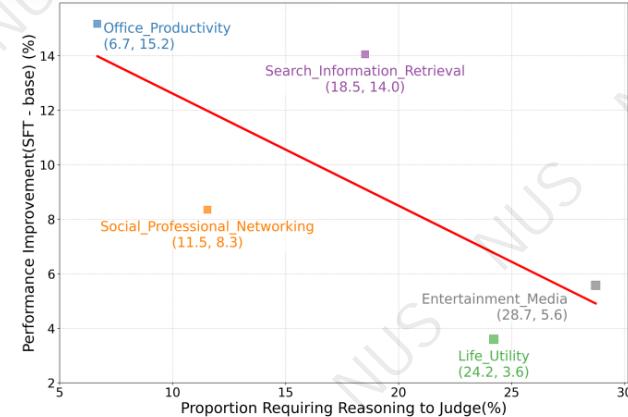


Figure 1: Judgment performance improvement vs. reasoning requirement across domain. Y-axis shows F1 score improvement (SFT – base) based on Qwen2.5-7B-instruct; X-axis shows the proportion of tasks requiring reasoning. Each point represents a domain. A negative linear trend ($y = -0.41x + 16.72$, $R^2 = 0.53$) suggests that domains with more reasoning-heavy tasks benefit less from SFT alone. Sample counts across domains: 105 (Office_Productivity), 108 (Search_Information_Retrieval), 195 (Entertainment_Media), and estimated 108 (Social_Professional_Networking), 190 (Life_Utility).

[1] Yidong Wang, Zhuohao Yu, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, Xing Xie, Wei Ye, Shikun Zhang, and Yue Zhang. PandaLM: An Automatic Evaluation Benchmark for LLM Instruction Tuning Optimization. ICLR 2024..



How about Reinforcement Learning (RL)?

Group Relative Policy Optimization^[1] (GRPO)

- ▶ GRPO grouping samples and normalizing **intra-group** advantages
 - ▶ reduce variance and improve **stability** in strategy updates
 - ▶ reduce **training resources** by replacing the value model in PPO
 - ▶ reduce **query-type biases** to ensure balanced performance across domains

Methods	Data Source	Reward Function
SFT	$q, o \sim P_{sft}(Q, O)$	-
RFT	$q \sim P_{sft}(Q), o \sim \pi_{sft}(O q)$	Rule
DPO	$q \sim P_{sft}(Q), o^+, o^- \sim \pi_{sft}(O q)$	Rule
Online RFT	$q \sim P_{sft}(Q), o \sim \pi_\theta(O q)$	Rule
PPO	$q \sim P_{sft}(Q), o \sim \pi_\theta(O q)$	Model
GRPO	$q \sim P_{sft}(Q), \{o_i\}_{i=1}^G \sim \pi_\theta(O q)$	Model

[1] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024.

GRPO V.S. PPO/SFT on Judges

- ▶ GRPO adaptively adjust the strategy optimization of different groups
 - ▶ avoids biasing the model **towards one certain type**, because the reasoning ability required by different categories **varies** 😊
- ▶ GRPO stabilizes in reasoning-intensive tasks by **structural and content rewards**
 - ▶ encourages the model to learn distinguishing good/bad within problem groups and make relative judgments, **rather than pursuing absolute high rewards** 😊

[1] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024.

| Stimulate LM Reasoning on Judges

- ▶ Apply GRPO to learn judge data^[1] by outcome-driven rewards
- ▶ Learn start from Qwen Instruct
- ▶ No any reasoning details guided!

$$\begin{aligned}\mathcal{J}_{GRPO}(\theta) &= \mathbb{E}[q \sim P(Q), \{o_i\}_{i=1}^G \sim \pi_{\theta_{old}}(O|q)] \\ &\quad \frac{1}{G} \sum_{i=1}^G \frac{1}{|o_i|} \sum_{t=1}^{|o_i|} \left\{ \min \left[\frac{\pi_\theta(o_{i,t}|q, o_{i,<t})}{\pi_{\theta_{old}}(o_{i,t}|q, o_{i,<t})} \hat{A}_{i,t}, \text{clip} \left(\frac{\pi_\theta(o_{i,t}|q, o_{i,<t})}{\pi_{\theta_{old}}(o_{i,t}|q, o_{i,<t})}, 1 - \varepsilon, 1 + \varepsilon \right) \hat{A}_{i,t} \right] - \beta \mathbb{D}_{KL} [\pi_\theta || \pi_{ref}] \right\}\end{aligned}$$

[1] Lianghui Zhu, Xinggang Wang, and Xinlong Wang. Judgelm: Fine-tuned large language models are scalable judges, ICLR 2025 Spotlight.

[2] Tian Xie, Zitian Gao, Qingnan Ren, Haoming Luo, Yuqian Hong, Bryan Dai, Joey Zhou, Kai Qiu, Zhirong Wu, Chong Luo.

Logic-RL: Unleashing LLM Reasoning with Rule-Based Reinforcement Learning 2025.

Reasoning Details on knights and knaves^[2]

Question:

A very special island is inhabited only by knights and knaves. Knights always tell the truth, and knaves always lie. Five residents (William, Chloe, Lily, Jack, and Logan) made the following statements: (1) William said: "William is a knight and Chloe is a knight." (2) Chloe said: "Lily is a knight." (3) Lily said: "William is a knave and Jack is a knight." (4) Jack said: "William is a knave." (5) Logan said: "Lily is a knave if and only if William is a knave." So who is a knight and who is a knave?

Response: <think>

To solve this, let's start analyzing the statements one by one, beginning with William's statement since it's a good starting point due to its self-referential nature.

William's statement: "William is a knight and Chloe is a knight."

- If William is telling the truth (i.e., he is a knight), then both parts of his statement must be true. However, if William is a knight, the statement "William is a knight" is true, but the statement "Chloe is a knight" cannot determine the truth about Chloe at this point since we haven't analyzed Chloe's statement yet. (*Hesitate*) Therefore, William cannot be a knight because it would lead to a contradiction.

...

Let's test both possibilities. (*Explore*)

Testing Chloe as a knight: If Chloe is a knight, then her statement "Lily is a knight" is true, so Lily is a knight. Analyzing Lily's statement: "William is a knave and Jack is a knight."

...

Now, we have determined: (*Summarize*)

- William is a knave.
- Chloe is a knight.

...

Let's verify all statements: (*Verify*)

- "William is a knight and Chloe is a knight" (False, since William is a knave).
- "Lily is a knight" (True).

...

| Stimulate LM Reasoning on Judges

- ▶ Judge data^[1] includes pair answer score in 1-10
 - ▶ It builds the source of rewards.
 - ▶ Score ≤ 5 : unreliable answers with significant mistakes
 - ▶ Score around 7: answers partially correct, meeting usable standards
 - ▶ Score ≥ 9 : detailed and almost flawless answers
 - ▶ Pair-wise evaluations
 - ▶ Helpfulness
 - ▶ Relevance
 - ▶ Accuracy
 - ▶ Detail

[1] Lianghui Zhu, Xinggang Wang, and Xinlong Wang. Judgelm: Fine-tuned large language models are scalable judges, ICLR 2025 Spotlight.

| Stimulate LM Reasoning on Judges

► Apply GRPO^[1] to learn judge data^[2] by **outcome-driven** rewards

► Structural Reward

- Format: <think>...</think>
<answer>score1</answer> <answer>score2</answer>

► Content Reward

- Relation: correct partial order of judge
- Judge-wise absolute: accurate score1 and score2 value
- Judge-wise confidence: larger score differences than gold

$$r_i = \underbrace{\mathcal{R}_{\text{format}}}_{\mathcal{R}_{\text{struct}}} + \underbrace{r_{\text{relation}} + r_{\text{absolute}} + r_{\text{confidence}}}_{\mathcal{R}_{\text{content}}}$$

$$\begin{aligned} J_{\text{GRPO}}(\theta) = & \mathbb{E}_{Q \sim \mathcal{D}} \mathbb{E}_{(s_1, s_2) \sim \pi_\theta} \left[\min \left(\frac{\pi_\theta(s_1, s_2 | Q, A_1, A_2)}{\pi_{\text{old}}(s_1, s_2 | Q, A_1, A_2)} A_i(s_1, s_2 | Q), \right. \right. \\ & \left. \left. \text{clip} \left(\frac{\pi_\theta(s_1, s_2 | Q, A_1, A_2)}{\pi_{\text{old}}(s_1, s_2 | Q, A_1, A_2)}, 1 - \epsilon, 1 + \epsilon \right) A_i(s_1, s_2 | Q) \right) - \beta D_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}}) \right] \end{aligned}$$

[1] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024.

[2] Lianghui Zhu, Xinggang Wang, and Xinlong Wang. Judgelm: Fine-tuned large language models are scalable judges, ICLR 2025 Spotlight.

Can We Rely on RL More Than SFT?

Can We Rely on RL More Than SFT?

Since the learning objective is only from 2 scores...

JudgeLRM: RL > SFT !

Dataset	JudgeLM (GPT-4 as ground truth)				PandaLM (Human as ground truth)			
	Agreement	Precision	Recall	F1	Agreement	Precision	Recall	F1
<i>Existing Baseline. (* from original paper)</i>								
GPT-3.5*	73.83	70.70	52.80	52.85	62.96	61.95	63.59	58.20
GPT-4*	-	-	-	-	66.47	66.20	68.15	61.80
PandaLM-7B	68.61	40.75	38.82	39.41	59.26	57.28	59.23	54.56
Auto-J-13B*	74.86	61.65	57.53	58.14	-	-	-	-
JudgeLM-7B	81.11	69.67	78.39	72.21	65.07	66.89	71.95	61.92
JudgeLM-13B*	84.33	73.69	80.51	76.17	68.97	68.21	74.15	65.12
JudgeLM-33B*	89.03	80.97	84.76	82.64	75.18	69.30	74.93	69.73
PandaLM-70B*	-	-	-	-	66.87	74.02	66.87	69.23
<i>Base Models.</i>								
Qwen2.5-3B-Instruct	72.29	80.19	64.07	71.23	68.50	50.92	56.13	51.57
Qwen2.5-7B-Instruct	76.85	78.71	77.85	78.28	63.96	61.95	67.61	59.81
Qwen2.5-3B-Instruct-Judge-SFT	83.58	75.25	76.12	75.05	70.57	67.09	73.36	66.10
Qwen2.5-7B-Instruct-Judge-SFT	82.00	84.43	81.74	83.06	73.57	67.31	72.23	67.98
Deepseek-R1	-	-	-	-	78.67	77.51	69.97	72.48
<i>Ours.</i>								
JudgeLRM-3B	83.72	86.31	82.98	84.61	77.68	74.26	70.86	72.12
JudgeLRM-7B	83.74	85.84	83.65	84.73	78.28	74.90	75.74	75.05
- w/o. $r_{abs} + r_{conf}$	-	-	-	-	75.78	69.09	73.69	70.36
- w. \mathcal{R}_{length}	-	-	-	-	78.28	75.81	69.19	71.34

Table 1: Performance on JudgeLM and PandaLM. Qwen2.5-Instruct-Judge-SFT means the finetuned Qwen2.5-Instruct on JudgeLM train set. As pairwise comparisons rarely yield ties, we exclude tie cases ($\sim 10\%$ of the test set) for more interpretable evaluation, which JudgeLRM-7B's F1 reaches 83.47, all metrics on PandaLM surpass DeepSeek-R1. (see Table 7 for details). Notably, **JudgeLRM-7B matches or even surpasses Deepseek-R1 on PandaLM**.

[1] Yidong Wang, Zhuohao Yu, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, Xing Xie, Wei Ye, Shikun Zhang, and Yue Zhang. PandaLM: An Automatic Evaluation Benchmark for LLM Instruction Tuning Optimization. ICLR 2024..

[2] Lianghui Zhu, Xinggang Wang, and Xinlong Wang. Judgelm: Fine-tuned large language models are scalable judges, ICLR 2025 Spotlight.

► Both learn from Qwen2.5-7B Instruct

JudgeLM: RL > SFT !

Dataset	JudgeLM (GPT-4 as ground truth)				PandaLM (Human as ground truth)			
	Agreement	Precision	Recall	F1	Agreement	Precision	Recall	F1
<i>Existing Baseline. (* from original paper)</i>								
GPT-3.5*	73.83	70.70	52.80	52.85	62.96	61.95	63.59	58.20
GPT-4*	-	-	-	-	66.47	66.20	68.15	61.80
PandaLM-7B	68.61	40.75	38.82	39.41	59.26	57.28	59.23	54.56
AutoJ-13B*	74.86	61.65	57.53	58.14	-	-	-	-
JudgeLM-7B	81.11	69.67	78.39	72.21	65.07	66.89	71.95	61.92
JudgeLM-13B*	84.33	73.69	80.51	76.17	68.97	68.21	74.15	65.12
JudgeLM-33B*	89.03	80.97	84.76	82.64	75.18	69.30	74.93	69.73
PandaLM-70B*	-	-	-	-	66.87	74.02	66.87	69.23
<i>Base Models.</i>								
Qwen2.5-3B-Instruct	72.29	80.19	64.07	71.23	68.50	50.92	56.13	51.57
Qwen2.5-7B-Instruct	76.85	78.71	77.85	78.28	63.96	61.95	67.61	59.81
Qwen2.5-3B-Instruct-Judge-SFT	83.58	75.25	76.12	75.05	70.57	67.09	73.36	66.10
Qwen2.5-7B-Instruct-Judge-SFT	82.00	84.43	81.74	83.06	73.57	67.31	72.23	67.98
Deepseek-R1	-	-	-	-	78.67	77.51	69.97	72.48
<i>Ours.</i>								
JudgeLM-3B	83.72	86.31	82.98	84.61	77.68	74.26	70.86	72.12
JudgeLM-7B	83.74	85.84	83.65	84.73	78.28	74.90	75.74	75.05
- w/o. $r_{abs} + r_{conf}$	-	-	-	-	75.78	69.09	73.69	70.36
- w. \mathcal{R}_{length}	-	-	-	-	78.28	75.81	69.19	71.34

Table 1: Performance on JudgeLM and PandaLM. Qwen2.5-Instruct-Judge-SFT means the finetuned Qwen2.5-Instruct on JudgeLM train set. As pairwise comparisons rarely yield ties, we exclude tie cases ($\sim 10\%$ of the test set) for more interpretable evaluation, which JudgeLM-7B's F1 reaches 83.47, all metrics on PandaLM surpass DeepSeek-R1. (see Table 7 for details). Notably, JudgeLM-7B matches or even surpasses Deepseek-R1 on PandaLM.

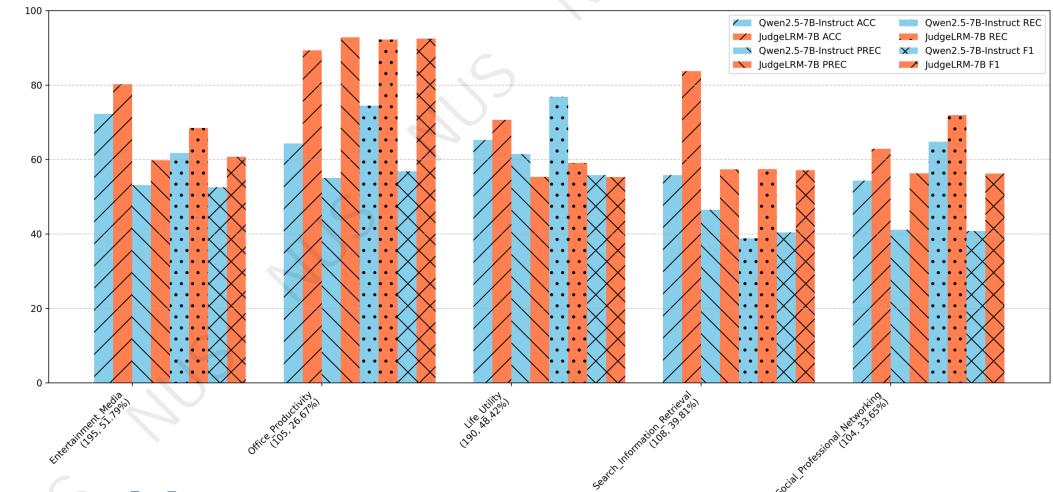
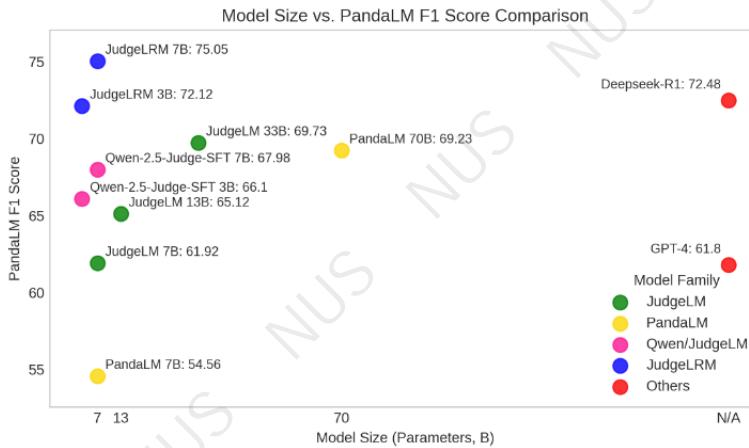
[1] Yidong Wang, Zhuohao Yu, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, Xing Xie, Wei Ye, Shikun Zhang, and Yue Zhang. PandaLM: An Automatic Evaluation Benchmark for LLM Instruction Tuning Optimization. ICLR 2024..

[2] Lianghui Zhu, Xinggang Wang, and Xinlong Wang. Judgelm: Fine-tuned large language models are scalable judges, ICLR 2025 Spotlight.

- ▶ Both learn from Qwen2.5-7B Instruct
- ▶ RL methods significantly surpass SFT
- ▶ JudgeLM-3B surpasses GPT-4!
- ▶ JudgeLM-7B outperforms DeepSeek-R1!

JudgeLRM: RL > SFT !

- ▶ JudgeLRM success under deployable size^[1]
- ▶ RL methods significantly surpass SFT^[2]



[1] Score of JudgeLRMs and baselines

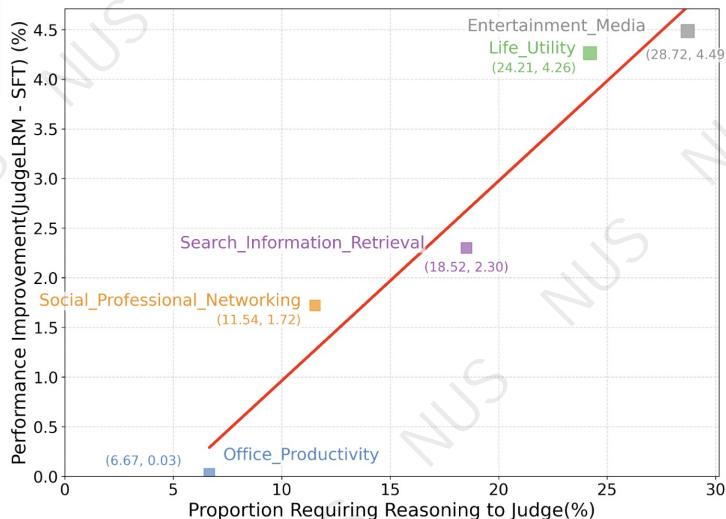
[2] Reasoning-Intensive domain-specific Improvement compared RL and SFT

Analysis: the Source of JudgeLRM F1 Gain

- ▶ The **correlation** shown between performance improvement and the proportion of reasoning-needed instances?

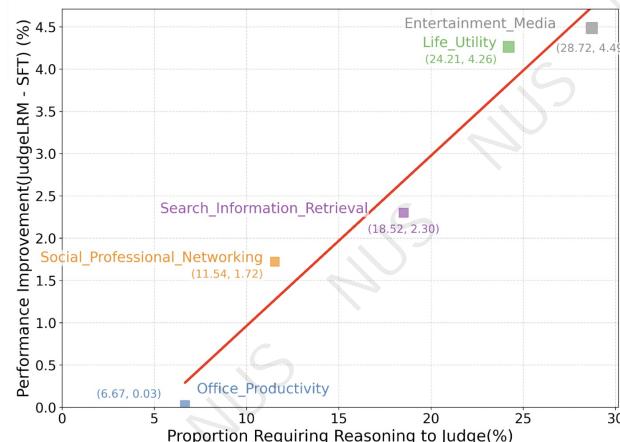
Analysis: the Source of JudgeLRM F1 Gain

- Clear positive correlation shown between performance improvement and the proportion of reasoning-needed instances



Analysis: the Source of JudgeLRM F1 Gain

- ▶ Clear **positive correlation** shown between performance improvement and the proportion of reasoning-needed instances
 - ▶ -> Evidence for the reasoning-intensive nature of judge tasks
 - ▶ -> Evidence for the effectiveness of RL in enhancing judgement capabilities through reasoning.



Analysis: CoT Case Study

- ▶ Long chain reasoning impact judge quality
- ▶ Deep reasoning behavior emerges through RL learning
 - ▶ Verification
 - ▶ Subgoal setting
 - ▶ Reflection
 - ▶ ...

Question:

(Instruction:) Include important study notes and key points that someone should know about the given subject. (Input:) history of the USA

[Assistant 1's Answer] The history of the United States is one of the most influential and influential countries in the world. It has had a profound impact on the world and continues to shape the world today. The United States has had a long and storied history, from its founding in 1775 to the present day. It has had a tumultuous and tumultuous history, from the Revolutionary War to the Civil War and beyond. The United States has had a tumultuous and tumultuous history, from the Revolutionary War to the Civil War and beyond. The United States has had a long and storied history, from its founding in 1775 to the present day. It has had a profound impact on the world and continues to shape the world today.

[Assistant 2's Answer] 1. The United States of America was founded in 1776. 2. The Declaration of Independence was signed in 1776. 3. The Constitution of the United States of America was signed in 1787. 4. The Civil War began in 1861. 5. The Emancipation Proclamation was issued in 1863. 6. The 13th Amendment was ratified in 1865. 7. The 14th Amendment was ratified in 1868. 8. The 15th Amendment was ratified in 1870. 9. The 16th Amendment was ratified in 1913. 10. The 17th Amendment was ratified in 1913. 11. The 18th Amendment was ratified in 1919. 12. The 19th Amendment was ratified in 1920. 13. The 20th Amendment was ratified in 1933. 14. The 21st Amendment was ratified in 1933.

Response: <think>

Assistant 1 provided a very broad, somewhat repetitive statement on the history of the United States which did not include specific key events, dates, or figures that would be considered important in the narrative of U.S. history. This kind of vague and repetitive phrasing might make it difficult for someone seeking specific historical information to gain an accurate or clear understanding. While it correctly noted the overall influence and significant historical periods like the Revolutionary War and the Civil War, this approach lacks specificity which can be misleading or uninformative.

In contrast, Assistant 2's answer is quite accurate and relevant in terms of providing a list of critical events and amendments in U.S. history. Each point listed is accurate, such as the dates when key documents were signed, the years of major constitutional changes, and significant legislative acts like the Emancipation Proclamation and various constitutional amendments related to civil rights and representation. (*Verification*) By including these specific details, the accuracy and relevance of the information provided by Assistant 2 is significantly higher, making it easier for someone to trace specific aspects of American history through these milestones. (*Subgoal Setting*)

Accuracy is also a concern in Assistant 1's answer since the statement "from its founding in 1775" is inaccurate; the actual founding date recognized in the United States is July 4, 1776, with the signing of the Declaration of Independence, rather than 1775. (*Verification, Double Checking*) Additionally, the repetition might confuse readers, suggesting a lack of clarity or precision in how the history is presented, which could mislead people who are looking for a clear timeline or significant events. (*Reflection, Error Identification*)

Both answers need to be more precise, but given the task, the importance lies in identifying key events, dates, and significant changes that shaped the nation. (*Subgoal Setting*) Thus, the accuracy and relevance in listing specific, factual events and their exact timing in history would be a more appropriate way to address the question asked. (*Decision Justification*)

Given these considerations, while both answers are somewhat relevant to the history of the United States, the accuracy and level of detail in the information provided by Assistant 2 would likely be seen as more accurate and relevant because they pinpoint specific, significant events and changes in the country's development.</think>

<answer>2</answer><answer>9</answer>

Analysis: Incentivizing Lengthy Response

- ▶ How about simply incentivizing slightly longer answers?

Analysis: Incentivizing Lengthy Response

- ▶ Simply incentivizing slightly longer answers degraded results of about 3
 - ▶ -> Encouraging verbosity does not benefit the judge task.

$$\mathcal{R}_{\text{length}} = \begin{cases} 0.2, & \text{if the reasoning chain exceeds 120 tokens} \\ -1.0, & \text{if the maximum token limit is reached} \\ 0, & \text{otherwise} \end{cases}$$

Dataset	<i>PandaLM (Human as ground truth)</i>			
Criteria	Agreement	Precision	Recall	F1
JudgeLRM-7B	78.28	74.90	75.74	75.05
- w. $\mathcal{R}_{\text{length}}$	78.28	75.81	69.19	71.34

JudgeLRM Demo

► Compare any Hugging Face LMs by asking any questions!

Demo



Paper



Github



[1] <https://huggingface.co/spaces/nuojohnchen/JudgeLRMDemo>

Spaces • nuojohnchen | JudgeLRMDemo | like 1 | Running on ZERO | × 1 | App Files Community

JudgeLRM

This Space demonstrates the JudgeLRM model, designed to evaluate the quality of two AI assistant responses. JudgeLRM is a family of judgment-oriented LLMs trained using reinforcement learning (RL) with judge-wise, outcome-driven rewards. JudgeLRM models consistently outperform both SFT-tuned and state-of-the-art reasoning models. Notably, JudgeLRM-3B surpasses GPT-4, and JudgeLRM-7B outperforms DeepSeek-R1 by 2.79% in F1 score, particularly excelling in judge tasks requiring deep reasoning.

Enter an instruction and two responses, and the model will think, reason and score them on a scale of 1-10 (higher is better).

You can also select Hugging Face models to automatically generate responses for evaluation.

Auto-Generate Responses Manual Evaluation

Select Judge Model: JudgeLRM-7B

Instruction/Question: Will a computer science PhD graduate be unemployed?

Select Model 1: Qwen/Qwen2.5-7B-Instruct Or enter custom model path: Qwen/Qwen2.5-7B-Instruct

Select Model 2: 01-ai/Yi-6B-Chat Or enter custom model path: 01-ai/Yi-6B-Chat

Parameters

Generate Responses and Evaluate

Response from Model 1: No, a computer science PhD graduate is not guaranteed employment, but they typically have strong job prospects due to their specialized education and skills. Many PhD graduates find employment in academia, research institutions, or industry as software developers, data scientists, researchers, or technical leaders. The specific job market conditions, skills, and experiences of the individual can impact their employment prospects. It's also worth noting that some PhD graduates may choose to continue working on their research or pursue other career paths. Overall, while there is no guarantee of employment, a computer science PhD can provide excellent opportunities for career advancement and job security. </end>

Response from Model 2: It is difficult to predict with certainty whether a computer science PhD graduate will be unemployed, as the job market for PhDs in computer science can vary greatly depending on several factors, including the graduate's area of specialization, experience, location, and the overall economic conditions.

Some computer science PhD graduates go on to become professors in universities, while others find employment in research positions in industry, government, or academia. Additionally, many PhDs in computer science work in high-tech companies as software engineers, data scientists, or researchers.

However, it's important to note that the job market for PhDs in any field, including computer science, can be competitive, and there may be periods of time when certain graduates are unable to find employment immediately after completing their degrees.

To increase the likelihood of finding employment, PhD graduates in computer science can pursue additional skills such as machine learning, data analysis, and programming languages that are in demand in the industry. They can also network, attend job fairs, and consider alternative career paths such as postdoctoral research or non-academic roles.

In summary, while a PhD in computer science can open many doors, the job market for PhDs is not guaranteed, and graduates must actively seek employment and be flexible in their career choices.

Evaluation Results

<think>

Both answers provided by Assistant 1 and Assistant 2 are quite accurate and relevant

Examples

Instruction/Question	Select Model 1	Select Model 2
Write a short poem about artificial intelligence	Qwen/Qwen2.5-7B-Instruct	01-ai/Yi-6B-Chat
我听说有些人有高血压却没有任何症状。这是真的吗？	Freedominelligence/Apollo-7B	openchat/openchat-3.5-0106

Conclusion

- ▶ Take away: judge is reasoning-intensive, SFT struggles
- ▶ Raise LRM as a judge, apply GRPO to stimulate judge reasoning
- ▶ JudgeLRM-7B surpass Deepseek R1 on judge
- ▶ Take away: successful judgment involves complex reasoning behaviors

Reference

- [1] Nuo Chen, Zhiyuan Hu, Qingyun Zou Jiaying Wu Qian Wang Bryan Hooi Bingsheng He. JudgeLrm: Large reasoning models as a judge, 2025. URL <https://arxiv.org/abs/2504.00050>.
- [2] OpenAI. Introducing openai o1, 2024
- [3] DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025
- [4] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging Ilm-as-a-judge with mt-bench and chatbot arena. NIPS 2024.
- [5] Yidong Wang, Zhuohao Yu, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, Xing Xie, Wei Ye, Shikun Zhang, and Yue Zhang. PandaLM: An Automatic Evaluation Benchmark for LLM Instruction Tuning Optimization. ICLR 2024..
- [6] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024.
- [7] Lianghui Zhu, Xinggang Wang, and Xinlong Wang. Judgelm: Fine-tuned large language models are scalable judges, ICLR 2025 Spotlight.
- [8] Center for AI Safety, Scale AI. Humanity's Last Exam. 2025.
- [9] Yue Liu, Jiaying Wu, Yufei He, Hongcheng Gao, Hongyu Chen, Baolong Bi, Jiaheng Zhang, Zhiqi Huang, Bryan Hooi. Efficient Inference for Large Reasoning Models: A Survey. 2025.
- [10] Tian Xie, Zitian Gao, Qingnan Ren, Haoming Luo, Yuqian Hong, Bryan Dai, Joey Zhou, Kai Qiu, Zhirong Wu, Chong Luo. Logic-RL: Unleashing LLM Reasoning with Rule-Based Reinforcement Learning 2025.

THANK YOU