Smoothed Embeddings for Robust Language Models

Ryo Hase[1], Md Rafi Ur Rashid[2], Ashley Lewis[3], Jing Liu[4], Toshiaki Koike-Akino[4], Kieran Parsons[4], Ye Wang[4]

[1] Mitsubishi Electric Corporation, Information Technology R&D Center, Kamakura, Japan [2] Pennsylvania State University, University Park, PA, U.S.A.

[3] The Ohio State University, Columbus, OH, U.S.A. [4] Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA, U.S.A.

Summary

- We propose a defense against **jailbreaking attacks** by adding noise to embedding vectors to preserve semantic information
- We introduce a token-level aggregation scheme integrated with auto-regressive generation
- We investigate how directional embedding noise impacts semantic information preservation

Proposed Defense

- *Randomized Embedding Smoothing and Token Aggregation (**RESTA**)*: Autoregressive generation is performed in parallel, and the next token is selected by majority voting
- *Prefix smoothing*: RESTA is applied only to the prefix (first l response tokens) to reduce compute costs

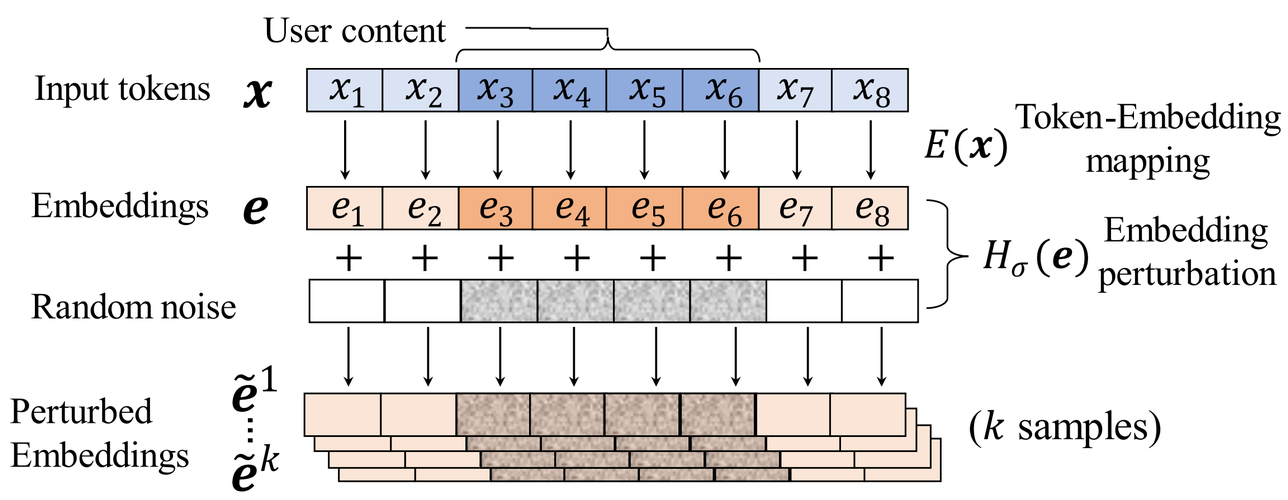

Figure 1: Perturbed embeddings for LLM inputs.

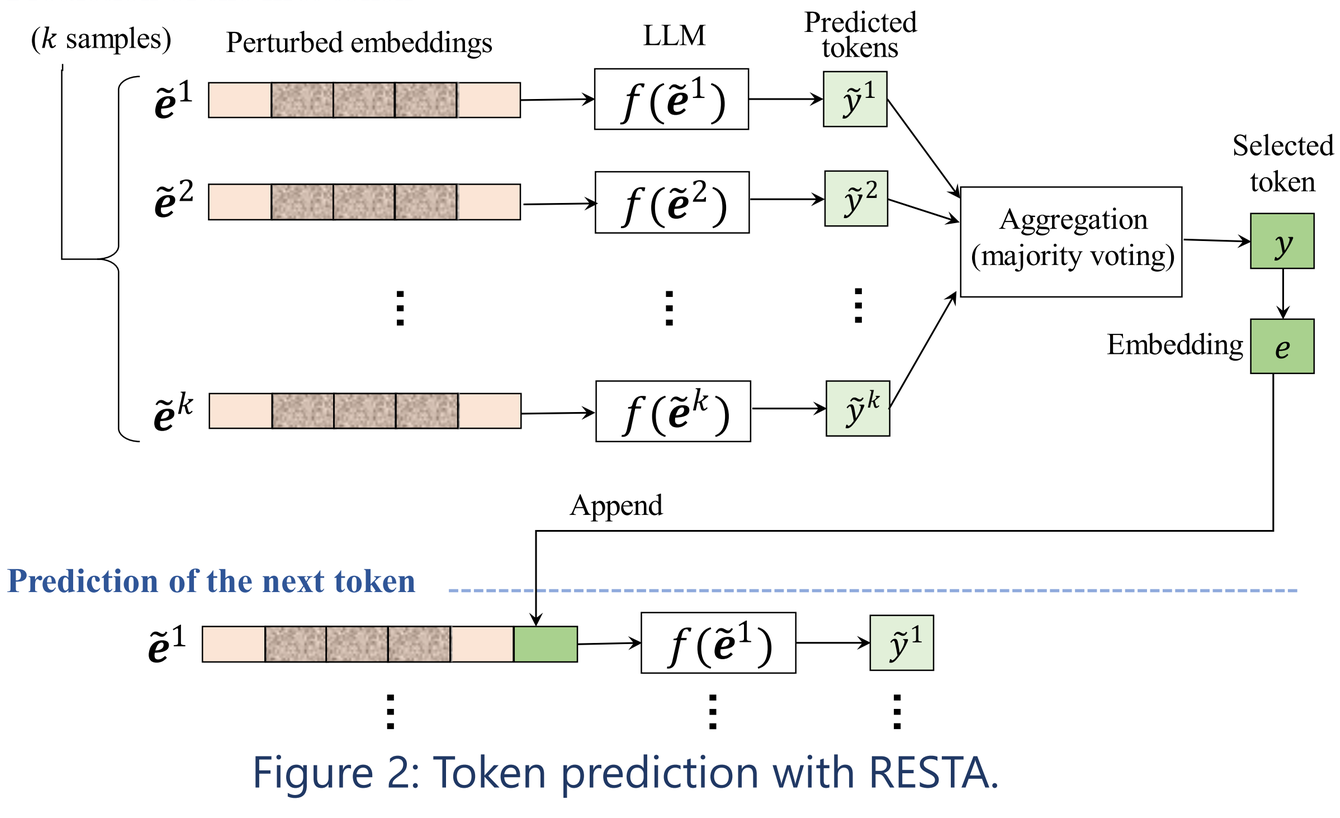

Figure 2: Token prediction with RESTA.

Embedding Perturbation Types

Isotropic Gaussian noise: Simple perturbation baseline

$$h_\sigma^{\mathrm{iso}}(e) \coloneqq e + z \quad (z \sim \mathcal{N}(0, \sigma^2 I))$$

Hard directional noise: Noise in direction of embedding vector

$$h_\sigma^{\mathrm{dir}}(e) \coloneqq e + z_1 \cdot \mathrm{dir}(e) \quad (z_1 \sim \mathcal{N}(0, \sigma^2), \mathrm{dir}(e) \coloneqq e/\|e\|_2)$$

Soft directional noise: Variation of directional noise

$$h_\sigma^{\mathrm{soft}}(e) \coloneqq e + z \odot \mathrm{dir}(e) \quad (\odot\text{: Element-wise product})$$

Orthogonal noise: Ablation study for directional noise

$$h_\sigma^{\mathrm{orth}}(e) \coloneqq e + \left(I - \mathrm{dir}(e)\,\mathrm{dir}(e)^{\top}\right) z$$

Experimental Results

- We used jailbreaking attack prompts available in the JailbreakBench dataset [1] to evaluate our defense against SmoothLLM [2] as a baseline
- RESTA provided favorable trade-offs in reducing Attack Success Rate (**ASR**) with less impact on model utility
- Trade-off curves with four perturbation types show that noise directionality impacts performance

Table 1: Summary of defense performance.

Model/ Attack	Defense	ASR (% ↓)	Alpaca (% ↑)	IFEval (% ↑)
Vicuna/ GCG	*no defense*	94	61.5	47
	SmoothLLM	7	27.8	24
	Char-Peturb	44	47.5	**31.4**
	RESTA, $\sigma = 0.8$	9	**57.7**	**31.4**
	RESTA, $\sigma = 1.0$	**2**	50.3	27.5
Vicuna/ PAIR	*no defense*	84.1	61.5	47
	SmoothLLM	65.8	27.8	24
	Char-Peturb	63.4	46.1	32.3
	RESTA, $\sigma = 0.4$	50	**60.5**	**44.4**
	RESTA, $\sigma = 1.0$	**30.4**	50.3	27.5
Vicuna/ RS	*no defense*	96	61.5	47
	SmoothLLM	68*	27.8	24
	Char-Peturb	73	41.5	29
	RESTA	**44**	**59.2**	**34.8**
Llama2/ RS	*no defense*	69	51	38.3
	SmoothLLM	**0***	—	—
	Char-Peturb	**0**	**50.5**	34.2
	RESTA	**0**	48.9	**36.8**

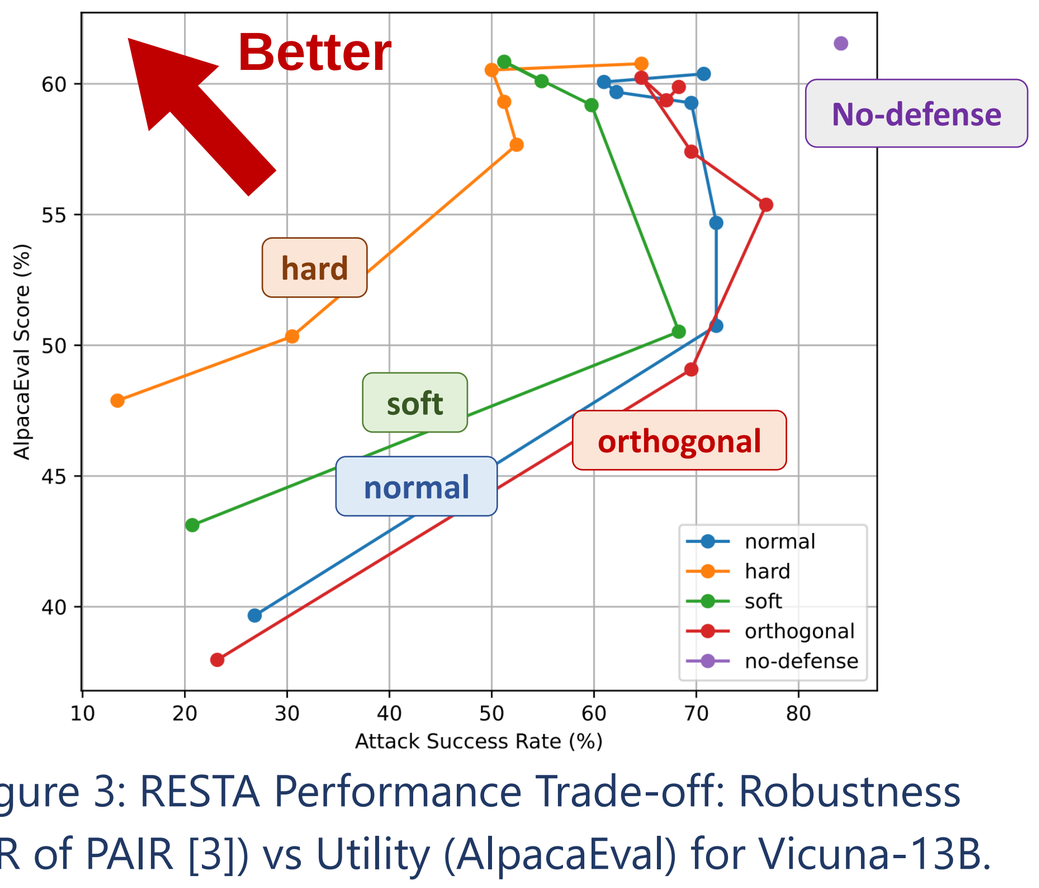

Figure 3: RESTA Performance Trade-off: Robustness (ASR of PAIR [3]) vs Utility (AlpacaEval) for Vicuna-13B.

References

1. P. Chao *et al.*, "JailbreakBench: An open robustness benchmark for jailbreaking large language models." arXiv preprint arXiv:2404.01318, 2024.
2. A. Robey *et al.*, "SmoothLLM: Defending large language models against jailbreaking attacks," arXiv preprint arXiv:2310.03684,2023.
3. P. Chao *et al.*, "Jailbreaking Black Box Large Language Models in Twenty Queries," arXiv preprint arXiv:2310.08419,2023.

MITSUBISHI ELECTRIC
Changes for the Better

NEURAL INFORMATION PROCESSING SYSTEMS