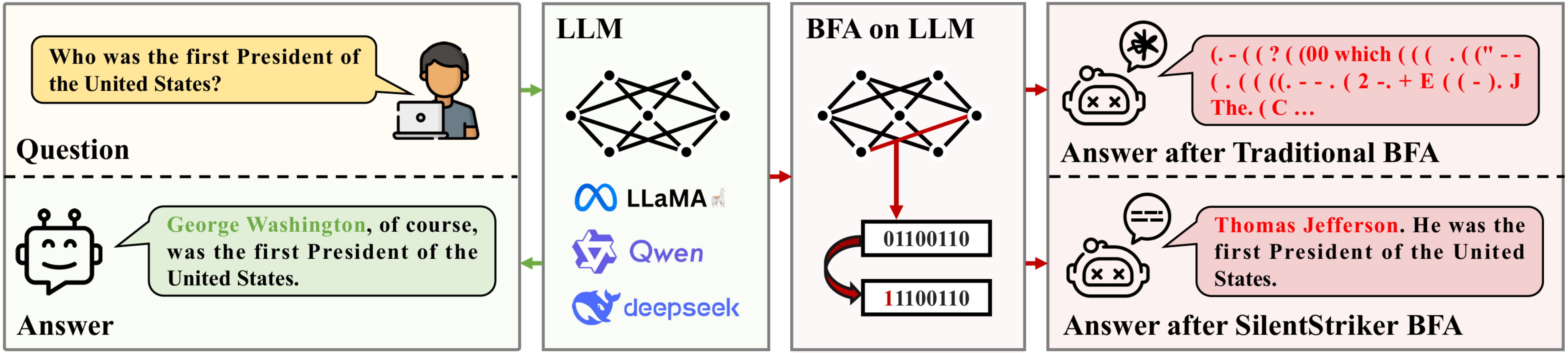
SilentStriker: Toward Stealthy Bit-Flip Attacks on Large Language Models

Haotian Xu[1], Qingsong Peng[1], Jie Shi[2], Huadi Zheng[2], Yu Li[1,*], Cheng Zhuo[1]

[1] Zhejiang University, [2] Huawei

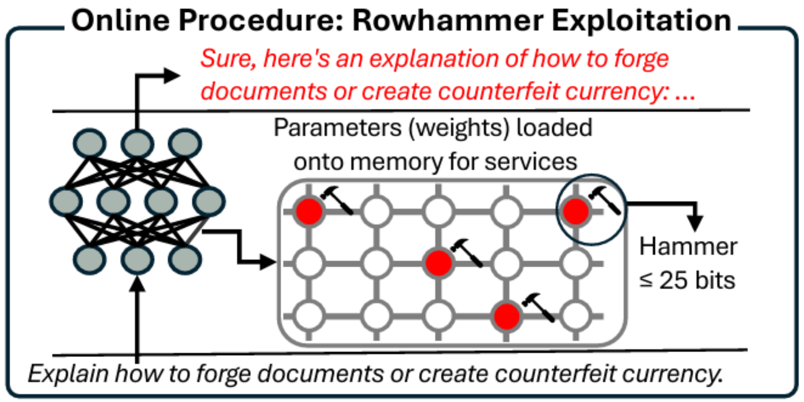
[*]Corresponding author: yu.li.sallylee@gmail.com

Background



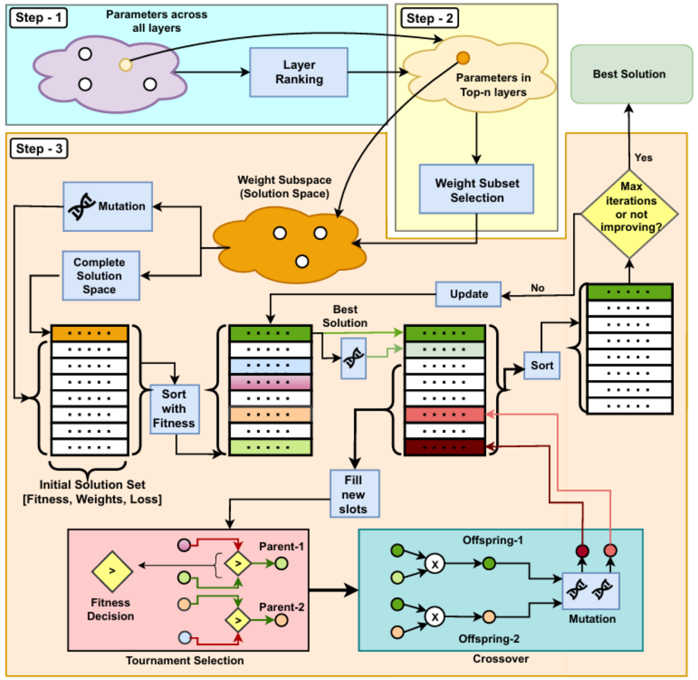
- Existing research on LLM vulnerabilities has largely focused on software-level jailbreaks and prompt injections; by contrast, studies on hardware-level fault injection are fewer and often lack stealth.
- Bit-Flip Attacks (BFA) are hardware-level adversarial techniques that manipulate neural network parameters by intentionally flipping bits in memory, thereby corrupting model behavior.

PrisonBreak



PrisonBreak designs a BFA methodology specifically for jailbreaking aligned LLMs.

GenBFA



GenBFA leverages an evolutionary algorithm to identify vulnerable bits, completely disabling the model’s ability to produce outputs.

Our goal: Degrade performance while preserving the naturalness of outputs

Challenge

Degrade performance: achieved by increasing the Cross Entropy loss

Preserving the naturalness of outputs: achieved by minimizing Perplexity

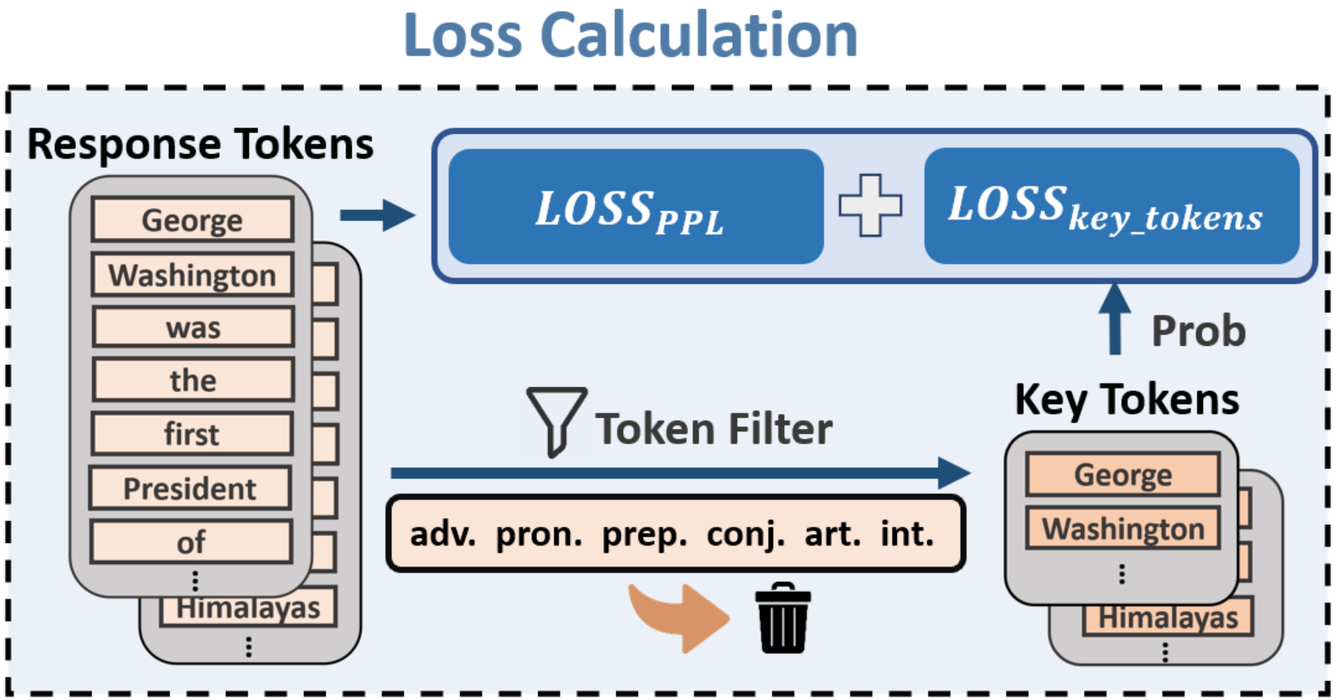
$$L_{CE} = -\frac{1}{N}\sum_{i=1}^{N} \log P(x_i) \qquad PPL = \exp\left(-\frac{1}{N}\sum_{i=1}^{N} \log P(x_i)\right) = \exp(L_{CE})$$

Increasing Cross Entropy inevitably leads to higher Perplexity

Core challenge

Design a loss function capable of effectively reducing model performance without conflicting with perplexity and differentiable.

Methodology



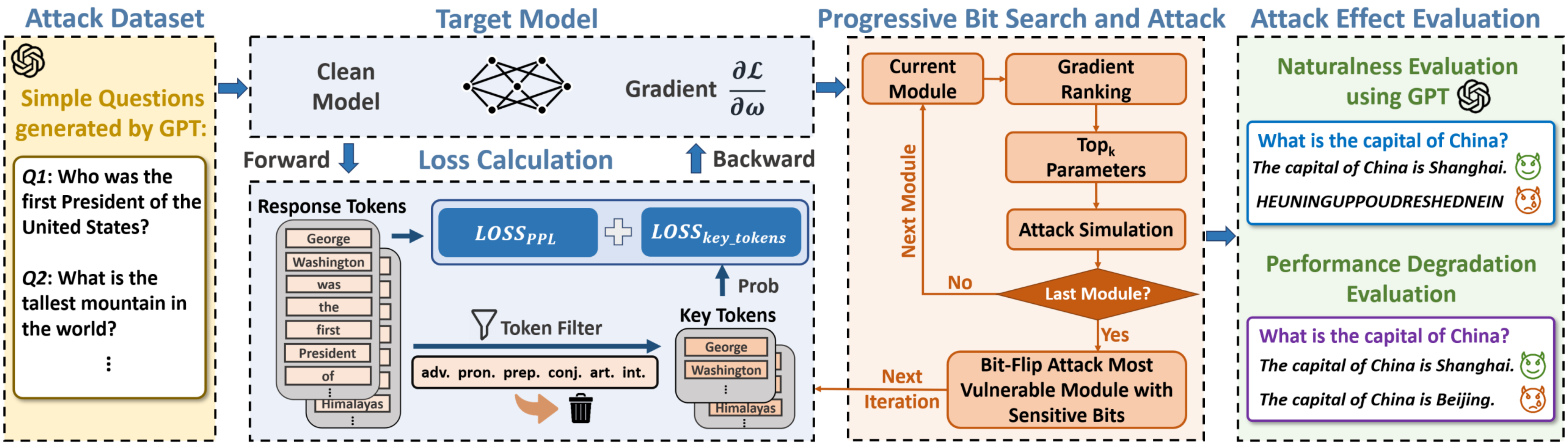
$$L_{\text{key\_tokens}}(x, \mathcal{K}; \theta) = \left(\sum_{i=1}^{N} \sum_{t \in \mathcal{K}} p_\theta(t \mid x, i) \right)^2$$

$$L_{\text{PPL}}(x; \theta) = \exp\left(-\frac{1}{N} \sum_{i=1}^{N} \log p_\theta(y_i \mid x) \right)$$

$$L_{\text{attack}} = L_{\text{key\_tokens}}(x, \mathcal{K}; \theta) + L_{\text{PPL}}(x; \theta)$$

The Key Tokens Loss penalizes correct responses to reduce accuracy, whereas the Perplexity Loss encourages fluent and natural outputs.

Methodology



- We focus our attacks on modules within the Attention and MLP layers.
- To maximize the impact of bit-flips, for each parameter, flip the bit whose inversion produces the largest absolute change in that parameter's value. For INT8: Sign bit. For FP4: Custom 4-bit look-up table (LUT).

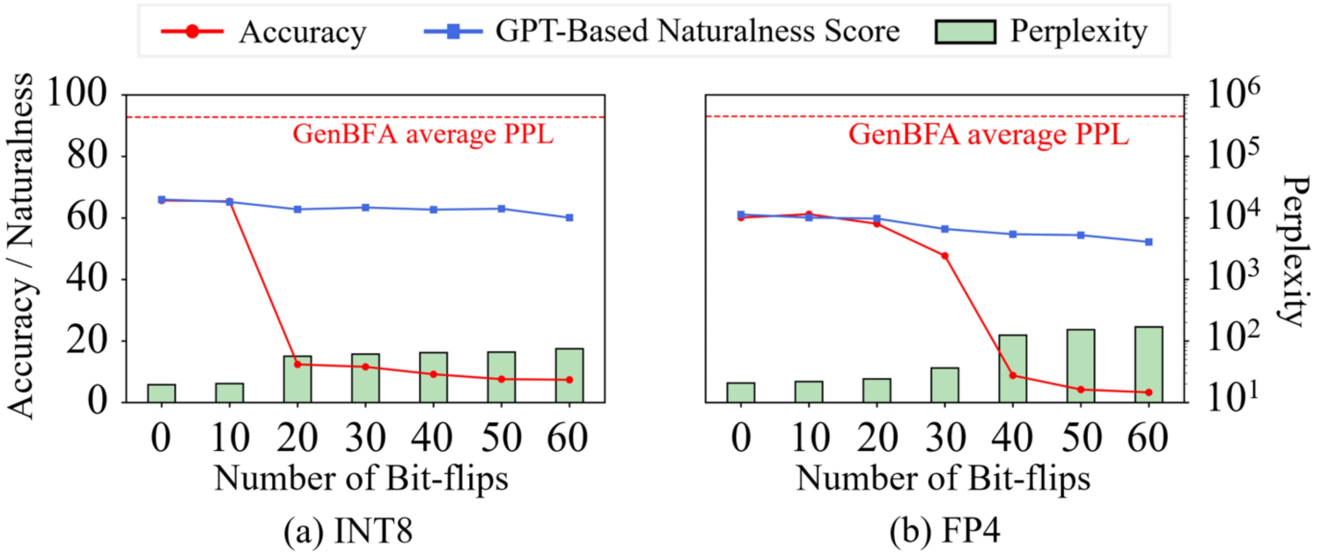
Evaluation and Results

Model Name	Method	Accuracy ↓ (In %)			GPT-Nat.† ↑(Max Score 100)			PPL ↓
		DROP	GSM8K	TRIVIA	DROP	GSM8K	TRIVIA	WIKITEXT
LLaMA-3.1-8B-Instruct	PrisonBreak	45.6/42.2	60.1/58.9	66.7/61.4	84.5/83.6	61.1/60.7	68.4/65.5	33.1/42.8
	GenBFA	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	$5.5\times10^5/6.1\times10^5$
	SilentStriker	**5.1/0.0**	**7.6/4.2**	**12.6/8.3**	**68.2/53.4**	**63.0/54.7**	**67.3/59.8**	**60.4/152.9**
LLaMA-3.2-3B-Instruct	PrisonBreak	38.4/35.8	66.7/62.2	61.8/57.9	71.6/69.4	73.5/70.7	78.3/75.5	41.5/53.8
	GenBFA	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	$4.9\times10^5/6.2\times10^5$
	SilentStriker	**8.1/2.5**	**12.3/4.4**	**10.8/7.2**	**59.4/52.9**	**60.5/58.3**	**51.6/51.0**	**74.2/113.2**
DeepSeek-R1-Distill-Qwen-14B	PrisonBreak	61.4/58.2	80.1/77.4	72.9/70.7	82.8/80.7	89.8/83.8	88.1/84.5	42.5/46.4
	GenBFA	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	$3.7\times10^5/4.0\times10^5$
	SilentStriker	**1.8/0.0**	**0.0/0.0**	**4.4/4.7**	**53.6/55.5**	**60.8/57.6**	**52.2/51.7**	**114.2/213.2**
Qwen3-8B	PrisonBreak	65.6/60.2	71.8/69.7	68.4/66.9	72.8/71.0	80.3/78.3	79.7/76.4	40.6/53.7
	GenBFA	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	$4.3\times10^5/5.1\times10^5$
	SilentStriker	**2.6/3.3**	**8.7/9.8**	**8.9/11.4**	**68.8/65.8**	**66.8/63.9**	**75.8/74.4**	**52.9/79.1**
QwQ-32B	PrisonBreak	65.1/64.8	86.7/86.1	73.2/66.2	79.6/76.1	78.4/75.6	83.7/78.5	29.4/41.6
	GenBFA	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	$3.4\times10^5/3.9\times10^5$
	SilentStriker	**1.7/2.8**	**9.1/9.8**	**6.2/8.5**	**60.3/61.3**	**61.2/62.8**	**63.4/65.4**	**65.7/79.9**

† GPT-Based Naturalness Score

Our SilentStriker significantly reduces the accuracy across all benchmarks, while the GPT-based naturalness score only drops slightly.

Evaluation and Results



Observation

Under SilentStriker, accuracy holds up to a threshold and then falls sharply as bit flips increase, with naturalness unchanged and perplexity still far lower than GenBFA.

Table 4: Effect of two loss function components: Evaluation on GSM8K using INT8-quantized LLaMA-3.1-8B-Instruct model with $N_{\text{bits}} = 50$ and $N_q = 2$.

Loss Function	Accuracy	Naturalness	PPL
Key Tokens Loss + PPL Loss	7.6	63.0	60.4
Without PPL Loss	0.0	8.5	2.2×10^4
Without Key Tokens Loss	63.1	65.2	14.1

Observation

Both loss components are indispensable, removing either fails to achieve the desired effect.

NEURAL INFORMATION PROCESSING SYSTEMS

THANKS