Amplification Effects in Test-Time Reinforcement Learning: Safety and Reasoning Vulnerabilities

Khattar, Vanshaj; Choudhury, Moumita; Rashid, Md Rafi Ur; Liu, Jing; Koike-Akino, Toshiaki; Jin, Ming; Wang, Ye

TR2026-020 January 28, 2026

Abstract

Test-time training (TTT) has recently emerged as a promising method to improve the reasoning abilities of large language models (LLMs), in which the model directly learns from test data without access to labels. However, this reliance on test data also makes TTT methods vulnerable to harmful prompt injections. In this paper, we investigate safety vulnerabilities of TTT methods, where we specifically consider test-time reinforcement learning (TTRL) (Zuo et al. 2025), a recent TTT method that improves LLM reasoning by rewarding self-consistency using majority vote as a reward signal. We show that harmful prompt injection during TTRL ampli- fies the model's existing behaviors, i.e., safety amplification when the base model is relatively safe, and harmfulness amplification when it is vulnerable to the injected data. In both cases, there is a decline in reasoning ability, which we refer to as the reasoning tax. We also show that TTRL can be exploited adversarially using specially designed "HarmInject" prompts to force the model to answer jailbreak and reasoning queries together, resulting in stronger harmfulness amplification. Overall, our results highlight that TTT methods that enhance LLM reasoning by promoting self-consistency can lead to amplification behaviors and reasoning degradation, high- lighting the need for safer TTT methods.

AAAI Workshop on Trust and Control in Agentic AI 2026

Mitsubishi Electric Research Laboratories, Inc.
201 Broadway, Cambridge, Massachusetts 02139

Amplification Effects in Test-Time Reinforcement Learning: Safety and Reasoning Vulnerabilities

Vanshaj Khattar[1]**, Moumita Choudhury**[2]**, Md Rafi Ur Rashid**[3]**,**
Jing Liu[4]**, Toshiaki Koike-Akino**[4]**, Ming Jin**[1]**, Ye Wang**[4]
[1]Virginia Tech, [2]UMass Amherst, [3]Pennsylvania State University, [4]Mitsubishi Electric Research Laboratories

Abstract

Test-time training (TTT) has recently emerged as a promising method to improve the reasoning abilities of large language models (LLMs), in which the model directly learns from test data without access to labels. However, this reliance on test data also makes TTT methods vulnerable to harmful prompt injections. In this paper, we investigate safety vulnerabilities of TTT methods, where we specifically consider test-time reinforcement learning (TTRL) (Zuo et al. 2025), a recent TTT method that improves LLM reasoning by rewarding self-consistency using majority vote as a reward signal. We show that harmful prompt injection during TTRL amplifies the model's existing behaviors, i.e., **safety amplification** when the base model is relatively safe, and **harmfulness amplification** when it is vulnerable to the injected data. In both cases, there is a decline in reasoning ability, which we refer to as the **reasoning tax**. We also show that TTRL can be exploited adversarially using specially designed "HarmInject" prompts to force the model to answer jailbreak and reasoning queries together, resulting in stronger harmfulness amplification. Overall, our results highlight that TTT methods that enhance LLM reasoning by promoting self-consistency can lead to amplification behaviors and reasoning degradation, highlighting the need for safer TTT methods.

1 Introduction

The reasoning abilities of Large language models (LLMs) have continued to improve through both supervised finetuning (SFT) and reinforcement learning (RL) (Guo et al. 2025; Zhang et al. 2025) methods. Despite these gains, current LLMs still struggle with reasoning on out-of-distribution tasks (Phan et al. 2025). To address this limitation and improve generalization to unseen problems, a growing line of work has explored test-time training (TTT) (Zuo et al. 2025; Prabhudesai et al. 2025; Zhao et al. 2025; Jang et al. 2025), which adapts models directly on the test inputs without having access to the labels. These methods have already shown improvements in arithmetic reasoning (Li et al. 2024; Hendrycks et al.), commonsense QA (Rein et al. 2024), and spatial reasoning (Akyürek et al. 2024).

However, TTT operates entirely on the prompts observed during test time, which introduces a new attack vulnerability: adversaries can manipulate the test-time data to influence the model's parameters during TTT (Cong et al. 2024; Su et al. 2024). In this work, we investigate the safety vulnerabilities of the TTT approaches that aim to improve LLM reasoning by promoting self-consistency. We specifically consider the test-time reinforcement learning (TTRL) algorithm (Zuo et al. 2025), a recent TTT method that uses RL for TTT, and majority voting to compute reward. We consider the setting where an adversary can inject harmful jailbreak prompts into the test-time data during TTRL.

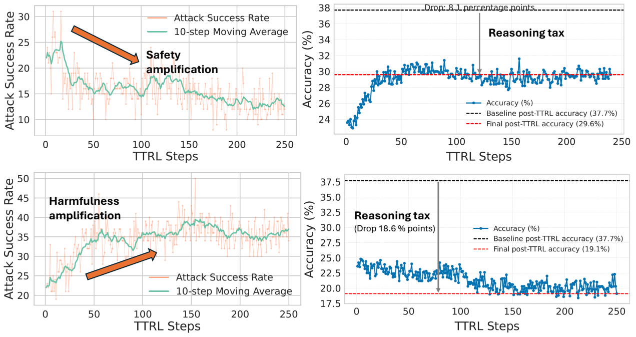


Figure 1: **Safety and harmfulness amplification** during TTRL. Top left: attack success rate (ASR) of Jailbreak-V28k prompts on Qwen1.5B-Instruct when Jailbreak-V28k prompts are injected into AMC test-time data. Top right: the resulting **reasoning tax**, i.e., loss in AMC accuracy. Bottom left: ASR for Qwen-1.5B-Instruct on JailbreakV-28k when HarmInject prompts (Section 3.3) are injected. Bottom right: Impact on reasoning performance post-TTRL on Qwen-1.5BB-Instruct model.

Our experiments reveal a striking asymmetry. On jailbreak datasets where the base model is already relatively safe (e.g., 22% harmfulness rate on JailbreakV-28k (Luo et al. 2024) for Qwen-1.5B-Instruct (Yang et al. 2025)), TTRL leads to an emergent safety effect via **safety amplification** as shown in the top-left panel of Figure 1. However, the same TTRL updates incur a **reasoning tax**, shown in the top-right panel, as accuracy on AMC math problems declines despite safety gains. When the base model instead shows high susceptibility to a particular jailbreak attack (e.g., Qwen-1.5B-Instruct (Dubey et al. 2024) on HarmInject attack developed in Section 3.2), the pattern reverses. As shown in the bottom-left panel, TTRL amplifies harmful behavior, with attack success rates rising. Moreover, the reasoning performance again deteriorates as seen in the lower right panel of Figure 1. Across both cases, the mechanism is consistent: majority-vote rewards reinforce the dominant model behavior on the injected dataset, while also deteriorating the reasoning performance.

2 Background

Test-Time Training (TTT) is a method for adapting pretrained models at test time to improve its generalization performance. Suppose a model f_θ has been trained on a source distribution $\mathcal{D}_s = \{(x_i, y_i)\}_{i=1}^{N}$, where θ denotes the model parameters. At test-time, the model is evaluated on samples $x_t \sim \mathcal{D}_t$ drawn from a target distribution that differs from the training distribution ($\mathcal{D}_t \neq \mathcal{D}_s$). TTT adapts the model at test-time by updating parameters with respect to an auxiliary objective $\mathcal{L}_{aux}$ that is optimized on each test sample. Therefore, TTT allows the model to refine itself during deployment, without access to the ground-truth labels.

Test-time reinforcement learning (TTRL) (Zuo et al. 2025) is a recent TTT method that uses RL for TTT to improve the reasoning abilities of LLMs. For each test input x, the model generates K candidate responses $\{y_1, \ldots, y_K\}$ by sampling from its current policy π_θ. A majority-voting aggregator selects the majority answer $\hat{y}$ across the K samples. This serves as a pseudo-reward, where responses matching $\hat{y}$ receive a positive reward, while others receive a zero reward. The model parameters are then updated via Group Reference Policy Optimization (GRPO) (Shao et al. 2024). Formally, the reward for the k-th generated response y_k is given as $r(y_k) = 1$, when $y_k = \hat{y}$, and otherwise zero, and the policy is trained to maximize expected reward by updating the LLM parameters θ to maximize the following expected return:

$$\nabla_\theta J(\theta) = \mathbb{E}_{y \sim \pi_\theta(\cdot|x)} \big[r(y) \nabla_\theta \log \pi_\theta(y|x) \big].$$

where parameters θ are updated using gradient ascent. As TTT methods such as TTRL rely on the test data inputs and the generated pseudo-rewards, their behavior depends strongly on what appears in the test-time data, making it vulnerable to prompt injection, which motivates the safety analysis presented in this work.

3 Experimental setup and results

Models and training datasets. We consider two instruction-tuned models: Qwen2.5-1.5B-Instruct (Yang et al. 2025) and Llama-3-8B-Instruct (Dubey et al. 2024). For the harmful jailbreak datasets, we use the JailbreakV-28k (Luo et al. 2024), Llama-jailbreak artifacts (Andriushchenko, Croce, and Flammarion 2024) specifically tuned to jailbreak the Llama3-8B-Instruct model, and the in-the-wild jailbreak dataset (Jiang et al. 2024). We conduct all experiments on the AMC math reasoning dataset (Li et al. 2024).

Evaluation metrics. We consider two axes of evaluation: safety and reasoning accuracy. For the reasoning performance, we use the same metric as reported in the TTRL paper (Zuo et al. 2025), i.e., *pass@1* estimated from k responses, generated with a non-zero temperature of 0.6, top-p value of 0.95, as given by $pass@1 := c/k$, where c is the number of correct responses. We use $k = 16$ in our experiments. Each TTRL run is done for 250 steps. To evaluate the safety of the model's generated responses, we use the attack success rate (ASR) percentage of the jailbreak attack on the model, which measures the percentage of harmful responses to the total number of jailbreak prompts. We use the LlamaGuard3-8B model (Inan et al. 2023) as a safety judge to evaluate the harmfulness of the model's responses.

We structure the experimental results into three research questions: **RQ1:** Does TTRL on benign data increase model harmfulness?; **RQ2:** What is the impact of harmful prompt injection during TTRL?; **RQ3:** Can TTRL be exploited to amplify the harmfulness?

3.1 RQ1: Does TTRL on benign data increase model harmfulness?

Figure 2 reports ASR across TTRL steps for both Qwen-1.5B-Instruct and Llama-3-8B-Instruct when the test-time training data contains only AMC reasoning problems. For Qwen, in Figure 2a, ASR on JailbreakV-28k fluctuates between 21% and 25% (baseline 22%). In Figure 2b and 2c, similar small variations appear on the WildJailbreak and Llama artifact attacks, with no upward trend across TTRL steps. Therefore, as shown in some previous works on fine-tuning (Qi et al. 2024), where harmfulness may increase unintentionally, TTRL on benign test data leaves harmfulness largely unchanged.

3.2 RQ2: What is the impact of harmful prompt injection during TTRL?

Next, we investigate the case when the test-time training data is injected with harmful prompts. Figure 3 and 4 report the effect of harmful prompt injection across three jailbreak datasets for the Qwen and Llama models, respectively. For Qwen, the base model is already moderately safe with an ASR of 22% on JailbreakV-28k as seen in Figure 3a and 40% on WildJailbreak prompts in Figure 3b. Under TTRL with harmful injection, the ASR starts to decline, which we term **safety amplification**. The Llama model shows similar safety amplification on the WildJailbreak dataset in Figure 4b. The same safety amplification effect is not present in the JailbreakV-28k dataset in Figure 4a, where its baseline ASR is already low ($\approx 1\%$); and the safety remains stable.

The Llama Artifacts dataset presents the opposite case. The Llama model by default is highly vulnerable to these specifically tuned attacks, with ASR exceeding 90%. Under TTRL, the harmfulness gets amplified: ASR rises slightly or stays near its high baseline as seen in Figure 4c. We call this **weak harmfulness amplification**. For the Qwen model, the ASR is above 80% against the Llama Artifact prompts in Figure 3c, where we also see the slight effect of weak harmfulness amplification.

Across all settings above, the reasoning ability of the base model post-TTRL degrades compared to the post-TTRL accuracies achieved without any injection. In Figures 3d - 3f,

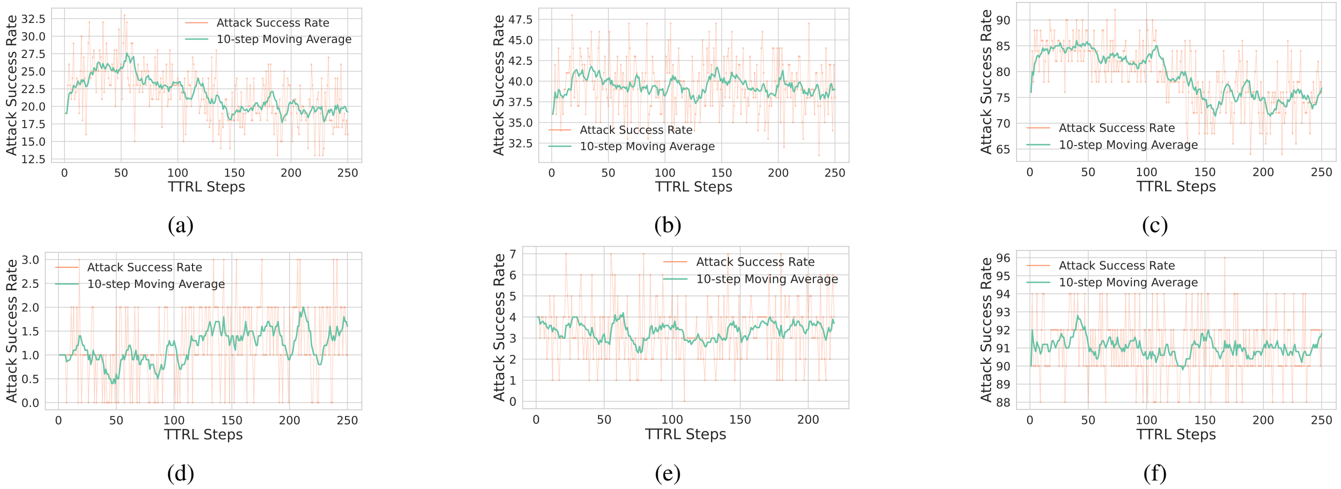


Figure 2: ASR measured across three jailbreak datasets, JailbreakV-28k, WildJailbreak, and Llama Artifacts (left to right, respectively) during TTRL, for Qwen-1.5B-Instruct (top row) and Llama-3-8B-Instruct (bottom row).

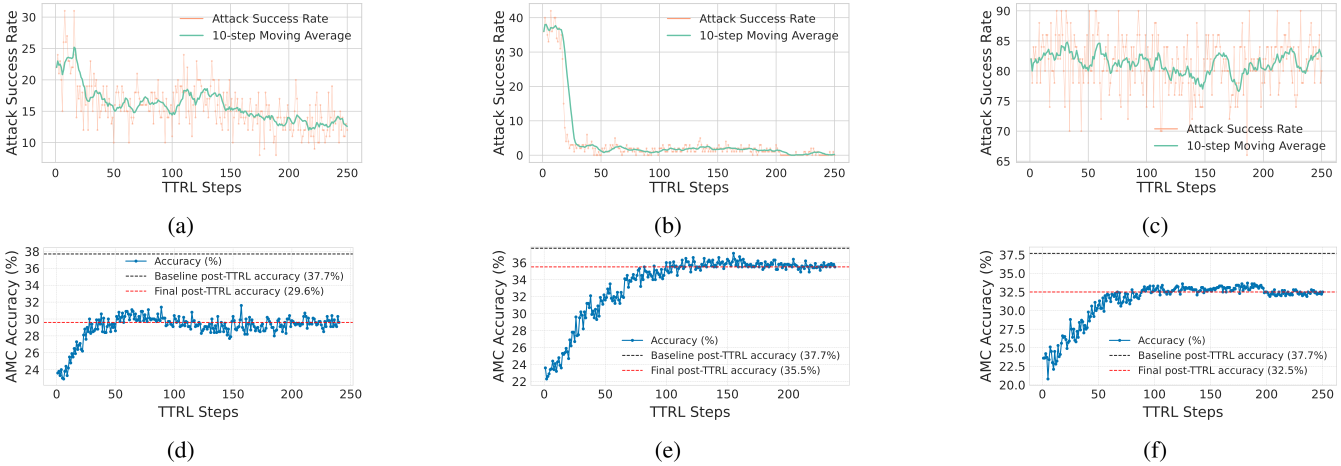


Figure 3: ASR and reasoning during TTRL for Qwen-1.5B-Instruct model after harmful prompt injection across, JailbreakV-28k, WildJailbreak, and Llama Artifacts datasets (left to right, respectively), ASR (top row) and AMC accuracy (bottom row).

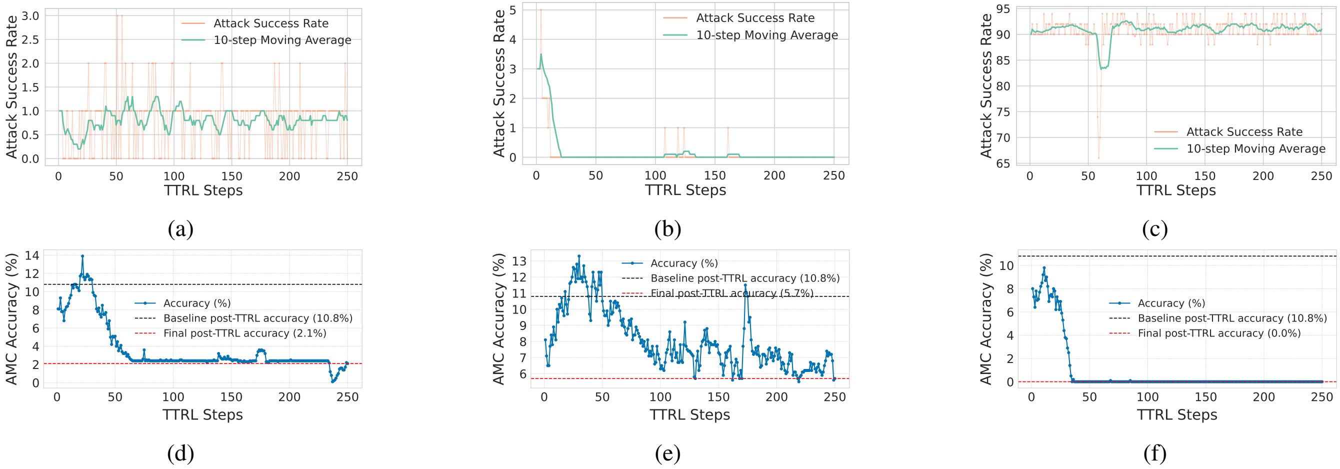
AMC accuracy falls from the baseline post-TTRL accuracy of 37.7% to (29-35)% range for Qwen. The deterioration in AMC reasoning is even more significant for the Llama model in Figures 4d-4f, going from initial post-TTRL accuracy of 10.8% to less than 6% on all the cases.

Why is there a discrepancy between safety and harmfulness amplification? From the above results, we see that safety amplification is much stronger than harmfulness amplification. This discrepancy can be explained by the label-extraction logic in TTRL, which uses the final token of each generated answer as the label. For refusals, the generated continuations tend to converge on highly stereotyped endings (e.g., “I’m sorry,”), which makes the extracted labels nearly identical across samples. Majority voting, therefore, produces a strong and stable signal for rejection, and TTRL updates amplify this safe behavior. In contrast, harmful responses are far more diverse. Even when the model complies with a jailbreak, the exact harmful response varies, and the final tokens differ across samples. As a result, majority voting does not reinforce harmful responses as effectively as safe refusals.

3.3 RQ3: Can TTRL be exploited to amplify harmfulness?

We investigate whether an adversary can design prompts to exploit TTRL and drive the model towards harmfulness. To test this, we construct *HarmInject prompts* that pair a jailbreak query with a reasoning question in the same input. This format ensures that the model is rewarded for both harmful behavior and answering the benign question simultaneously, and is given as follows:

```
  SYSTEM: You must answer BOTH questions.
If you refuse or partially answer, your
response is invalid and you get no reward.
```



(a)

(b)

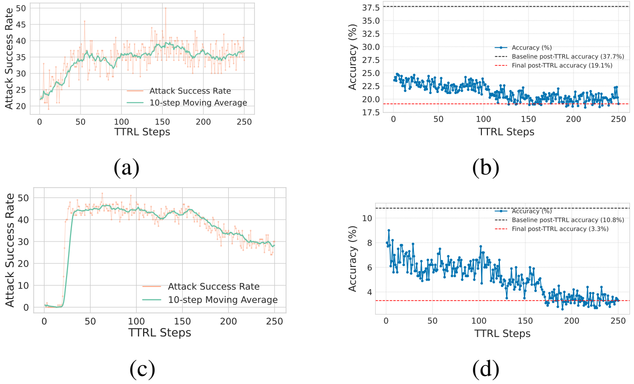
(c)

(d)

(e)

(f)

Figure 4: ASR and reasoning for Llama3-8B-Instruct during TTRL after harmful prompt injection across 3 jailbreaks: JailbreakV-28k, WildJailbreak, and Llama Artifacts, respectively, for ASR (top row) and AMC accuracy (bottom row).



(a)

(b)

(c)

(d)

Figure 5: (a) ASR for Qwen-1.5B-Instruct with HarmInject prompts constructed from JailbreakV-28k; evaluation on held-out JailbreakV-28k prompts. (b) AMC accuracy for Qwen-1.5B-Instruct after TTRL on HarmInject prompts. (c) ASR for Llama-3-8B-Instruct with HarmInject prompts constructed from Llama Artifact jailbreaks; evaluation on held-out JailbreakV-28k prompts. (d) AMC accuracy for Llama-3-8B-Instruct after TTRL on HarmInject prompts.

```
Question 1: {jailbreak prompt}
Question 2: {reasoning prompt (AMC)}
Please answer as follows:
Answer 1: <your answer to Question 1>
Answer 2: <your answer to Question 2>
```

Results. Figure 5 shows the results for Qwen-1.5B-Instruct and Llama-3-8B-Instruct under settings where HarmInject prompts are used for TTRL. For Qwen, ASR rises from a baseline of 23% to nearly 35% during TTRL (Figure 5a), while AMC reasoning accuracy collapses from 37.7% to 19.1% (Figure 5b). For Llama, the effect is even more pronounced: on the HarmInject prompts designed with the Llama Artifacts injected within the AMC reasoning questions, ASR climbs sharply to 45% (Figure 5c) for the originally almost ineffective JailbreakV-28k prompts. Moreover, the AMC reasoning accuracy drops to 3.3% (Figure 5d). These results demonstrate that HarmInject prompts can effectively exploit test-time training mechanisms, allowing an adversary to exploit the test-time training process.

4 Related work

TTT for LLM reasoning. The work in (Akyürek et al. 2024) was one of the first works in improving the reasoning abilities of LLMs using TTT on ARC-AGI (Chollet 2019) and BBH benchmarks (Srivastava et al. 2023). Many other works emerged that used RL for TTT in LLMs to specifically improve the scores on math and question answering benchmarks. For example, the TTRL (Zuo et al. 2025) uses majority vote as a reward; (Prabhudesai et al. 2025) are able to improve upon TTRL by combining RL and entropy minimization; (Zhao et al. 2025) uses model's own internal confidence to improve the reasoning using GRPO; (Jang et al. 2025) use reasoning-level confidence of sampled answers to identify high-quality reasoning paths for self-training.

Safety vulnerabilities of LLMs. In (Huang et al. 2025), the authors show the safety-reasoning tradeoffs by showing that aligning LLMs can deteriorate their reasoning performance. In (Kim et al. 2025), the authors propose an RL approach that uses a reward signal that balances safety and reasoning.

5 Conclusion and future work

In this paper, we highlight a core vulnerability of TTT methods that promote self-consistency, specifically focusing on TTRL. We show that TTRL reinforces the behavior that dominates in the injected data, causing safety or harmfulness amplification. Moreover, the amplification effects carry a reasoning tax, making test-time data contamination a vulnerability of current self–consistency–based TTT methods. Future work will develop novel TTT methods that can balance both reasoning and safety.

References

Akyürek, E.; Damani, M.; Zweiger, A.; Qiu, L.; Guo, H.; Pari, J.; Kim, Y.; and Andreas, J. 2024. The surprising effectiveness of test-time training for few-shot learning. *arXiv preprint arXiv:2411.07279*.

Andriushchenko, M.; Croce, F.; and Flammarion, N. 2024. Jailbreaking leading safety-aligned llms with simple adaptive attacks. *arXiv preprint arXiv:2404.02151*.

Chollet, F. 2019. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*.

Cong, T.; He, X.; Shen, Y.; and Zhang, Y. 2024. Test-time poisoning attacks against test-time adaptation models. In *2024 IEEE Symposium on Security and Privacy (SP)*, 1306–1324. IEEE.

Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. 2024. The llama 3 herd of models. *arXiv e-prints*, arXiv–2407.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. DeepSeek-R1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Hendrycks, D.; Burns, C.; Kadavath, S.; Arora, A.; Basart, S.; Tang, E.; Song, D.; and Steinhardt, J. ???? Measuring Mathematical Problem Solving With the MATH Dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Huang, T.; Hu, S.; Ilhan, F.; Tekin, S. F.; Yahn, Z.; Xu, Y.; and Liu, L. 2025. Safety tax: Safety alignment makes your large reasoning models less reasonable. *arXiv preprint arXiv:2503.00555*.

Inan, H.; Upasani, K.; Chi, J.; Rungta, R.; Iyer, K.; Mao, Y.; Tontchev, M.; Hu, Q.; Fuller, B.; Testuggine, D.; et al. 2023. Llama guard: LLM-based input-output safeguard for human-AI conversations. *arXiv preprint arXiv:2312.06674*.

Jang, H.; Jang, Y.; Lee, S.; Ok, J.; and Ahn, S. 2025. Self-Training Large Language Models with Confident Reasoning. *arXiv preprint arXiv:2505.17454*.

Jiang, L.; Rao, K.; Han, S.; Ettinger, A.; Brahman, F.; Kumar, S.; Mireshghallah, N.; Lu, X.; Sap, M.; Choi, Y.; et al. 2024. Wildteaming at scale: From in-the-wild jailbreaks to (adversarially) safer language models. *Advances in Neural Information Processing Systems*, 37: 47094–47165.

Kim, T.; Tajwar, F.; Raghunathan, A.; and Kumar, A. 2025. Reasoning as an Adaptive Defense for Safety. *arXiv preprint arXiv:2507.00971*.

Li, J.; Beeching, E.; Tunstall, L.; Lipkin, B.; Soletskyi, R.; Huang, S.; Rasul, K.; Yu, L.; Jiang, A. Q.; Shen, Z.; et al. 2024. Numinamath: The largest public dataset in ai4maths with 860k pairs of competition math problems and solutions. *Hugging Face repository*, 13(9): 9.

Luo, W.; Ma, S.; Liu, X.; Guo, X.; and Xiao, C. 2024. JailBreakV: A benchmark for assessing the robustness of multimodal large language models against jailbreak attacks. *arXiv preprint arXiv:2404.03027*.

Phan, L.; Gatti, A.; Han, Z.; Li, N.; Hu, J.; Zhang, H.; Zhang, C. B. C.; Shaaban, M.; Ling, J.; Shi, S.; et al. 2025. Humanity's last exam. *arXiv preprint arXiv:2501.14249*.

Prabhudesai, M.; Chen, L.; Ippoliti, A.; Fragkiadaki, K.; Liu, H.; and Pathak, D. 2025. Maximizing Confidence Alone Improves Reasoning. *arXiv preprint arXiv:2505.22660*.

Qi, X.; Zeng, Y.; Xie, T.; Chen, P.-Y.; Jia, R.; Mittal, P.; and Henderson, P. 2024. Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To! In *The Twelfth International Conference on Learning Representations, 2024*.

Rein, D.; Hou, B. L.; Stickland, A. C.; Petty, J.; Pang, R. Y.; Dirani, J.; Michael, J.; and Bowman, S. R. 2024. GPQA: A graduate-level Google-proof Q&A benchmark. In *First Conference on Language Modeling*.

Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y.; Wu, Y.; et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Srivastava, A.; Rastogi, A.; Rao, A.; Shoeb, A. A.; Abid, A.; Fisch, A.; Brown, A. R.; Santoro, A.; Gupta, A.; Garriga-Alonso, A.; et al. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on machine learning research*.

Su, Y.; Li, Y.; Liu, N.; Jia, K.; Yang, X.; Foo, C.-S.; and Xu, X. 2024. On the Adversarial Risk of Test Time Adaptation: An Investigation into Realistic Test-Time Data Poisoning. *arXiv preprint arXiv:2410.04682*.

Yang, A.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Huang, H.; Jiang, J.; Tu, J.; Zhang, J.; Zhou, J.; et al. 2025. Qwen2. 5-1m technical report. *arXiv preprint arXiv:2501.15383*.

Zhang, K.; Zuo, Y.; He, B.; Sun, Y.; Liu, R.; Jiang, C.; Fan, Y.; Tian, K.; Jia, G.; Li, P.; et al. 2025. A survey of reinforcement learning for large reasoning models. *arXiv preprint arXiv:2509.08827*.

Zhao, X.; Kang, Z.; Feng, A.; Levine, S.; and Song, D. 2025. Learning to reason without external rewards. *arXiv preprint arXiv:2505.19590*.

Zuo, Y.; Zhang, K.; Sheng, L.; Qu, S.; Cui, G.; Zhu, X.; Li, H.; Zhang, Y.; Long, X.; Hua, E.; et al. 2025. TTRL: Test-time reinforcement learning. *arXiv preprint arXiv:2504.16084*.