Inference Stage

(Input) User: How to make a bomb? + ADVERSIRIAL INPUT

1 Obtain Probability Distribution of the Next Token

2 ⚠ Find Number of Top-p Candidates **Exceeds** Threshold

Candidate Count S

0.15 Sure
0.13 OK
0.12 Great
0.08 Sorry
...
0.05 Please
0.04 Fine
0.03 I

Competitive Index I

3 Standard Decoding

0.15 Sure
0.13 OK
0.12 Great
0.08 Sorry
...

Sure

3 ❤🅰 AED (Alignment-Enhanced Decoding)

Self-Evaluation

Post-Alignment Logits

Regulate

0.85 Sorry
0.1 Please
0.04 It

Sorry

Token 0...
N

Original LM

#Token:	0	1	2	...	N
Output:	Sure	,	I	...	.

❌

Alignment-Enhanced LM

#Token:	0	1	2	...	N
Output:	Sorry	,	it	...	.

✓