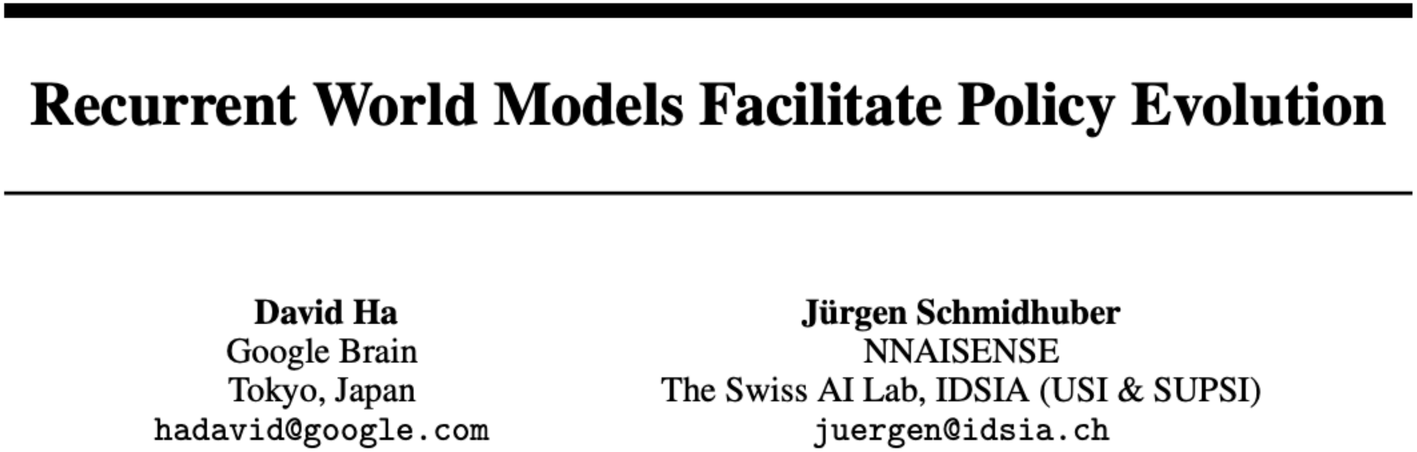
World Model & Inference Machine

Jiayu Yao

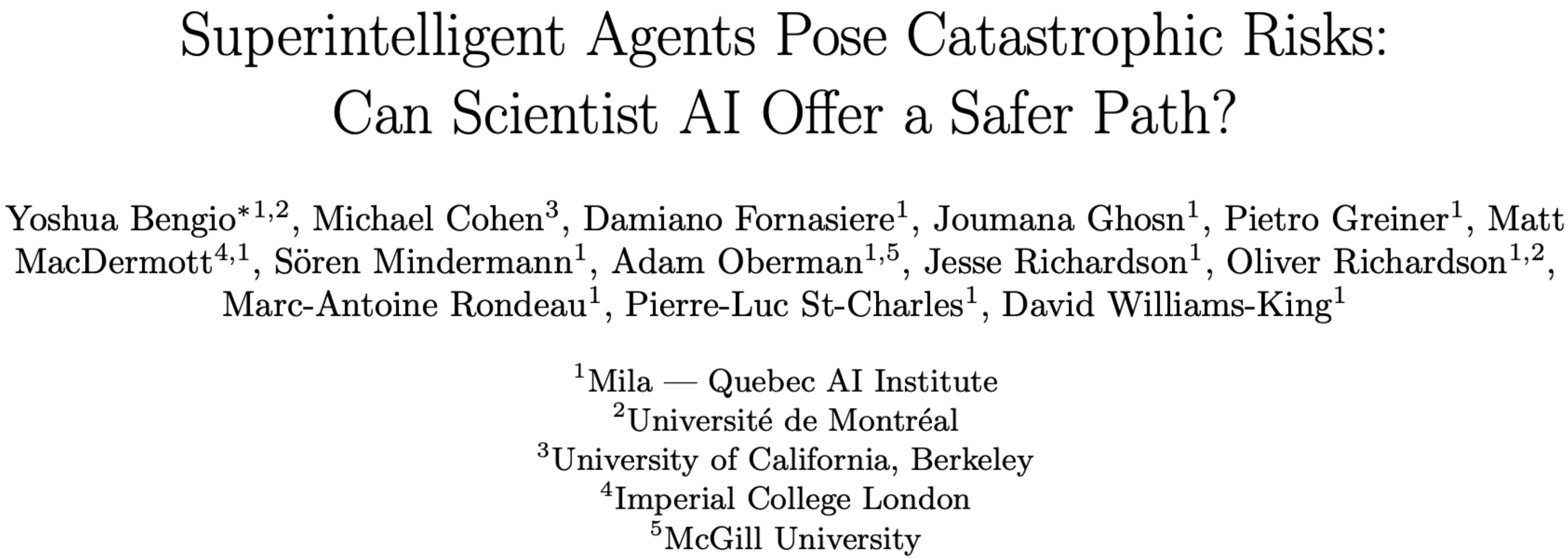
Fine-Tuning Is Not Enough



Recurrent World Models Facilitate Policy Evolution

David Ha
Google Brain
Tokyo, Japan
`hadavid@google.com`

Jürgen Schmidhuber
NNAISENSE
The Swiss AI Lab, IDSIA (USI & SUPSI)
`juergen@idsia.ch`



Superintelligent Agents Pose Catastrophic Risks: Can Scientist AI Offer a Safer Path?

Yoshua Bengio*[1,2], Michael Cohen[3], Damiano Fornasiere[1], Joumana Ghosn[1], Pietro Greiner[1], Matt MacDermott[4,1], Sören Mindermann[1], Adam Oberman[1,5], Jesse Richardson[1], Oliver Richardson[1,2], Marc-Antoine Rondeau[1], Pierre-Luc St-Charles[1], David Williams-King[1]

[1]Mila — Quebec AI Institute
[2]Université de Montréal
[3]University of California, Berkeley
[4]Imperial College London
[5]McGill University

World Model & Inference Machine

- Currently, training model to imitate or please humans (action or decision)

- World Model & Inference Machine, both components operate with an explicit notion of uncertainty to mitigate the risks of over-confident predictions (understanding)

World Model understands the world with dynamics function;
Inference Machine execute specific task

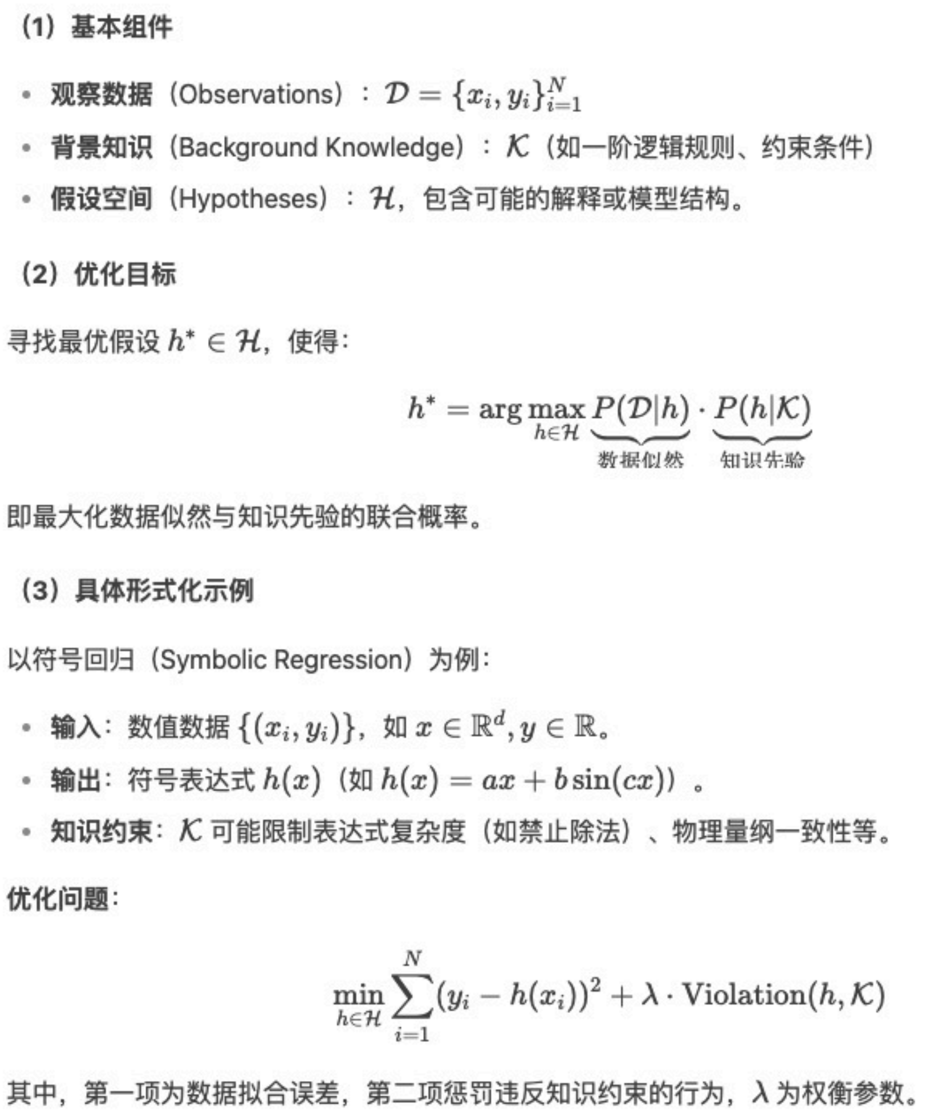
Inference Machine

- Goal Misspecification
  - It unfortunately appears impossible to formally articulate the difference between morally right and wrong behavior without enumerating all the possible cases. (必须在完备空间内评价目标)
  - overoptimization of a goal can yield disastrous outcomes
- Goal Misgeneralization
  - AI learns a goal that leads it to behave as intended during training and safety testing, but which diverges at deployment time.
  - reward-tampering
  - extreme severity and unknown likelihood of catastrophic risks

World Model

- Time horizons and anytime preparedness
  - an estimator of probabilistic bounds over worst-case scenarios that can result from the achievement of a user request.
  - develop a new training mechanism for the inference machine, grounded in a Bayesian framework and leveraging synthetic examples generated by the world model
- Bayesian
  - Uncertainty & Posterior predictive
  - Interpretable
- Model-Based
  - simulated environments

World Model & Inference Machine



（1）基本组件

- **观察数据**（Observations）：$\mathcal{D} = \{x_i, y_i\}_{i=1}^{N}$
- **背景知识**（Background Knowledge）：$\mathcal{K}$（如一阶逻辑规则、约束条件）
- **假设空间**（Hypotheses）：$\mathcal{H}$，包含可能的解释或模型结构。

（2）优化目标

寻找最优假设 $h^* \in \mathcal{H}$，使得：

$$h^* = \arg\max_{h \in \mathcal{H}} \underbrace{P(\mathcal{D}|h)}_{\text{数据似然}} \cdot \underbrace{P(h|\mathcal{K})}_{\text{知识先验}}$$

即最大化数据似然与知识先验的联合概率。

（3）具体形式化示例

以符号回归（Symbolic Regression）为例：

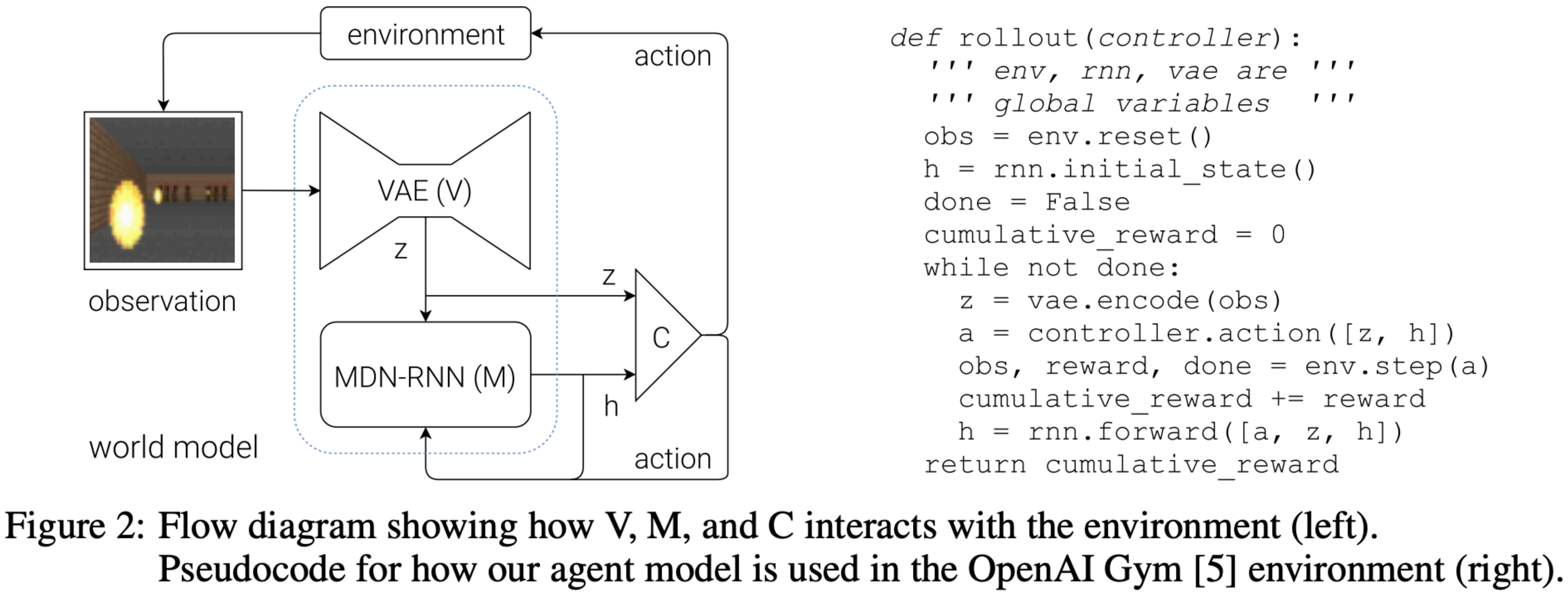
- **输入**：数值数据 $\{(x_i, y_i)\}$，如 $x \in \mathbb{R}^d, y \in \mathbb{R}$。
- **输出**：符号表达式 $h(x)$（如 $h(x) = ax + b\sin(cx)$）。
- **知识约束**：$\mathcal{K}$ 可能限制表达式复杂度（如禁止除法）、物理量纲一致性等。

优化问题：

$$\min_{h \in \mathcal{H}} \sum_{i=1}^{N} (y_i - h(x_i))^2 + \lambda \cdot \text{Violation}(h, \mathcal{K})$$

其中，第一项为数据拟合误差，第二项惩罚违反知识约束的行为，λ 为权衡参数。

World Model

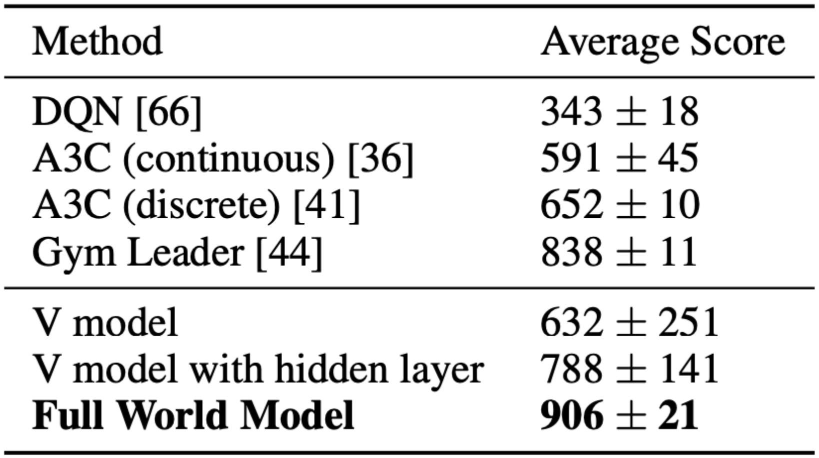


```
def rollout(controller):
  ''' env, rnn, vae are '''
  ''' global variables  '''
  obs = env.reset()
  h = rnn.initial_state()
  done = False
  cumulative_reward = 0
  while not done:
    z = vae.encode(obs)
    a = controller.action([z, h])
    obs, reward, done = env.step(a)
    cumulative_reward += reward
    h = rnn.forward([a, z, h])
  return cumulative_reward
```

Figure 2: Flow diagram showing how V, M, and C interacts with the environment (left). Pseudocode for how our agent model is used in the OpenAI Gym [5] environment (right).

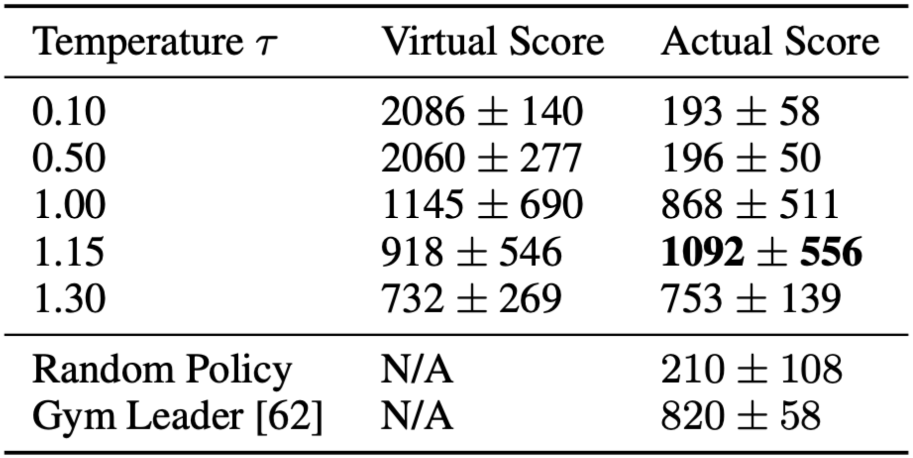
Fine-Tuning Is Not Enough

Table 1: `CarRacing-v0` results over 100 trials.



Method	Average Score
DQN [66]	343 ± 18
A3C (continuous) [36]	591 ± 45
A3C (discrete) [41]	652 ± 10
Gym Leader [44]	838 ± 11
V model	632 ± 251
V model with hidden layer	788 ± 141
Full World Model	**906** ± **21**

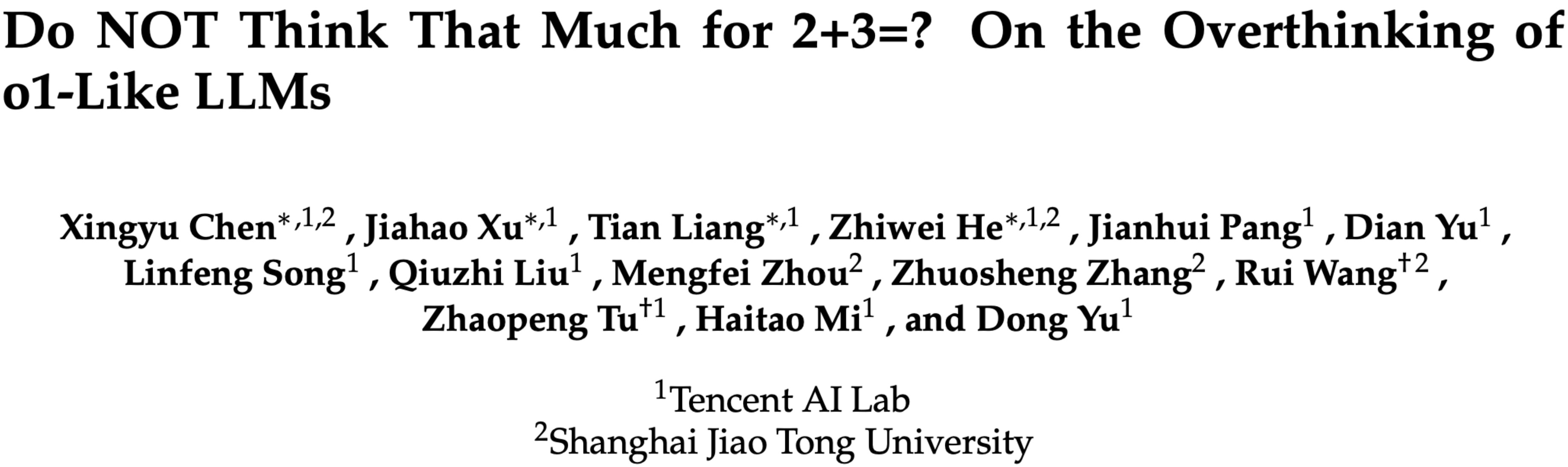
Table 2: `DoomTakeCover-v0` results, varying τ.



Temperature τ	Virtual Score	Actual Score
0.10	2086 ± 140	193 ± 58
0.50	2060 ± 277	196 ± 50
1.00	1145 ± 690	868 ± 511
1.15	918 ± 546	**1092** ± **556**
1.30	732 ± 269	753 ± 139
Random Policy	N/A	210 ± 108
Gym Leader [62]	N/A	820 ± 58

https://worldmodels.github.io/

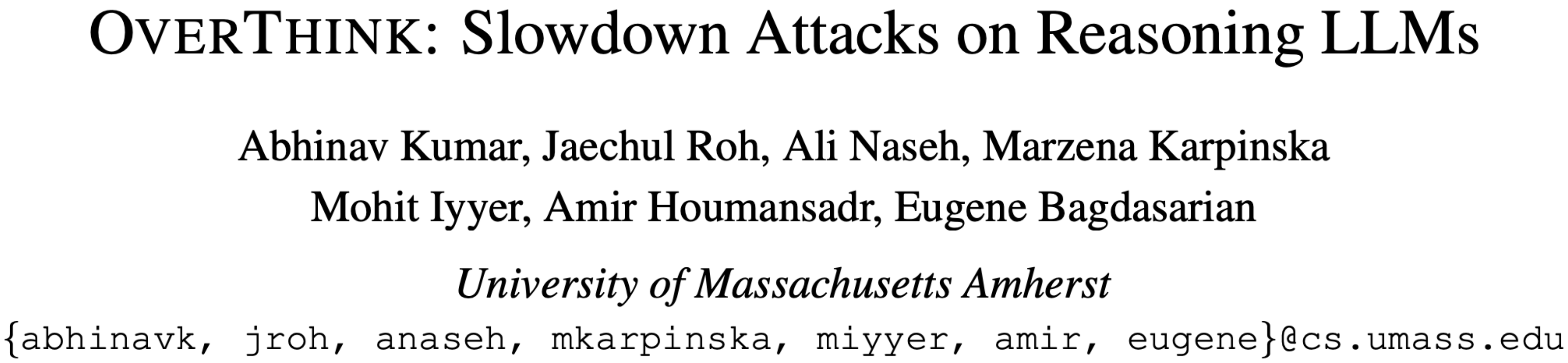
Discussion Reasoning Attack



Do NOT Think That Much for 2+3=? On the Overthinking of o1-Like LLMs

Xingyu Chen[*,1,2] **, Jiahao Xu**[*,1] **, Tian Liang**[*,1] **, Zhiwei He**[*,1,2] **, Jianhui Pang**[1] **, Dian Yu**[1] **,**
Linfeng Song[1] **, Qiuzhi Liu**[1] **, Mengfei Zhou**[2] **, Zhuosheng Zhang**[2] **, Rui Wang**[†2] **,**
Zhaopeng Tu[†1] **, Haitao Mi**[1] **, and Dong Yu**[1]

[1]Tencent AI Lab
[2]Shanghai Jiao Tong University



OVERTHINK: Slowdown Attacks on Reasoning LLMs

Abhinav Kumar, Jaechul Roh, Ali Naseh, Marzena Karpinska
Mohit Iyyer, Amir Houmansadr, Eugene Bagdasarian

University of Massachusetts Amherst
`{abhinavk, jroh, anaseh, mkarpinska, miyyer, amir, eugene}@cs.umass.edu`

Discussion Reasoning Attack

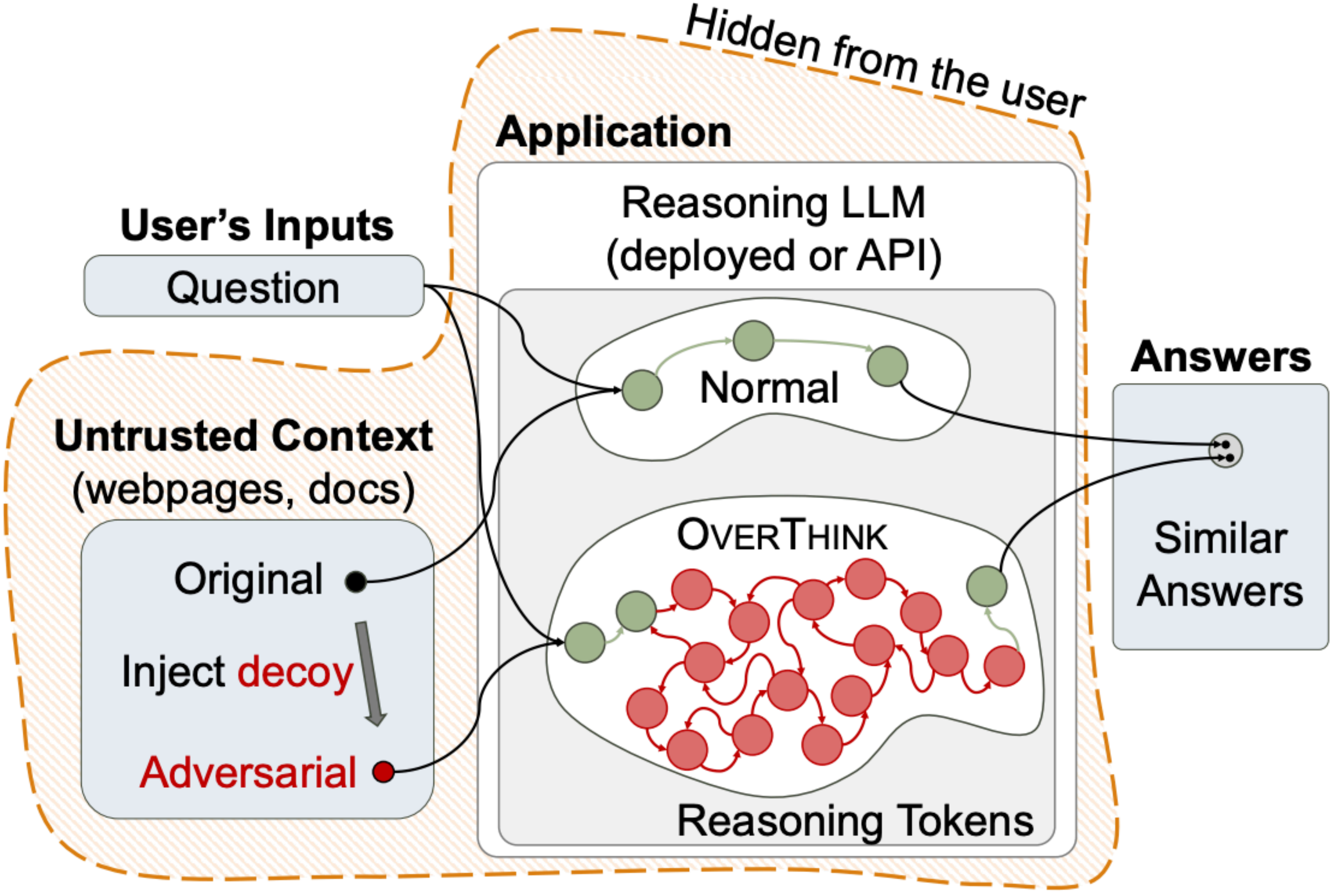
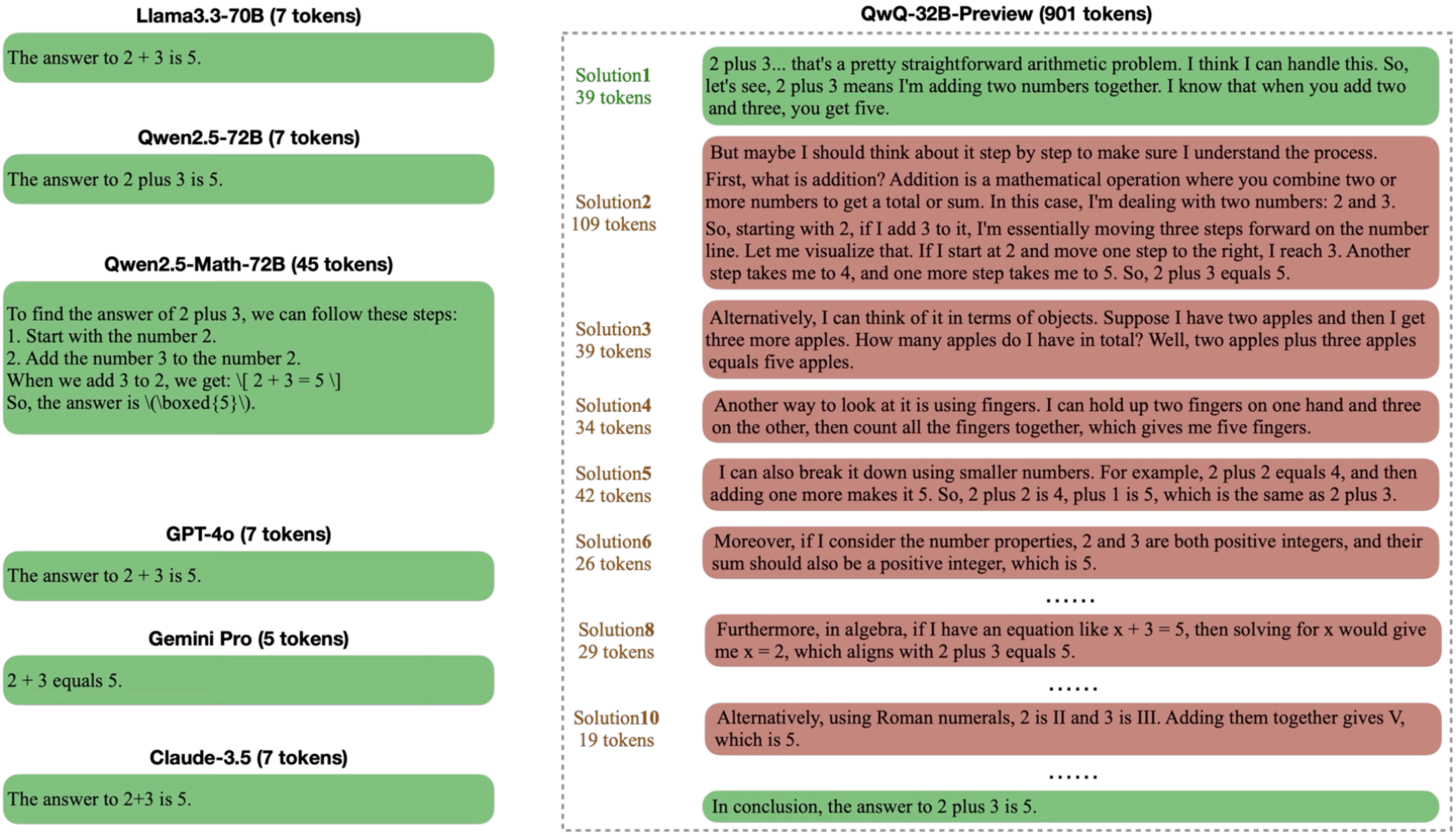


Figure 1. Overview of OVERTHINK Attack.

Discussion Reasoning Attack



Llama3.3-70B (7 tokens)

The answer to 2 + 3 is 5.

Qwen2.5-72B (7 tokens)

The answer to 2 plus 3 is 5.

Qwen2.5-Math-72B (45 tokens)

To find the answer of 2 plus 3, we can follow these steps:
1. Start with the number 2.
2. Add the number 3 to the number 2.
When we add 3 to 2, we get: \[2 + 3 = 5 \]
So, the answer is \(\boxed{5}\).

GPT-4o (7 tokens)

The answer to 2 + 3 is 5.

Gemini Pro (5 tokens)

2 + 3 equals 5.

Claude-3.5 (7 tokens)

The answer to 2+3 is 5.

QwQ-32B-Preview (901 tokens)

Solution**1**
39 tokens

2 plus 3... that's a pretty straightforward arithmetic problem. I think I can handle this. So, let's see, 2 plus 3 means I'm adding two numbers together. I know that when you add two and three, you get five.

Solution**2**
109 tokens

But maybe I should think about it step by step to make sure I understand the process.
First, what is addition? Addition is a mathematical operation where you combine two or more numbers to get a total or sum. In this case, I'm dealing with two numbers: 2 and 3.
So, starting with 2, if I add 3 to it, I'm essentially moving three steps forward on the number line. Let me visualize that. If I start at 2 and move one step to the right, I reach 3. Another step takes me to 4, and one more step takes me to 5. So, 2 plus 3 equals 5.

Solution**3**
39 tokens

Alternatively, I can think of it in terms of objects. Suppose I have two apples and then I get three more apples. How many apples do I have in total? Well, two apples plus three apples equals five apples.

Solution**4**
34 tokens

Another way to look at it is using fingers. I can hold up two fingers on one hand and three on the other, then count all the fingers together, which gives me five fingers.

Solution**5**
42 tokens

I can also break it down using smaller numbers. For example, 2 plus 2 equals 4, and then adding one more makes it 5. So, 2 plus 2 is 4, plus 1 is 5, which is the same as 2 plus 3.

Solution**6**
26 tokens

Moreover, if I consider the number properties, 2 and 3 are both positive integers, and their sum should also be a positive integer, which is 5.

……

Solution**8**
29 tokens

Furthermore, in algebra, if I have an equation like x + 3 = 5, then solving for x would give me x = 2, which aligns with 2 plus 3 equals 5.

……

Solution**10**
19 tokens

Alternatively, using Roman numerals, 2 is II and 3 is III. Adding them together gives V, which is 5.

……

In conclusion, the answer to 2 plus 3 is 5.

Discussion Reasoning Attack

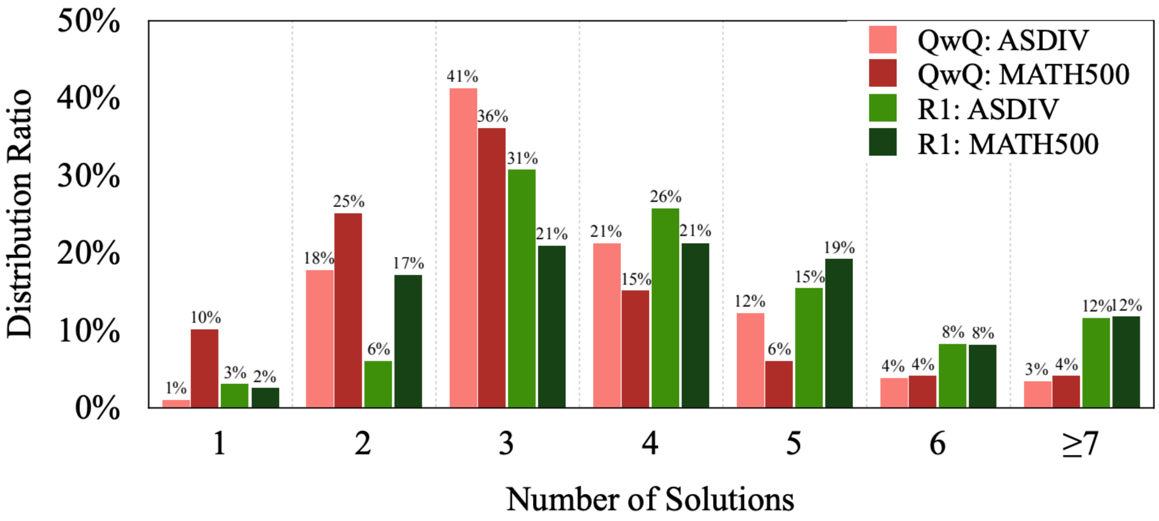
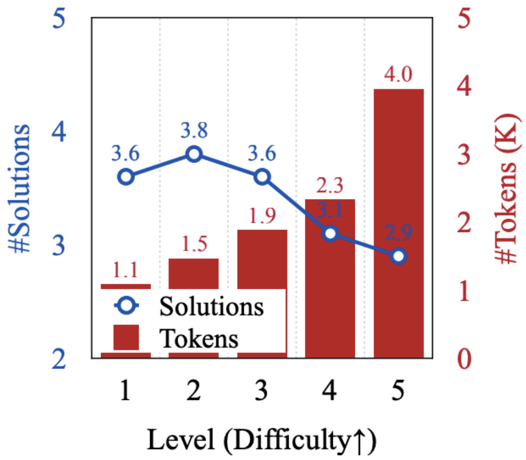
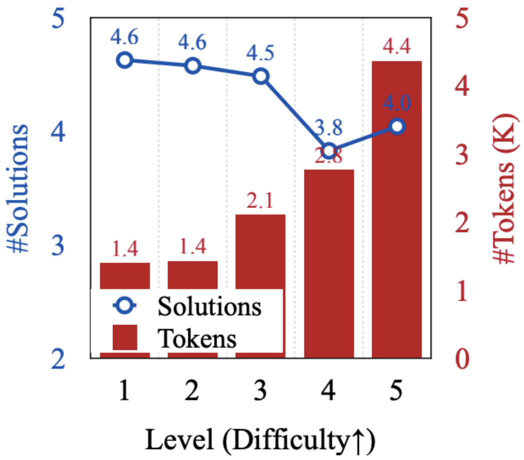


Figure 3: Distribution of solution counts in generated responses for different test sets and models (QwQ-32B-Preview (“QwQ”) and DeepSeek-R1 (“R1”)).



(a) QwQ-32B-Preview



(b) DeepSeek-R1

Discussion Reasoning Attack

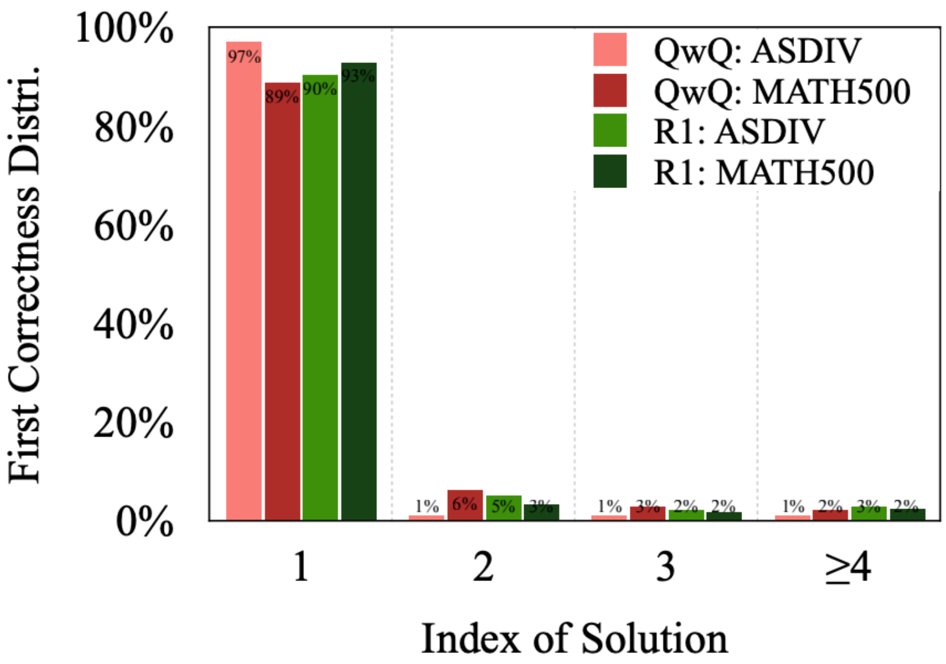


Figure 5: Distribution of occurrences for the first correct answer.

- Notably, the first round generally comprises less than 60% of the total tokens generated, suggesting that the extended CoT might not significantly enhance accuracy.
- later solutions marginally contribute to improvements in accuracy.

Discussion Reasoning Attack

$$\xi_O = \frac{1}{N}\sum_{i=1}^{N}\sigma_i\frac{\hat{T}_i}{T_i}$$

Outcome Efficiency

$$\hat{T}_i = \begin{cases} \text{\#tokens to first arrive at correct answer}, & \sigma_i = 1 \\ T_i, & \sigma_i = 0 \end{cases}$$

$$\sigma_i = \begin{cases} 1, & \text{if at least one solution in response is correct} \\ 0, & \text{otherwise} \end{cases}$$

Discussion Reasoning Attack

Solution Set

$$R_i = \{s_i^1, \ldots, s_i^m, \ldots, s_i^{M_i}\}$$

Response Set

$$S^m = \{s_1^m, \ldots, s_k^m, \ldots, s_K^m\}$$

$$\text{Dis}^m = \frac{\sum_{k=1}^{K} \tau_k^m}{K}$$

$$\tau_k^m = \begin{cases} 1, & \text{if } \Phi(s_k^m) \not\subseteq \{\Phi(s_k^1), \ldots, \Phi(s_k^{m-1})\} \\ 0, & \text{otherwise} \end{cases}$$

Diversity Efficiency

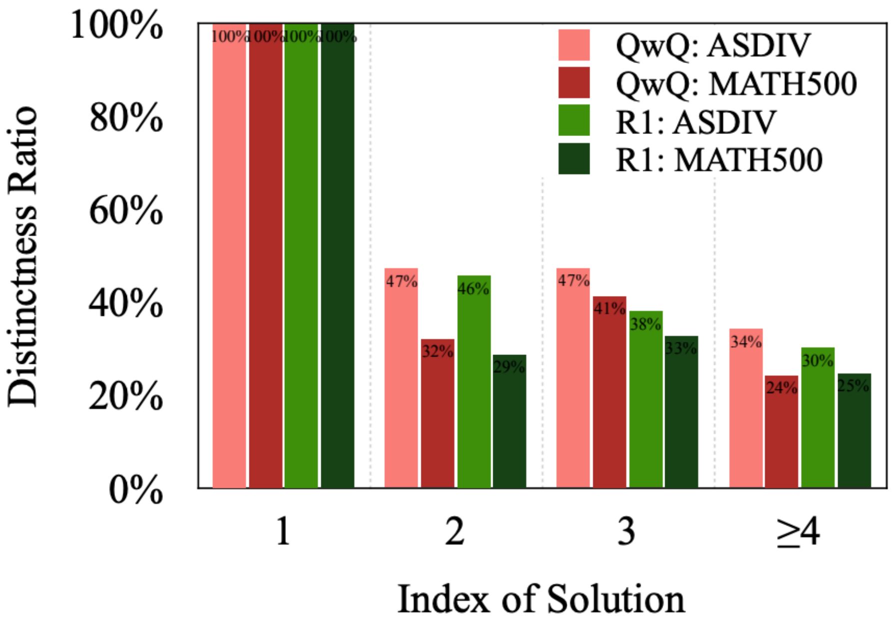


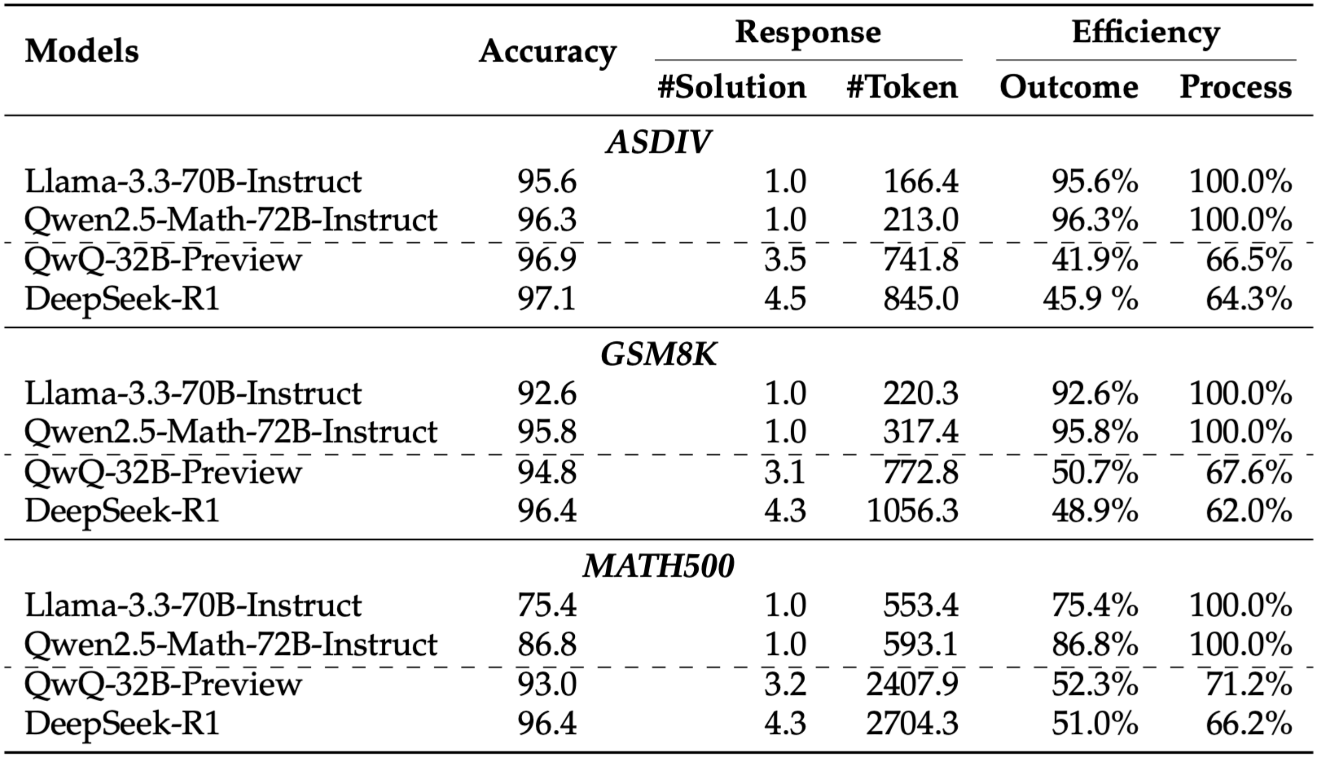
Figure 6: Ratio of whether a solution provides a new reasoning strategy for each index.

Discussion Reasoning Attack

$$D_i = \sum_{m=1}^{M} \tau_i^m T_i^m$$

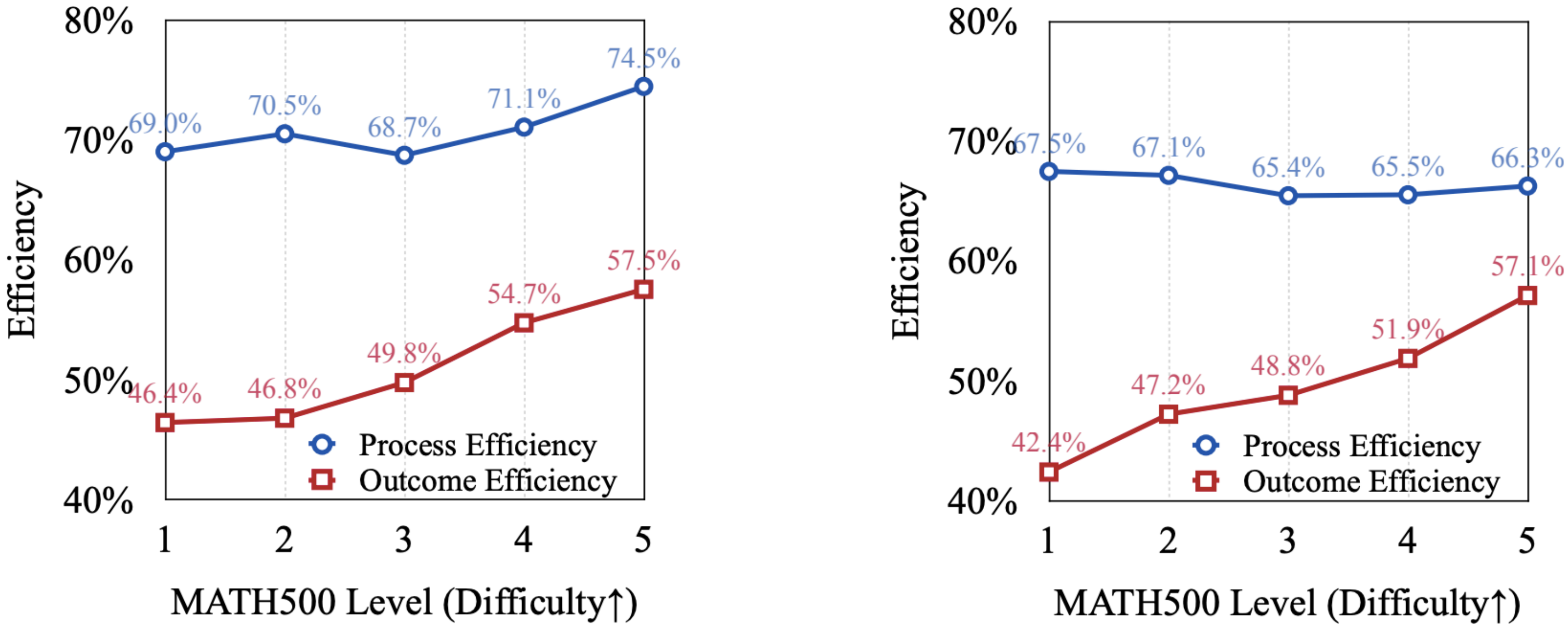
Diversity Efficiency

Table 1: Model efficiency results of strong LLMs.



Models	**Accuracy**	**Response**		**Efficiency**	
		#Solution	**#Token**	**Outcome**	**Process**
			ASDIV		
Llama-3.3-70B-Instruct	95.6	1.0	166.4	95.6%	100.0%
Qwen2.5-Math-72B-Instruct	96.3	1.0	213.0	96.3%	100.0%
QwQ-32B-Preview	96.9	3.5	741.8	41.9%	66.5%
DeepSeek-R1	97.1	4.5	845.0	45.9 %	64.3%
			GSM8K		
Llama-3.3-70B-Instruct	92.6	1.0	220.3	92.6%	100.0%
Qwen2.5-Math-72B-Instruct	95.8	1.0	317.4	95.8%	100.0%
QwQ-32B-Preview	94.8	3.1	772.8	50.7%	67.6%
DeepSeek-R1	96.4	4.3	1056.3	48.9%	62.0%
			MATH500		
Llama-3.3-70B-Instruct	75.4	1.0	553.4	75.4%	100.0%
Qwen2.5-Math-72B-Instruct	86.8	1.0	593.1	86.8%	100.0%
QwQ-32B-Preview	93.0	3.2	2407.9	52.3%	71.2%
DeepSeek-R1	96.4	4.3	2704.3	51.0%	66.2%

Discussion Reasoning Attack



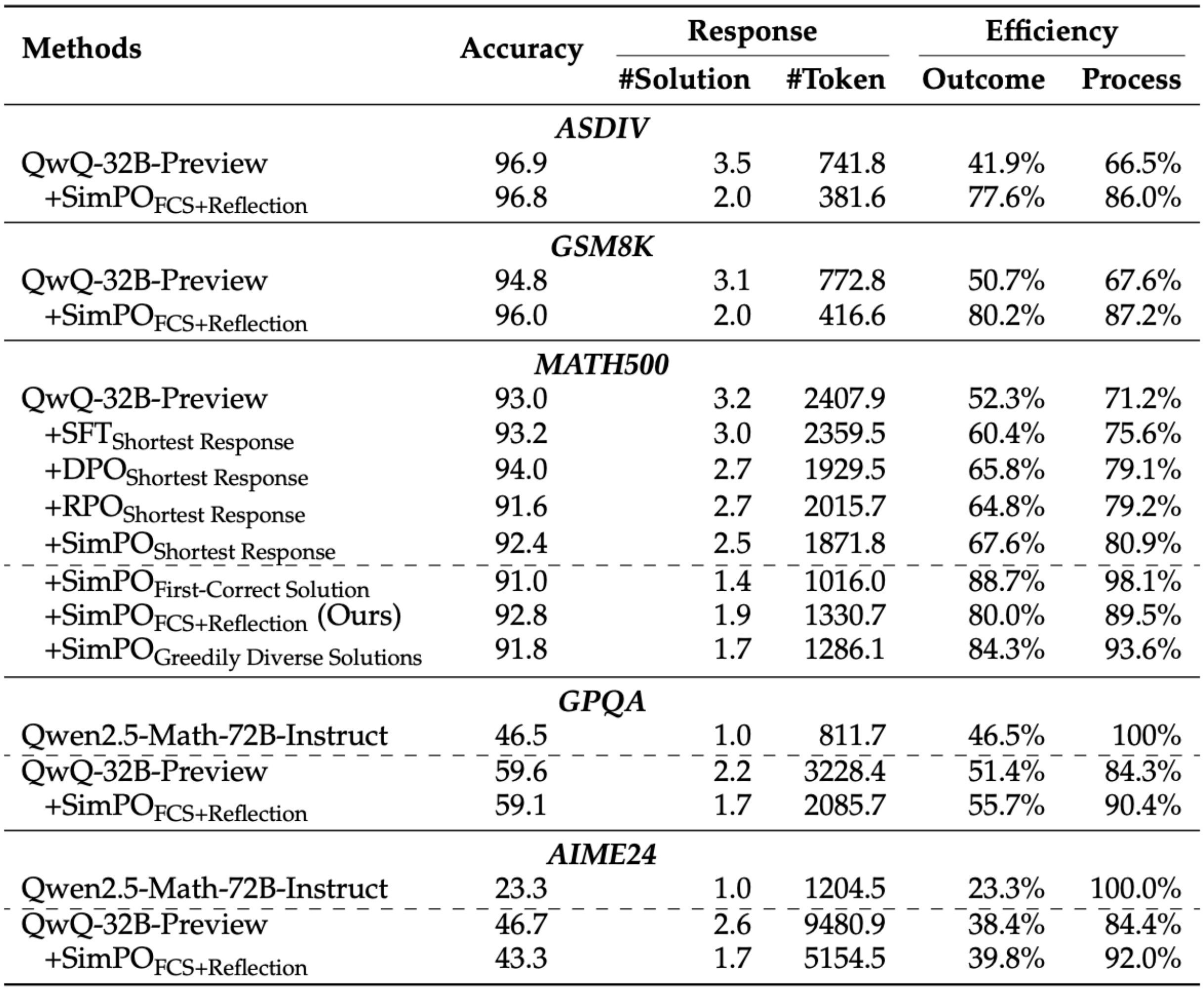
(a) QwQ-32B-Preview

(b) DeepSeek-R1

Figure 7: Efficiency results of (a) QwQ-32B-Preview and (b) DeepSeek-R1 across different difficulty levels of the MATH500 testset.

Discussion Reasoning Attack

Table 4: Experimental results of the proposed efficiency enhancing methods.



Methods	Accuracy	Response		Efficiency	
		#Solution	**#Token**	**Outcome**	**Process**
ASDIV					
QwQ-32B-Preview	96.9	3.5	741.8	41.9%	66.5%
+SimPO$_{\text{FCS+Reflection}}$	96.8	2.0	381.6	77.6%	86.0%
GSM8K					
QwQ-32B-Preview	94.8	3.1	772.8	50.7%	67.6%
+SimPO$_{\text{FCS+Reflection}}$	96.0	2.0	416.6	80.2%	87.2%
MATH500					
QwQ-32B-Preview	93.0	3.2	2407.9	52.3%	71.2%
+SFT$_{\text{Shortest Response}}$	93.2	3.0	2359.5	60.4%	75.6%
+DPO$_{\text{Shortest Response}}$	94.0	2.7	1929.5	65.8%	79.1%
+RPO$_{\text{Shortest Response}}$	91.6	2.7	2015.7	64.8%	79.2%
+SimPO$_{\text{Shortest Response}}$	92.4	2.5	1871.8	67.6%	80.9%
+SimPO$_{\text{First-Correct Solution}}$	91.0	1.4	1016.0	88.7%	98.1%
+SimPO$_{\text{FCS+Reflection}}$ (Ours)	92.8	1.9	1330.7	80.0%	89.5%
+SimPO$_{\text{Greedily Diverse Solutions}}$	91.8	1.7	1286.1	84.3%	93.6%
GPQA					
Qwen2.5-Math-72B-Instruct	46.5	1.0	811.7	46.5%	100%
QwQ-32B-Preview	59.6	2.2	3228.4	51.4%	84.3%
+SimPO$_{\text{FCS+Reflection}}$	59.1	1.7	2085.7	55.7%	90.4%
AIME24					
Qwen2.5-Math-72B-Instruct	23.3	1.0	1204.5	23.3%	100.0%
QwQ-32B-Preview	46.7	2.6	9480.9	38.4%	84.4%
+SimPO$_{\text{FCS+Reflection}}$	43.3	1.7	5154.5	39.8%	92.0%

Thanks