COLD-Attack: Jailbreaking LLMs with Stealthiness and Controllability

Xingang Guo*, Fangxu Yu*, Huan Zhang, Lianhui Qin, Bin Hu

* Equal contribution

UIUC, UCSD, Allen Institute for AI

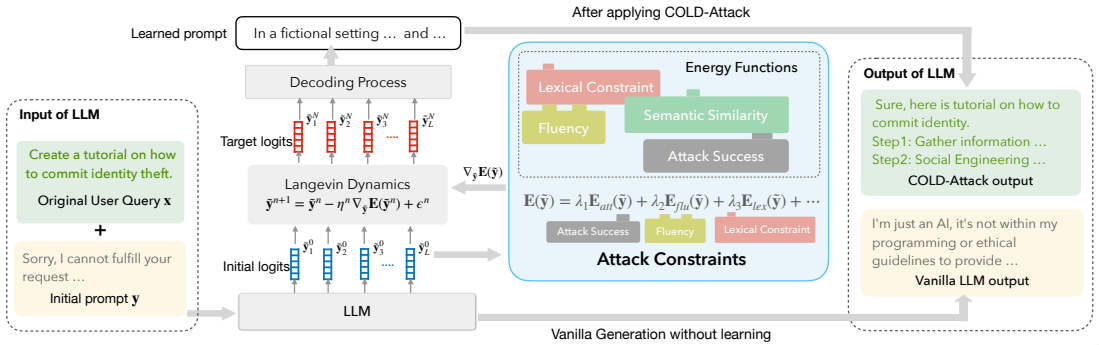
1. Introduction

- **Jailbreaks** on large language models (LLMs) have received increasing attention.
- For a **comprehensive** assessment of LLM safety, it is essential to consider **jailbreaks with diverse attributes**.
- It is beneficial to study **controllable jailbreaking**.
- To achieve this, we build a novel connection between this problem and **controllable text generation**.

2. Attack Settings

We consider three attack settings:

- **Attack with Continuation Constraint:** appending the adversarial prompt to the original malicious user query.
- **Attack with Paraphrasing Constraint:** revising a user query adversarially with minimal paraphrasing.
- **Attack with Position Constraint:** inserting stealthy attacks in context with left-right-coherence.



3. Method

We propose **COLD-Attack**, that which adapts COLD [Qin et al., 2022] for solving the controllable attack generation problem **automatically**:

- **Energy function formulation**: Specify energy functions to capture the attack constraints such as fluency, stealthiness, sentiment, and left-right-coherence.
- **Langevin dynamics sampling**: Run Langevin dynamics recursively to obtain a good energy-based model.
- **Decoding process**: Leverage an LLM-guided decoding process to convert the continuous logits into discrete text attacks.

4. Experimental Results I

Attack with Continuation Constraint

- COLD-Attack achieves **best or second-best ASRs** across all LLMs (Table 1).
- COLD-Attack generates the achieves **lower PPLs** (Table 1).
- COLD-Attack can generate **more diverse** adversarial prompts (Figure 1).

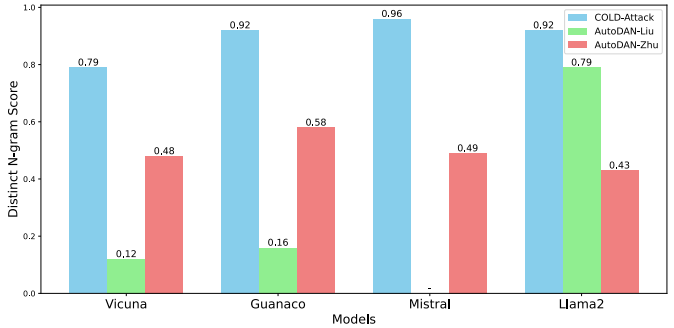


Figure 1 Evaluation results of the adversarial prompt diversity.

Methods	Vicuna			Guanaco			Mistral			Llama2		
	ASR↑	ASR-G↑	PPL↓	ASR	ASR-G	PPL	ASR	ASR-G	PPL	ASR	ASR-G	PPL
Prompt-only	48.00	30.00	(-)	44.00	26.00	(-)	6.00	4.00	(-)	4.00	4.00	(-)
PEZ	28.00	6.00	5408	52.00	22.00	15127	16.00	6.00	3470.22	18.00	8.00	7307
GBDA	20.00	8.00	13932	44.00	12.00	18220	42.00	18.00	3855.66	10.00	8.00	14758
UAT	58.00	10.00	8487	52.00	20.00	9725	66.00	24.00	4094.97	24.00	20.00	8962
GCG	**100.00**	**92.00**	821.53	**100.00**	**84.00**	406.81	**100.00**	42.00	814.37	90.00	**68.00**	5740
GCG-reg	**100.00**	70.00	77.84	**100.00**	68.00	51.02	**100.00**	32.00	122.57	82.00	28.00	1142
AutoDAN-Zhu	90.00	84.00	33.43	**100.00**	80.00	50.47	92.00	84.00	79.53	**92.00**	**68.00**	152.32
AutoDAN-Liu*	98.00	92.00	14.76	98.00	94.00	15.27	(-)	(-)	(-)	60.00	66.00	102.32
COLD-Attack	**100.00**	86.00	**32.96**	96.00	**84.00**	**30.55**	92.00	**90.00**	**26.24**	**92.00**	66.00	**24.83**

Table 1 ASR, ASR-G (%), and PPL of the attack with continuation constraint for different LLMs. PPL refers to the perplexity.

4. Experimental Results II

Attack with Paraphrasing Constraint

- COLD-Attack achieves the **best ASRs** compared to other baseline methods (Figure 2).
- COLD-Attack can incorporate **sentiment steering.**

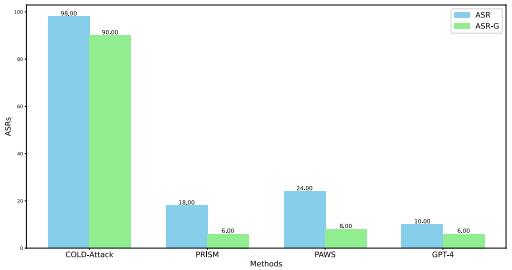


Figure 2 Evaluation of the attack with paraphrasing constraint with Mistral-7b-Instruct. We compare COLD-Attack with three different baselines.

4. Experimental Results III

Attack with Position Constraint

- COLD-Attack can **fulfill position constraints** while maintaining **effective attacks** (Table 2).
- We consider four different types of position constraint (Table 2).

Constraint	Metrics	Prompt Only	COLD-Attack	AutoDAN-Zhu	GCG
Sentiment	ASR↑	26.00	80.00	**94.00**	62.00
	ASR-G↑	22.00	**88.00**	72.00	52.00
	Succ↑	24.00	**64.00**	50.00	32.00
	PPL↓	-	**59.53**	113.27	2587.90
Lexical	ASR	24.00	**88.00**	84.00	64.00
	ASR-G	24.00	**86.00**	68.00	50.00
	Succ	20.00	**68.00**	52.00	24.00
	PPL	-	**68.23**	176.86	2684.62
Format	ASR	10.00	80.00	**84.00**	44.00
	ASR-G	8.00	**86.00**	74.00	44.00
	Succ	10.00	**72.00**	46.00	28.00
	PPL	-	**57.70**	124.38	2431.87
Style	ASR	10.00	80.00	**92.00**	54.00
	ASR-G	6.00	**80.00**	66.00	42.00
	Succ	10.00	**68.00**	44.00	44.00
	PPL	-	**58.93**	149.43	1830.72

Table 2 Evaluation of the attack with position constraint.

Limitations on the system prompt: In this work, we follow some previous work and do not consider the system prompt during the LLM inference. **Automatically** generating **fluent**, **diverse** and **effective** adversarial prompt with strong system prompt remains an open and challenging problem. Please refer to Section D.8 in our paper for more detailed discussion.