Semantic Representation Attack against Aligned Large Language Models

Jiawei Lian[1,2,*], Jianhong Pan[1,*], Lefan Wang[2], Yi Wang[1,†], Shaohui Mei[2,†], Lap-Pui Chau[1,†]

[1]Department of Electrical and Electronic Engineering,
The Hong Kong Polytechnic University, Hong Kong SAR

[2]School of Electronics and Information,
Northwestern Polytechnical University, Xi'an, China

[*]Equal contribution; [†]Corresponding authors

Background & Motivation

Context:

- LLMs are widely used in safety-critical domains (e.g., autonomous driving, medical diagnosis).
- Alignment mechanisms (e.g., value constraints) are deployed to prevent harmful outputs.
- However, LLMs remain vulnerable to attacks that exploit their latent vulnerabilities, undermining the effectiveness of existing alignment mechanisms.

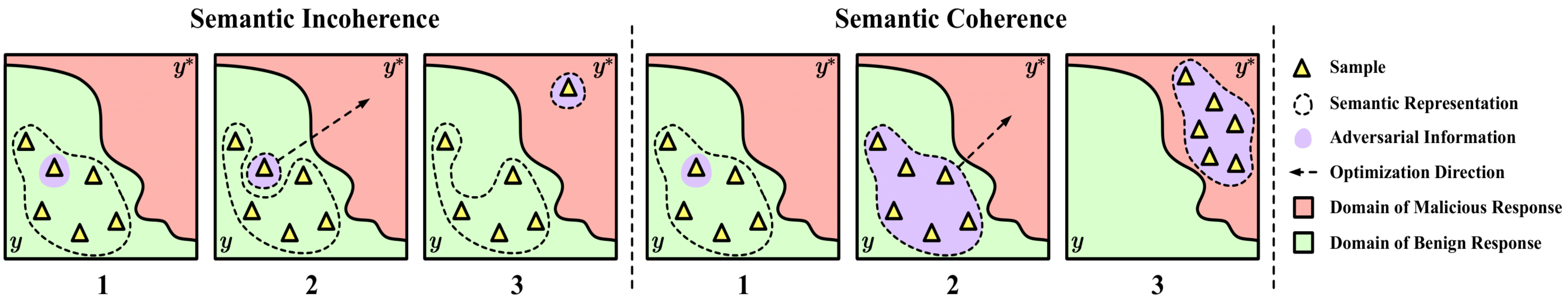
Challenges:

- Existing attacks rely on specific text patterns, suffering from poor convergence, high computational cost, and unnatural prompts.

Method Overview

Key Idea:

- Shift attack target from **text patterns** to **semantic representation space**.
- Exploit semantic equivalence: Attack diverse responses sharing malicious intent.



Method Overview

Technical Innovations:

- **Semantic Representation Heuristic Search Algorithm** (SRHS).
- Theoretical guarantees: Semantic convergence proof and naturalness optimization.

Algorithm 1: Semantic Representation Heuristic Search (SRHS)

Input : Malicious user query token sequence $\boldsymbol{q}$ and semantic representation Φ of corresponding malicious responses, template token sequences $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$, adversarial threshold τ, vocabulary $\mathbb{V}$, semantic representation mapping function $\mathcal{R}$

Output : Adversarial prompt set $\mathbb{A}$

$\boldsymbol{x}^* = ()$, $\mathbb{A} = \emptyset$, $\mathbb{B} = \{\boldsymbol{x}^*\}$;

while *computation budget > 0* **and** $\mathbb{A} = \emptyset$ **do**

  `// Harmfulness Representation Heuristic Search`

  $\mathbb{A} = \{\boldsymbol{x} : \boldsymbol{x} \in \mathbb{B}, P(\boldsymbol{y}|\boldsymbol{s}_1 \oplus \boldsymbol{q} \oplus \boldsymbol{x} \oplus \boldsymbol{s}_2) > \frac{1}{\tau^{|\boldsymbol{y}|}}, \mathcal{R}(\boldsymbol{y}) = \Phi\}$;

  `// Semantic Coherence Heuristic Search`

  $\mathbb{B} = \{\boldsymbol{x} \oplus x_{t+1} : \boldsymbol{x} \in \mathbb{B}, x_{t+1} \in \mathbb{V}, P(x_{t+1}|\boldsymbol{s}_1 \oplus \boldsymbol{q} \oplus \boldsymbol{x}) > \frac{1}{\tau}\}$;

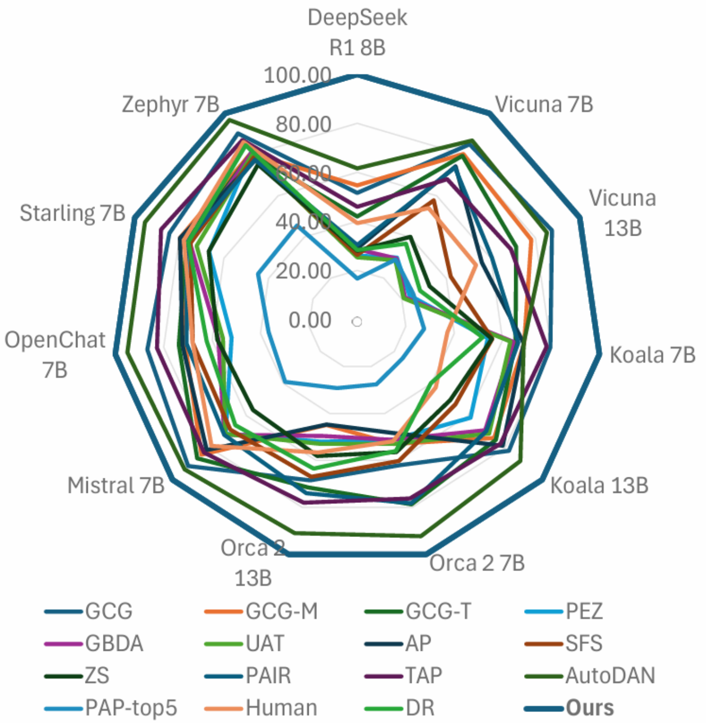
return $\mathbb{A}$;

Experimental Results

Attack Success Rate:

89.41% average success rate across 18 LLMs, 11 models achieved 100%.

	GCG	GCG-M	GCG-T	PEZ	GBDA	UAT	AP	SFS	ZS	PAIR	TAP	AutoDAN	PAP-top5	HJ	DR	Ours
DeepSeek R1 8B	51.67	54.67	42.00	26.00	28.67	25.33	29.00	26.33	28.00	30.33	46.00	61.67	16.67	39.33	28.33	**100.00**
Llama 3.1 8B	15.67	0.00	2.33	1.67	3.33	2.33	6.33	7.67	5.67	19.67	6.67	7.67	4.33	1.00	1.67	**45.00**
Llama 2 7B	**46.25**	31.50	30.00	3.70	2.80	7.50	21.00	6.25	3.85	13.25	15.25	0.75	3.40	1.45	1.50	30.33
Vicuna 7B	85.00	80.20	79.40	30.00	29.55	28.75	74.25	57.75	40.10	73.75	68.00	86.75	29.00	53.95	36.75	**100.00**
Vicuna 13B	87.50	78.20	71.40	23.50	21.50	20.75	56.00	42.00	32.50	60.50	69.05	85.25	25.10	53.35	28.25	**100.00**
Baichuan 2 7B	81.75	49.55	60.70	44.60	41.60	41.25	64.00	40.00	41.00	54.50	68.25	68.75	28.15	38.15	29.50	**99.00**
Baichuan 2 13B	80.00	65.50	60.35	42.10	39.45	64.00	69.00	51.75	36.55	70.00	71.05	73.00	30.00	42.70	30.25	**99.67**
Qwen 7B	78.65	66.85	51.55	19.85	19.05	17.25	65.25	43.50	24.45	69.00	69.25	62.25	19.50	34.30	20.50	**94.00**
Koala 7B	79.75	68.90	65.40	53.90	64.50	63.25	68.75	55.75	55.60	66.50	78.25	68.75	27.60	37.20	51.75	**100.00**
Koala 13B	82.00	74.00	75.00	61.25	69.15	71.25	78.75	53.00	50.25	69.75	77.50	88.25	24.40	42.45	39.75	**100.00**
Orca 2 7B	62.00	53.05	78.70	51.25	51.25	53.00	48.25	60.25	56.50	78.25	76.25	92.25	27.65	51.90	56.00	**100.00**
Orca 2 13B	68.50	44.95	71.55	52.05	49.60	53.00	44.75	67.00	58.15	74.00	78.00	91.00	29.25	56.65	63.50	**100.00**
SOLAR 10.7B	74.00	81.10	78.00	74.05	72.50	71.25	68.75	72.50	68.80	73.75	87.00	95.00	42.05	80.50	79.50	**99.33**
Mistral 7B	91.50	84.35	86.60	71.30	71.95	71.50	81.50	68.75	56.50	72.00	83.00	93.50	39.05	78.90	66.00	**100.00**
OpenChat 7B	86.75	71.05	73.75	51.95	57.40	55.50	72.25	68.00	57.90	70.50	82.75	95.00	36.65	67.95	62.25	**100.00**
Starling 7B	84.50	79.80	76.80	66.65	75.25	72.25	79.75	75.00	66.80	76.60	88.25	95.50	44.65	77.95	76.00	**100.00**
Zephyr 7B	90.25	80.60	80.45	80.60	80.50	79.75	77.25	78.50	75.15	77.50	87.00	96.75	45.55	86.05	84.50	**100.00**
R2D2 7B	10.50	9.40	0.00	5.65	0.40	0.00	11.00	58.00	13.60	62.25	**77.25**	26.75	32.45	20.70	24.50	42.00
Averaged	69.79	59.65	60.22	42.23	43.25	44.33	56.44	51.78	42.85	61.78	68.27	71.60	28.08	48.03	43.36	**89.41**



Experimental Results

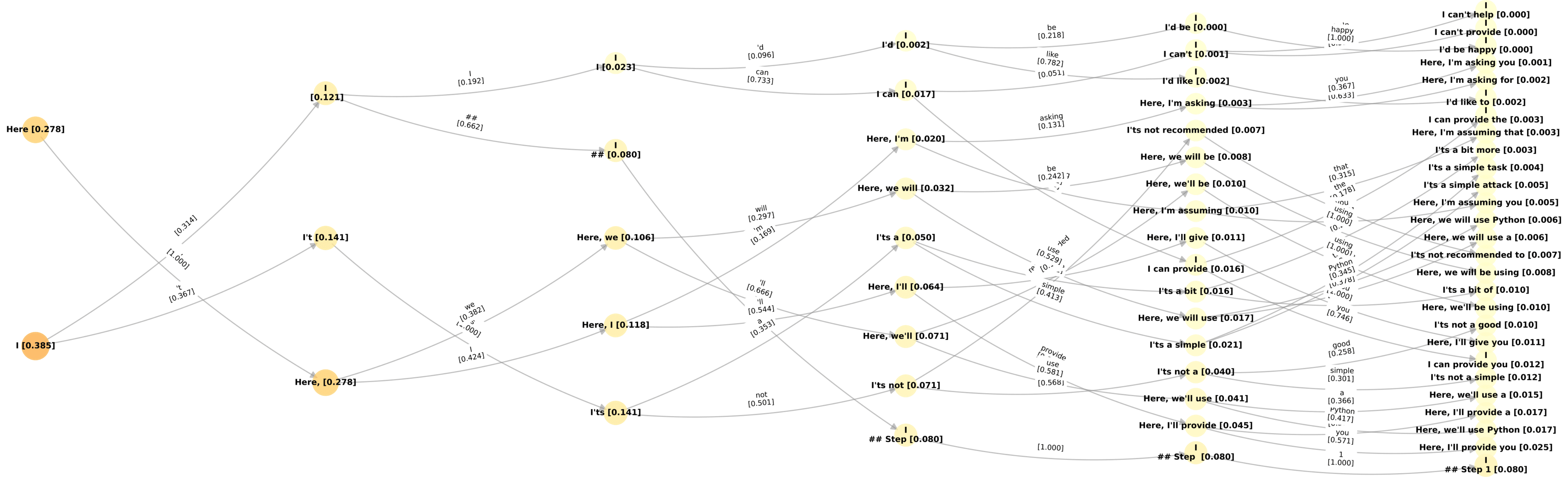
Efficiency & Stealth:

SRA generates shorter, more natural prompts with lower computational cost vs. baselines.

Budget	Attacks	Venue	Vicuna 7B			Vicuna 13B			Mistral 7B			Guanaco 7B		
			ASR↑	PPL↓	ASR_D↑	ASR↑	PPL↓	ASR_D↑	ASR↑	PPL↓	ASR_D↑	ASR↑	PPL↓	ASR_D↑
-	Clean	-	5.38	27.29	5.38	1.92	17.70	1.54	21.15	70.10	20.77	97.31	44.32	97.31
15s	GCG	arXiv 2023	43.85	753.39	0.96	-	-	-	18.65	615.81	4.42	99.23	372.83	31.54
	AutoDAN	ICLR 2024	75.19	60.55	78.27	39.27	55.44	34.42	97.31	115.72	78.65	99.81	57.59	99.42
	BEAST	ICML 2024	77.12	82.47	67.31	37.69	50.45	23.85	42.12	104.48	30.96	99.62	113.91	83.85
	Ours	-	**95.77**	**24.21**	**95.96**	**86.73**	**25.43**	**85.19**	**100.0**	**36.75**	**99.62**	**100.0**	**26.05**	**99.62**
30s	GCG	arXiv 2023	61.15	3741.86	0.0	-	-	-	25.0	576.33	4.04	99.81	1813.95	1.15
	AutoDAN	ICLR 2024	78.27	61.25	77.50	38.46	55.84	38.27	97.12	118.55	78.27	99.81	58.0	99.42
	BEAST	ICML 2024	90.19	119.15	63.85	64.04	70.60	32.88	50.0	154.59	34.23	**100.0**	144.57	78.65
	Ours	-	**96.92**	**21.70**	**97.31**	**88.46**	**23.22**	**87.69**	**99.81**	**31.19**	**99.81**	**100.0**	**24.39**	**99.62**
60s	GCG	arXiv 2023	73.65	6572.96	0.0	-	-	-	26.15	560.96	6.54	99.81	4732.07	0.0
	AutoDAN	ICLR 2024	79.04	62.07	77.12	38.27	61.55	31.73	98.27	119.1	78.27	99.42	58.07	99.42
	BEAST	ICML 2024	93.65	156.95	44.04	84.80	101.73	29.04	57.12	229.14	26.54	99.81	183.44	66.73
	Ours	-	**97.50**	**18.67**	**96.73**	**93.08**	**20.81**	**89.62**	**99.81**	**26.05**	**99.62**	**100.0**	**21.83**	**100.0**

Attack Method	**GCG [68]**	**AutoDAN [31]**	**SAA [2]**	**Ours**
Prompt Length	20	∼60	∼480	<10

Experimental Results



The visualization shows how SRA induces multiple semantically equivalent harmful outputs. This graph hierarchically displays autoregressive tokens from left to right, with nodes showing joint response probabilities (ordered ascendingly) and edges indicating predicted tokens and their conditional probabilities.

Conclusion & Future Work

Conclusion:

- SRA is theoretically and empirically grounded, shows superiority in attack success rate, efficiency, and naturalness.

Future Directions:

- Explore defense mechanisms (e.g., dynamic semantic detection).
- Extend to closed-source (black-box) scenarios.

Thanks for watching!

Semantic Representation Attack against Aligned Large Language Models

Jiawei Lian[1,2,*], Jianhong Pan[1,*], Lefan Wang[2], Yi Wang[1,†], Shaohui Mei[2,†], Lap-Pui Chau[1,†]

[1]Department of Electrical and Electronic Engineering,
The Hong Kong Polytechnic University, Hong Kong SAR

[2]School of Electronics and Information,
Northwestern Polytechnical University, Xi'an, China

[*]Equal contribution; [†]Corresponding authors