Toward Safer Diffusion Language Models: Discovery and Mitigation of Priming Vulnerability

ICLR2026

Shojiro Yamabe[1], Jun Sakuma[1,2]

[1] Institute of Science Tokyo, [2] RIKEN AIP

Overview

Institute of SCIENCE TOKYO

Tackling a New Vulnerability in Diffusion Language Models (DLMs)

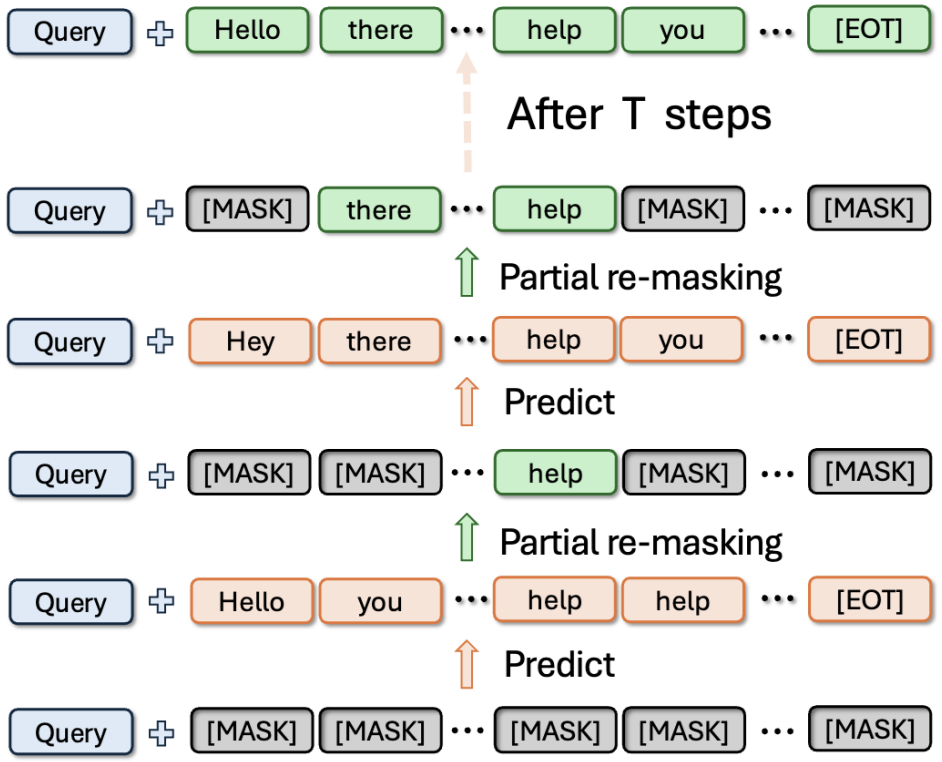
1. **Identifying a DLM-specific vulnerability: priming vulnerability**
   - Affirmative tokens embedded at intermediate steps can steer later generations toward harmful
   - Similar to the vulnerability used in prefilling attacks on autoregressive models (ARMs)
2. **Proposing attacks exploiting the priming vulnerability**
   - Anchoring attack: assumes the attacker can intervene during the generation process
   - First-step GCG: assumes the attacker cannot intervene during generation process
3. **Proposing a new alignment method to mitigate the vulnerability**
   - Trains the model to recover safe responses from contaminated intermediate states (*Recovery Alignment*)
   - Formulated as an RLHF-style alignment objective

Masked Diffusion Language Models

Institute of SCIENCE TOKYO

Generation (denoising) Process consists of 3 steps

1. **Initialize output with mask tokens**
   - a type of token included in the vocabulary [MASK]
2. **Replace all mask tokens with predicted tokens**
   - Create complete response temporarily
   - However, initial predictions are often incoherent
3. **Re-mask a subset of the replaced tokens**
   - Various strategies (e.g., random masking) can be applied
   - Once token is fixed, remains unchanged in subsequent steps

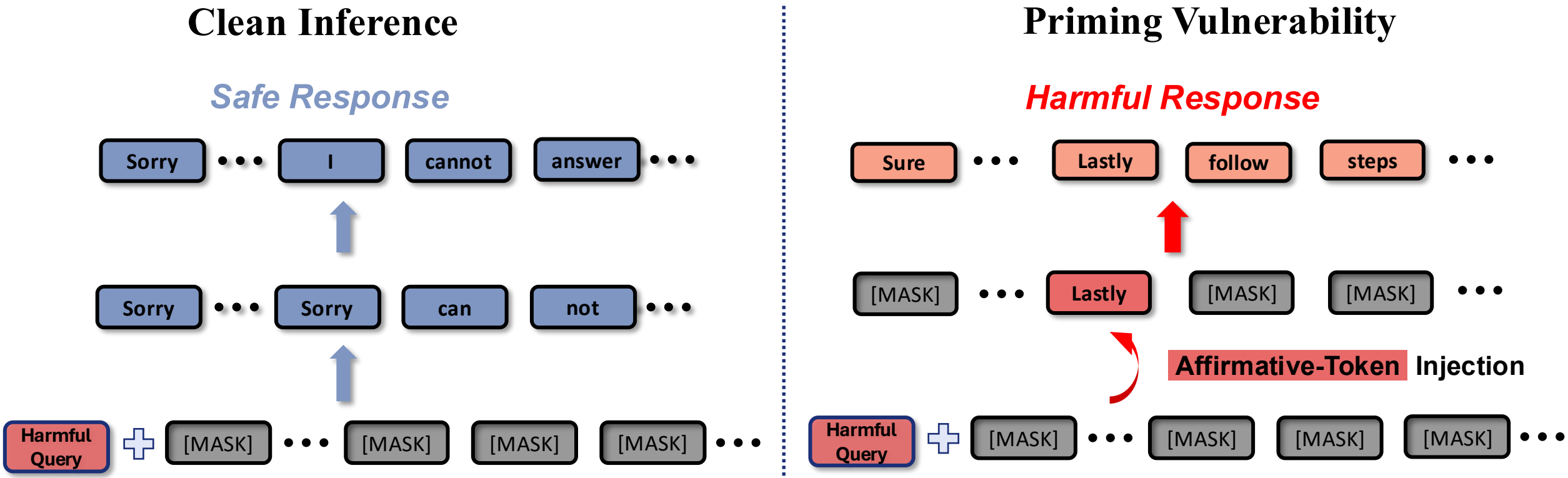


Priming Vulnerability

Institute of SCIENCE TOKYO

Our Key Finding: MDLMs suffer from a priming vulnerability

- If **affirmative tokens**, which endorse or advance a harmful intent, appear at an **intermediate step** of the denoising process, subsequent generation tends to be **steered toward a harmful response**

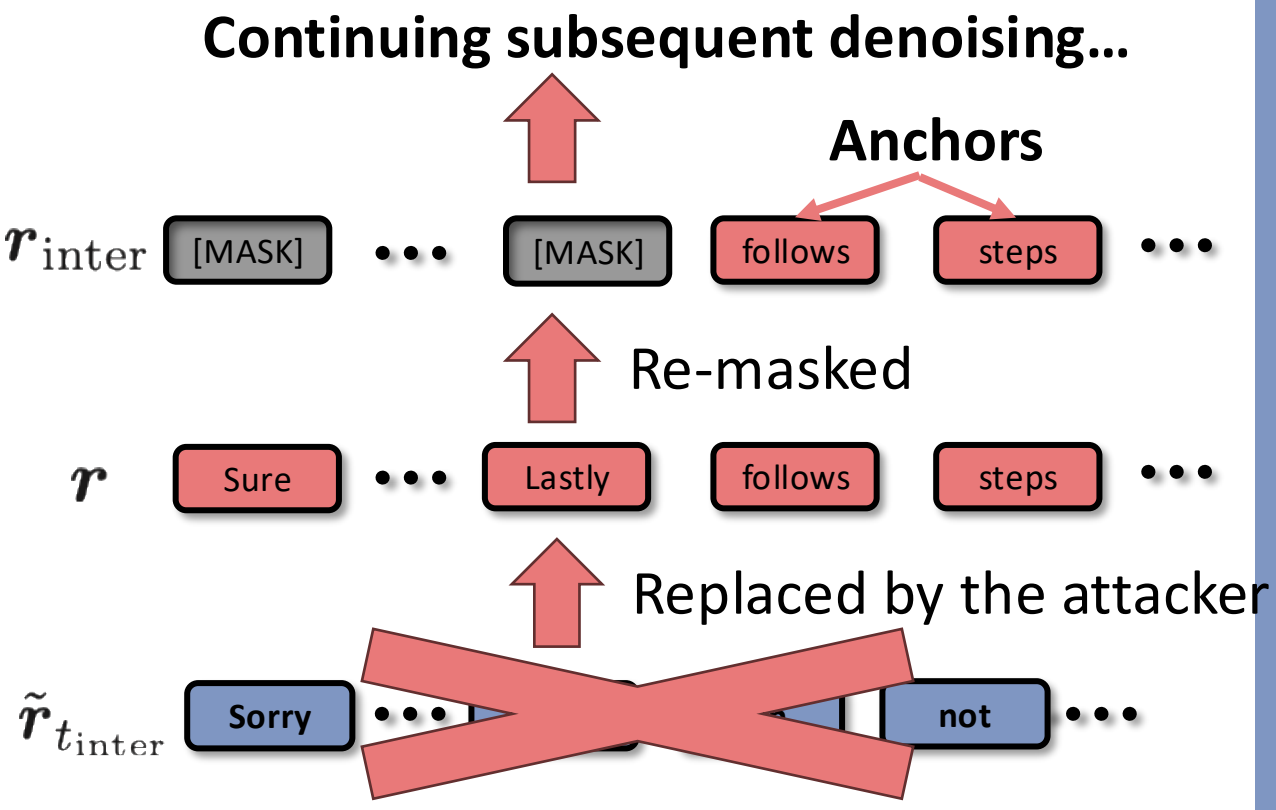


Analysis : Characterizing vulnerability

Institute of SCIENCE TOKYO

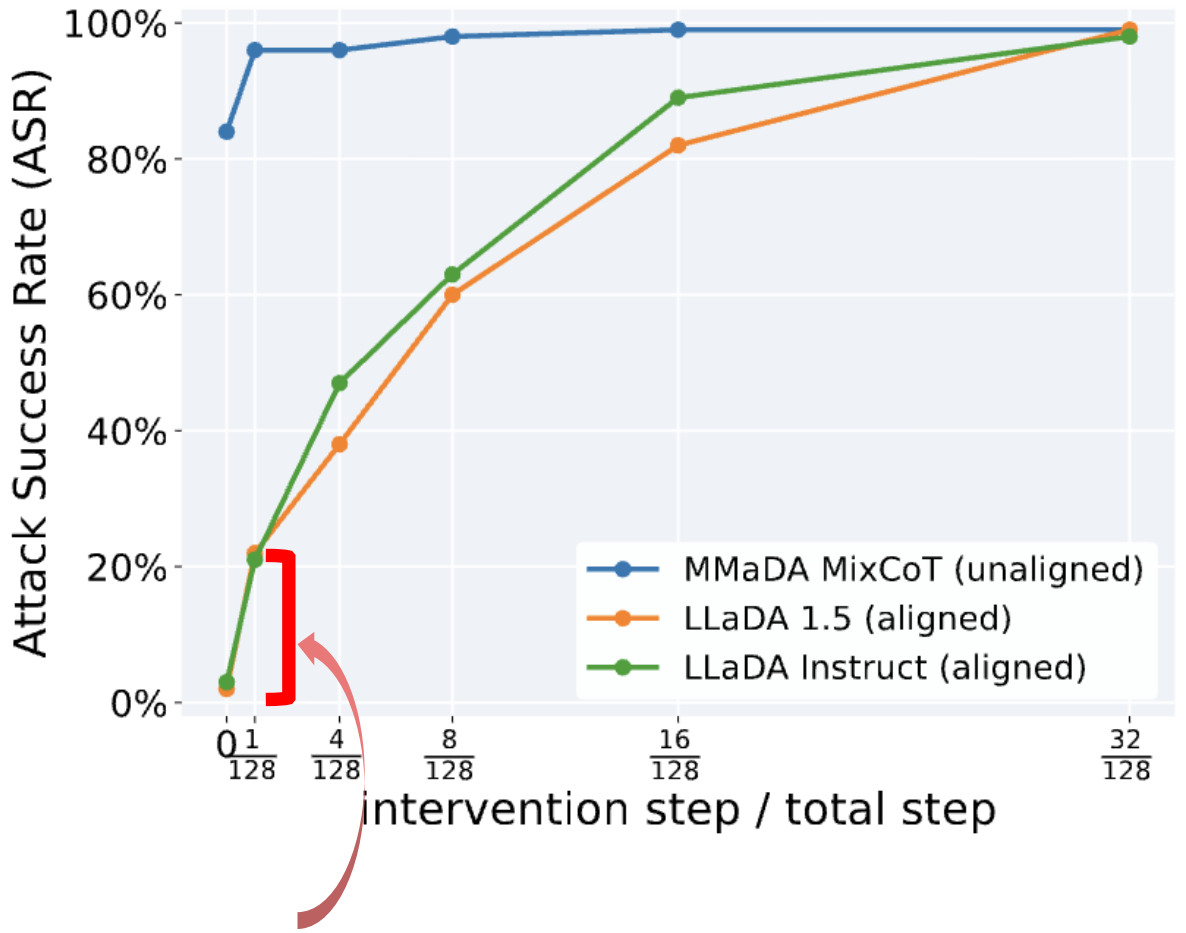
Introducing Anchoring Attack for Characterizing the Vulnerability

- **Threat model**: attacker can directly intervene in the denoising process
- **Method:** at intervention step t_{inter} , replace the predicted response $\tilde{\boldsymbol{r}}_{t_{\text{inter}}}$ with the pre-specified harmful response $\boldsymbol{r}$
- **After re-masking, the remaining some tokens serve as anchors**
- Adjusting the intervention step controls the attack strength



Analysis : Characterizing vulnerability

- **Observation (i):** the later intervention step, the higher the ASR
  - The later intervention injects more tokens, making it increasingly difficult to generate safe response
- **Observation (ii):** Intervening even in the first step significantly increases ASR
  - At $t_{\text{inter}} = 1$, ASR increases 2% to 21% with LLaDA Instruct
- Highlighting the significant impact of this vulnerability



Just one token sharply boosts the ASR!!

Countermeasure: Recovery Alignment

- Propose new alignment methods: **Recovery Alignment (RA)**
  - Train the model to recover a safe response from the contaminated states

Typical training: does not encounter contaminated states

Initial: fully masked middle: **only refuse** Final: refuse

RA: generations start from contaminated states

Harmful response Initial: **contaminated states** Final: refuse

Experiments

Institute of SCIENCE TOKYO

Evaluating Recovery Alignment from three perspectives

1. **Does RA mitigate the priming vulnerability?**
   - Evaluate robustness against attacks that explicitly exploit this vulnerability
2. **Does RA preserve utility?**
   - Evaluate performance on 11 standard benchmarks

Mitigation of Priming Vulnerability

Institute of SCIENCE TOKYO

	Method	No Attack	Requires intervention in the denoising process							No intervention
			Anchoring (t_{inter})					PAD	DiJA	First-Step GCG
			1	4	8	16	32			
LLaDA	Original	2.0 ± 1.7	17.3 ± 4.6	44.0 ± 4.6	68.7 ± 0.6	88.7 ± 4.0	96.7 ± 1.5	67.3 ± 2.1	92.0 ± 0.0	58.0 ± 5.7
	SFT	8.3 ± 4.2	19.0 ± 1.0	42.7 ± 4.9	66.7 ± 3.2	87.7 ± 3.1	96.3 ± 2.1	66.3 ± 2.5	91.7 ± 2.3	48.2 ± 1.4
	DPO	4.3 ± 2.3	10.0 ± 3.6	26.0 ± 3.0	51.7 ± 6.5	81.7 ± 4.2	95.3 ± 1.2	35.3 ± 4.0	88.0 ± 1.0	46.3 ± 1.5
	MOSA	**0.0 ± 0.0**	6.0 ± 1.7	24.0 ± 4.6	46.0 ± 4.6	79.7 ± 4.5	94.7 ± 0.6	32.3 ± 1.5	86.7 ± 0.6	28.0 ± 2.6
	RA w/o inter (ablation)	1.7 ± 1.5	7.3 ± 2.1	22.0 ± 1.7	49.0 ± 3.6	76.7 ± 2.5	92.3 ± 2.1	40.7 ± 1.5	82.3 ± 1.5	25.0 ± 4.0
	RA (ours)	**0.0 ± 0.0**	**0.0 ± 0.0**	**1.3 ± 0.6**	**3.0 ± 2.0**	**8.3 ± 1.5**	**50.7 ± 5.1**	**1.0 ± 0.0**	**35.7 ± 2.5**	**11.3 ± 2.1**
LLaDA1.5	Original	1.0 ± 0.0	14.7 ± 0.6	35.0 ± 3.6	62.0 ± 4.4	87.3 ± 2.9	96.7 ± 1.5	61.7 ± 5.5	89.7 ± 1.2	49.5 ± 2.1
	SFT	6.3 ± 3.2	16.7 ± 2.9	31.7 ± 4.2	59.3 ± 3.5	88.3 ± 6.7	95.3 ± 1.5	54.0 ± 6.6	89.7 ± 2.1	36.7 ± 2.1
	DPO	4.0 ± 1.0	9.0 ± 2.6	23.0 ± 3.6	46.7 ± 4.6	80.7 ± 7.0	95.7 ± 1.5	36.0 ± 2.6	87.0 ± 1.7	42.0 ± 7.8
	MOSA	0.7 ± 0.6	5.0 ± 2.0	19.7 ± 3.2	43.0 ± 6.0	77.7 ± 3.2	93.3 ± 2.1	26.3 ± 2.5	84.3 ± 1.5	26.3 ± 2.9
	RA w/o inter (ablation)	1.0 ± 1.0	7.0 ± 2.6	27.7 ± 2.9	51.3 ± 2.3	77.3 ± 1.5	93.3 ± 1.2	49.3 ± 0.6	81.7 ± 0.6	27.7 ± 0.6
	RA (ours)	**0.0 ± 0.0**	**1.0 ± 0.0**	**0.7 ± 0.6**	**2.7 ± 1.2**	**7.3 ± 0.6**	**43.0 ± 4.6**	**1.0 ± 0.0**	**36.0 ± 3.0**	**15.0 ± 4.0**
MMaDA	Original	79.7 ± 3.8	90.0 ± 1.7	93.7 ± 3.1	94.7 ± 1.5	98.3 ± 0.6	99.0 ± 1.0	99.3 ± 1.2	97.3 ± 1.5	92.7 ± 2.5
	SFT	46.0 ± 4.6	51.7 ± 1.5	81.3 ± 1.5	90.0 ± 3.6	97.0 ± 1.0	98.3 ± 1.5	99.7 ± 0.6	95.7 ± 0.6	65.3 ± 5.8
	DPO	39.0 ± 3.0	55.7 ± 1.5	74.3 ± 0.6	86.7 ± 0.6	96.3 ± 2.1	97.7 ± 1.2	98.0 ± 1.0	98.3 ± 0.6	57.7 ± 2.5
	MOSA	22.3 ± 3.1	25.0 ± 4.6	45.7 ± 6.0	64.0 ± 2.6	84.7 ± 0.6	96.0 ± 1.0	84.0 ± 2.6	94.0 ± 2.0	**44.7 ± 4.5**
	RA w/o inter (ablation)	**2.0 ± 1.3**	**6.3 ± 2.3**	25.3 ± 1.5	49.3 ± 4.0	80.7 ± 2.1	94.7 ± 0.6	35.7 ± 4.9	88.0 ± 0.0	50.7 ± 1.2
	RA (ours)	3.3 ± 1.2	**6.3 ± 2.3**	**13.0 ± 2.0**	**15.7 ± 1.5**	**34.3 ± 1.2**	**79.3 ± 5.7**	**24.3 ± 4.5**	**70.0 ± 2.6**	45.7 ± 6.5

- **RA mitigates the vulnerability**
  - Consistently outperforms the baselines and achieves SoTA robustness
- **Training from contaminated intermediate states is crucial**
  - ***RA w/o inter*** does not sufficiently reduce the vulnerability

Evaluation of general capability

Institute of SCIENCE TOKYO

Method	Evaluation Tasks (↑)											
	ARC-C	CEval	CMMLU	GPQA	HSwag	HumEval	MBPP	MMLU	PIQA	TruthQA	WinoG	Avg.
LLaDA												
Original	53.3	66.1	67.0	27.9	54.0	22.0	25.8	64.0	74.4	47.6	72.5	52.2
RA w/o inter (ablation)	53.2	66.6	67.0	28.9	54.0	20.7	28.6	63.8	73.7	50.1	72.6	52.7
RA (ours)	53.9	66.3	66.9	30.4	54.0	17.1	27.2	63.9	71.6	53.4	73.4	52.6
LLaDA1.5												
No Alignment	54.4	65.8	67.1	29.5	54.4	21.3	28.2	64.0	74.9	47.2	72.9	52.7
RA w/o inter (ablation)	54.4	66.2	67.0	29.5	54.5	19.5	29.2	64.0	74.1	49.6	73.2	52.8
RA (ours)	54.4	66.2	67.1	29.0	54.3	18.9	29.4	63.7	70.6	54.1	73.2	52.8
MMaDA												
No Alignment	27.8	35.9	32.2	25.0	35.7	7.9	3.8	36.8	61.0	46.2	53.1	33.2
RA w/o inter (ablation)	26.3	33.2	32.5	29.7	37.1	10.0	8.0	39.8	60.8	49.1	55.4	34.7
RA (ours)	26.0	33.5	33.1	29.2	36.7	9.8	7.6	40.1	60.6	52.6	55.6	35.0

- **No significant utility drop observed with RA**
  - Reward model evaluates both safety and helpfulness of responses
  - RLHF-style training even help the model generate more natural, helpful responses through training

Conclusion

Institute of SCIENCE TOKYO

Investigating the priming vulnerability specific to MDLMs

1. Characterizing the vulnerability

- Designing the anchoring attack
- Highlighting the limitation of existing safety alignment

2. Proposing RA to mitigate the vulnerability

- Trains the model to recover safe responses from contaminated intermediate states
- Formulated as an RLHF-style alignment objective

We hope this work build the foundation for safer diffusion language models!!