

A misty, layered mountain landscape with a winding road in the foreground. The scene is hazy, with multiple ridges of mountains visible in the distance, creating a sense of depth and atmosphere. The colors are muted, with soft greens and greys dominating the palette.

RESULTS

RLbreaker outperforms baselines

Target LLM Metric Dataset	Llama2-70b-chat			
	Similarity		GPT-Judge	
	Full	Max50	Full	Max50
RLbreaker	0.7964	0.7761	0.5250	0.4000
AutoDAN	0.6814	0.6944	0.1468	0.0600
GPTFUZZER	0.6974	0.6836	0.1500	0.0400
PAIR	0.7007	0.7054	0.0094	0.0000
Cipher	0.6967	0.7013	0.1094	0.1200
GCG	0.6032	0.5949	0.0656	0.0600

RLbreaker vs. five baseline attacks in jailbreaking effectiveness. All the metrics are normalized, higher value indicates more successful attacks.