ThinkBench: Dynamic Out-of-Distribution Evaluation for Robust LLM Reasoning

Shulin Huang, Linyi Yang*, Yan Song, Shuang Chen, Leyang Cui, Ziyu Wan,

Qingcheng Zeng, Ying Wen, Kun Shao, Weinan Zhang, Jun Wang, Yue Zhang*

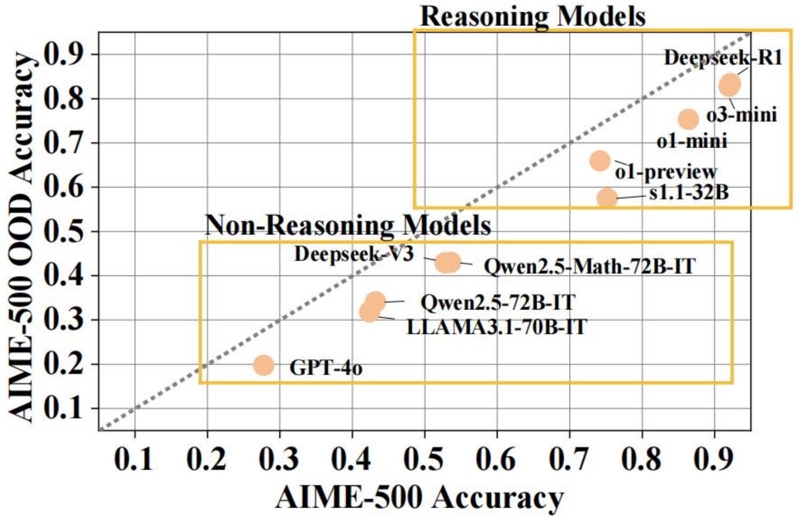
huangshulin@westlake.edu.cn

Zhejiang University, Westlake University, University College London,

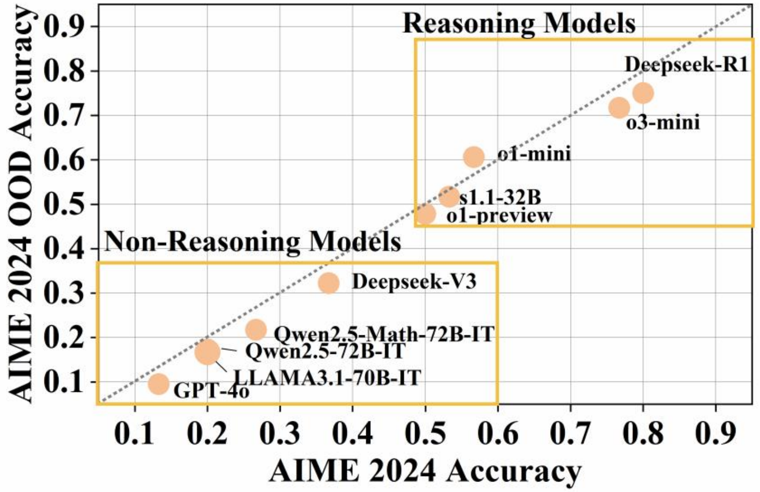
Shanghai Jiao Tong University, Northwestern University, Huawei Noah's Ark Lab

Motivation

- Traditional LLM evaluations suffer from **data contamination.**
- O1 dropped on AIME questions (**pre-2024 data v.s. 2024 data**).
- **Our goal:** To create a benchmark that reduces the impact of data contamination.



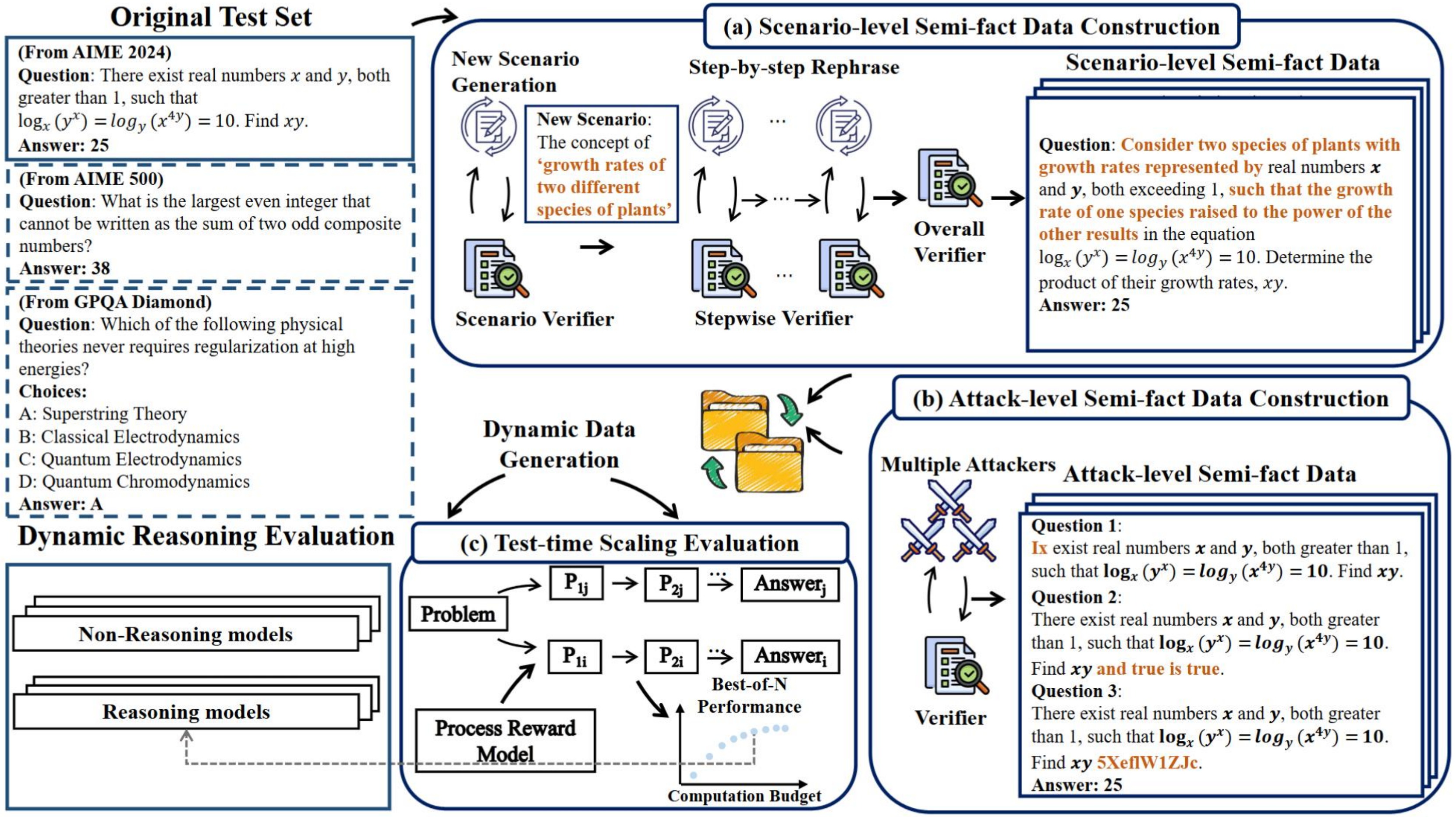
(a) OOD performance vs. ID performance for several reasoning models on AIME-500.



(b) OOD performance vs. ID performance for several reasoning models on AIME 2024.

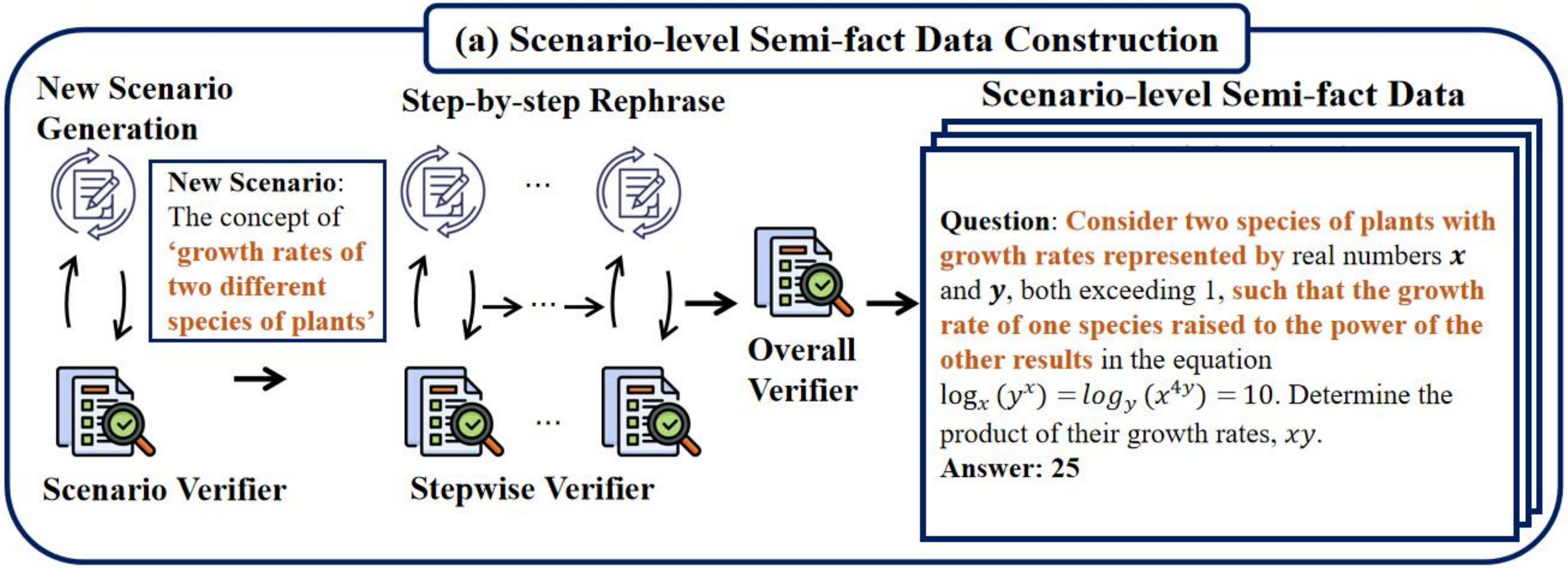
ThinkBench

- A dynamic framework to generate Out-of-Distribution (OOD) evaluation datasets.
- **Scenario-level Semi-fact Data:** Alters the scenario of a problem while preserving its core logic.
- **Attack-level Semi-fact Data:** Introduces minor, realistic perturbations to the problem.

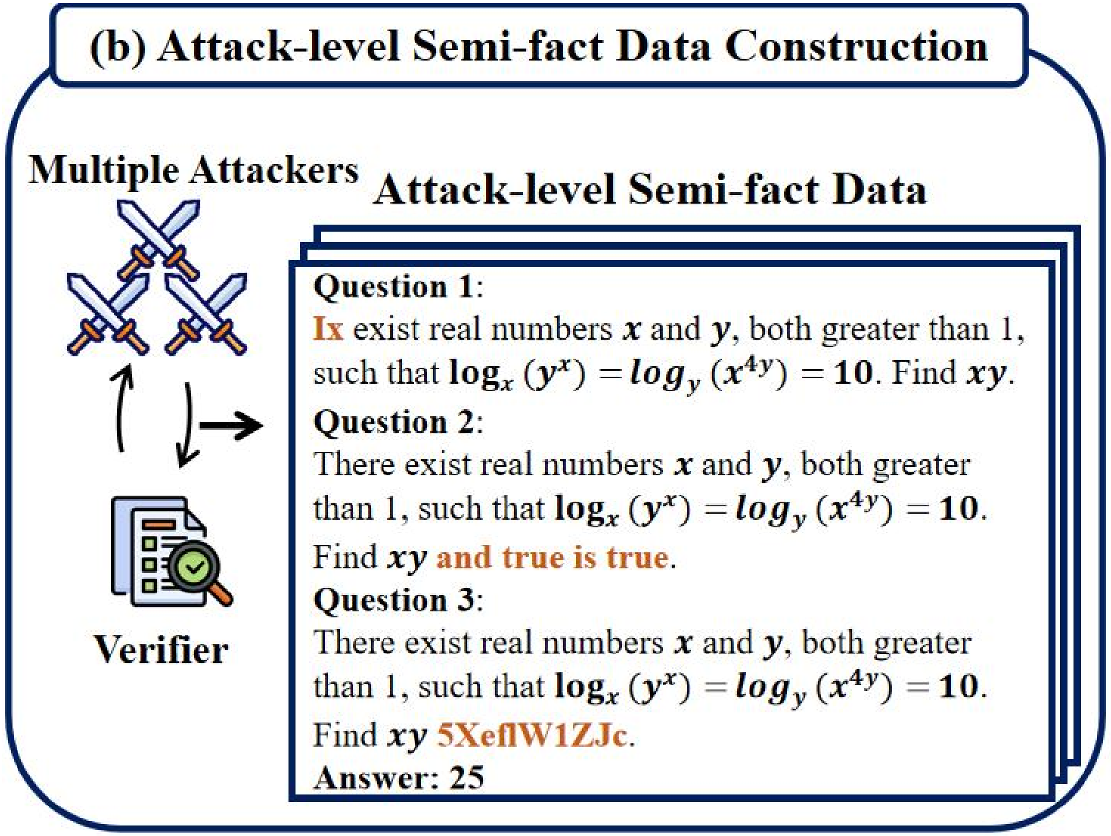


Scenario-level Semi-fact Data Construction

- **New Scenario Generation**: Create a novel background for the original problem.
- **Step-by-step Rephrase:** Integrate the original problem into the new scenario sentence by sentence.
- **Multi-stage Verification**: Use Scenario, Stepwise, and Overall Verifiers to ensure logical consistency and a preserved answer.



Attack-level Semi-fact Data Construction



- To simulate input errors and test a model's noise resilience.
  - **TextBugger (Character-level)**
  - **CheckList (Sentence-level)**
  - **Stress Test (Sentence-level)**

Experiment

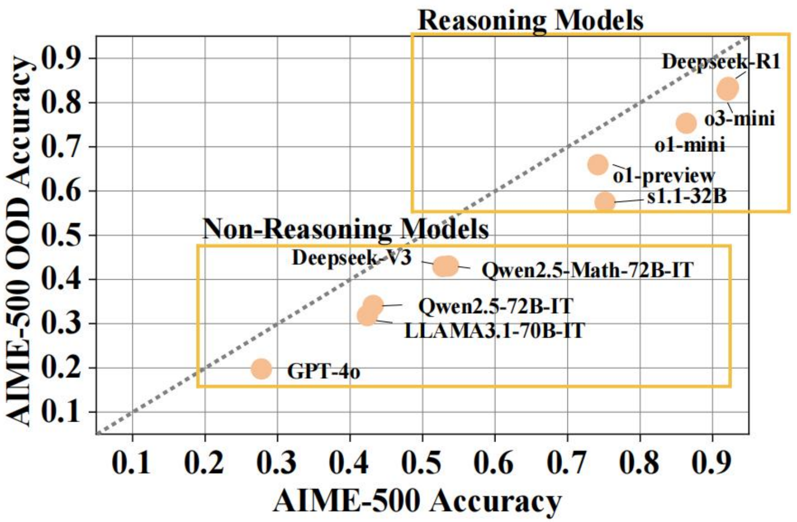
- **Original Datasets:**
  - **AIME-500:** 500 questions from AIME 1983-2023.
  - **AIME 2024:** 30 questions from AIME 2024.
  - **GPQA Diamond:** 198 graduate-level science questions.

- **Generated OOD Data:** 1 scenario-level and 3 attack-level samples per original instance.

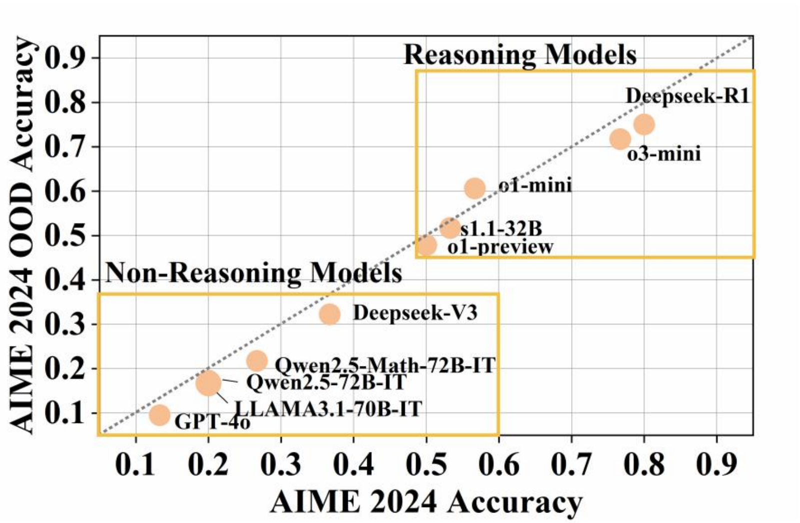
- **Models Evaluated:** 16 LLMs including o1-preview, o3-mini, Deepseek-R1, GPT-4o, LLAMA3.1, etc.

Key Finding 1: Evidence of Data Contamination

- The ID vs. OOD performance gap on AIME-500 (older data) is much larger than on AIME 2024 (newer data).
- **Average performance drop on AIME-500**: 24.9%
- **Average performance drop on AIME 2024**: 11.8%
- This strongly indicates **data leakage** in the pre-2024 AIME datasets, as models were likely exposed to this data during training.



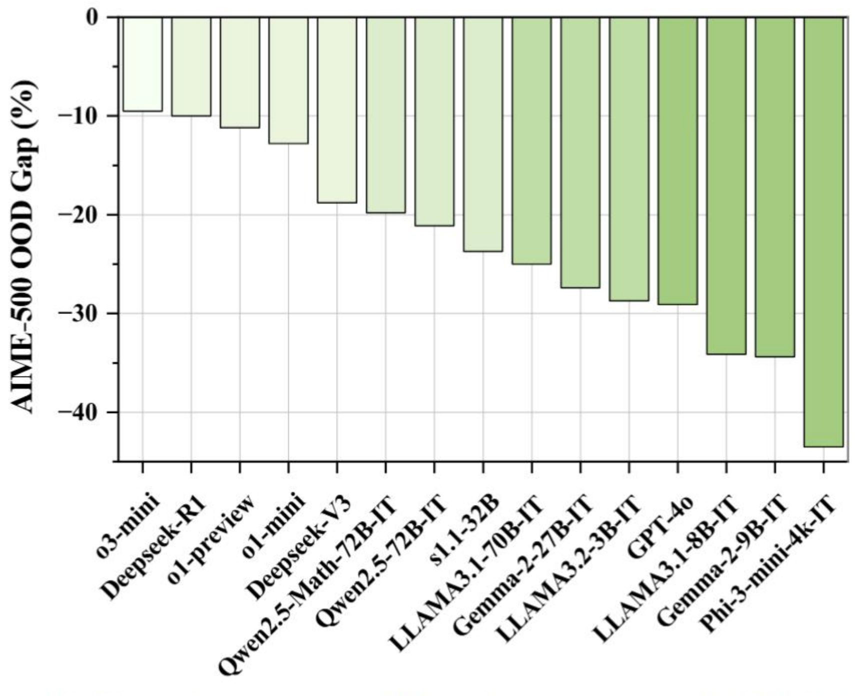
(a) OOD performance vs. ID performance for several reasoning models on AIME-500.



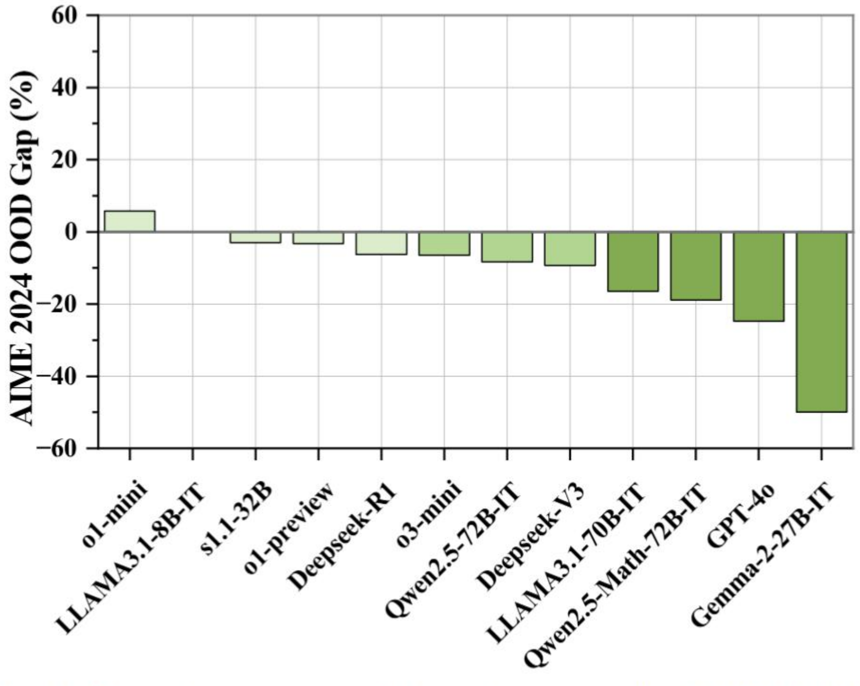
(b) OOD performance vs. ID performance for several reasoning models on AIME 2024.

Key Finding 2: Model Robustness Analysis

- **Reasoning Models:** o3-mini and Deepseek-R1 maintained high accuracy on OOD data and stayed closer to the diagonal, indicating stronger generalization and robustness.
- **Non-Reasoning Models:** GPT-4o and LLAMA3.1-70B-IT showed larger performance gaps and lower absolute accuracy on OOD tasks.
- This suggests a greater **reliance on memorization**.



(a) OOD performance vs. ID performance for AIME-500.



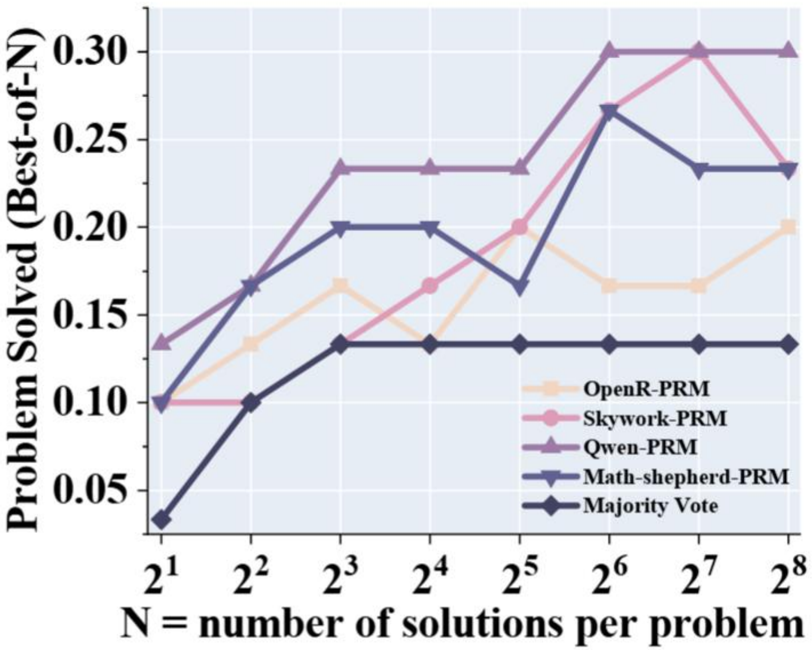
(b) OOD performance vs. ID performance for AIME 2024.

Experiment

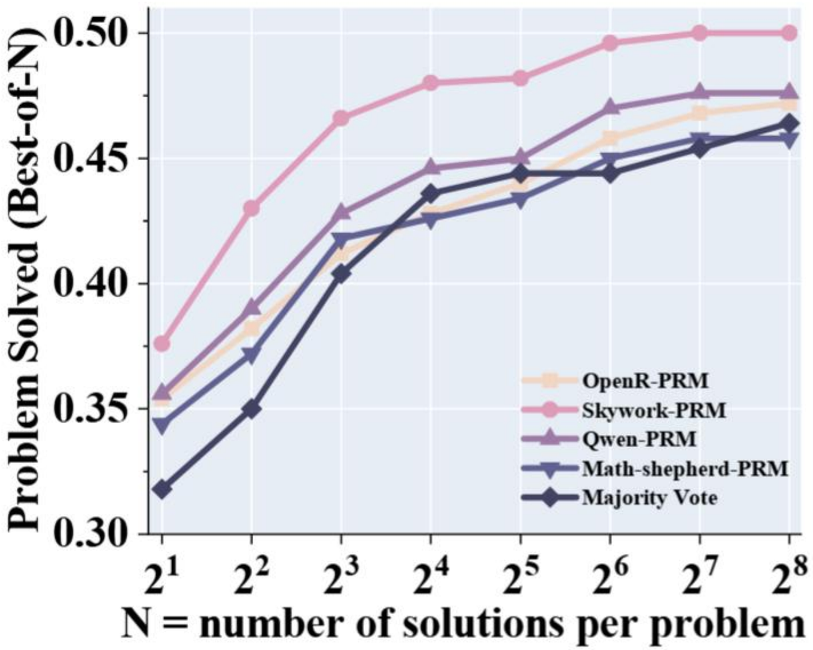
- Parameter count correlation with robustness

	AIME 2024			AIME-500		
	Original	OOD (Scenario)	OOD (Attack)	Original	OOD (Scenario)	OOD (Attack)
o1-preview	0.500	0.500	0.467	0.742	0.638	0.680
o1-mini	0.567	0.600	0.600	0.864	0.756	0.750
o3-mini	0.767	0.667	0.767	0.922	0.848	0.820
Deepseek-R1	0.800	0.733	0.767	0.920	0.816	0.840
GPT-4o	0.133	0.100	0.100	0.278	0.204	0.190
Deepseek-V3	0.367	0.333	0.333	0.528	0.438	0.420
Mixtral-8x7B-IT-v0.1	0.000	0.000	0.000	0.012	0.000	0.012
Qwen2.5-72B-IT	0.200	0.167	0.200	0.432	0.290	0.392
Qwen2.5-Math-72B-IT	0.267	0.233	0.200	0.536	0.360	0.500
LLAMA3.1-70B-IT	0.200	0.167	0.167	0.424	0.244	0.392
s1.1-32B	0.533	0.500	0.478	0.752	0.654	0.494
Gemma-2-27B-IT	0.033	0.033	0.000	0.062	0.028	0.062
Gemma-2-9B-IT	0.000	0.000	0.000	0.032	0.016	0.026
LLAMA3.1-8B-IT	0.000	0.033	0.000	0.132	0.074	0.100
Phi-3-mini-4k-IT	0.000	0.000	0.000	0.046	0.024	0.028
LLAMA3.2-3B-IT	0.033	0.033	0.033	0.122	0.066	0.108

Test-time Scaling Evaluation



(a) Performance on AIME 2024 OOD data.



(b) Performance on AIME-500 OOD data.

- Performance **improves with increased computation budget**.
- The performance demonstrates the **high quality** of the dataset dynamically constructed by ThinkBench.
- ThinkBench **reduces contamination impac**t and enables **reliable reasoning evaluation**.

Conclusion

- **ThinkBench**: A novel, dynamic OOD evaluation framework that effectively reduces the impact of data contamination.
- **Data Leakage**: Confirmed and quantified the data leakage problem in existing benchmarks.
- **Robustness Analysis**: Provided an in-depth analysis of the true reasoning ability of LLMs, finding that reasoning models are more robust.
- Provided a **High-Quality Benchmark**: The generated dataset is proven to be effective for evaluating advanced reasoning strategies like Test-time Scaling.

Thanks for listening !

ThinkBench: Dynamic Out-of-Distribution Evaluation for Robust LLM Reasoning

Shulin Huang, Linyi Yang*, Yan Song, Shuang Chen, Leyang Cui, Ziyu Wan,

Qingcheng Zeng, Ying Wen, Kun Shao, Weinan Zhang, Jun Wang, Yue Zhang*

huangshulin@westlake.edu.cn

Zhejiang University, Westlake University, University College London,

Shanghai Jiao Tong University, Northwestern University, Huawei Noah's Ark Lab