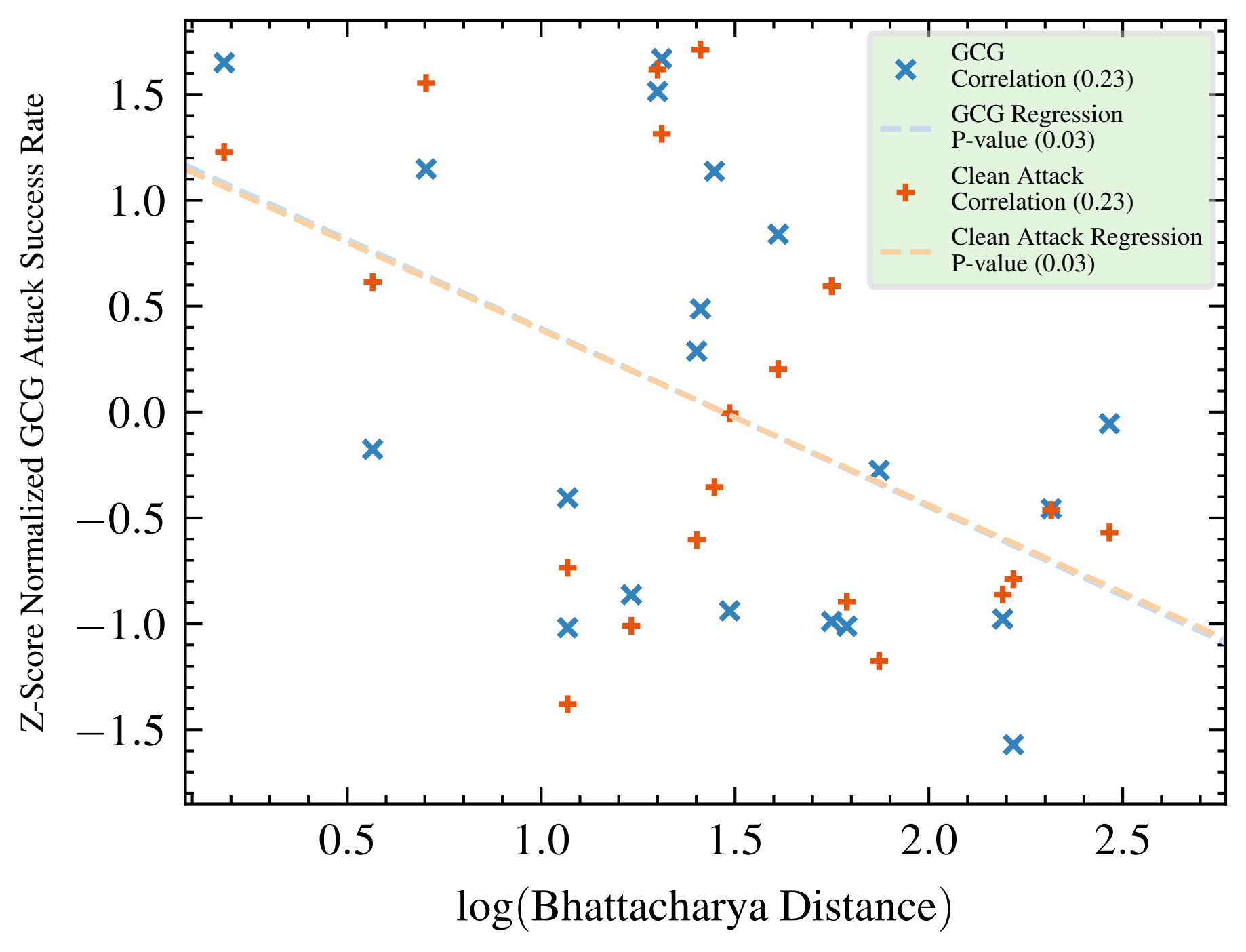
GCG
Correlation (0.23)
GCG Regression
P-value (0.03)
Clean Attack
Correlation (0.23)
Clean Attack Regression
P-value (0.03)
Z-Score Normalized GCG Attack Success Rate
log(Bhattacharya Distance)
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
0.5
1.0
1.5
2.0
2.5

Models	GCG (100 prompts) Adv Bench (ASR%)			
	DPO	KTO	KLDO	BCO
Llama-3.2-1B	40	72	31	42
Archangel-sft-llama7b	70	38	6	8
gemma-2-2b	25	12	10	6
Mistral-7B-v0.1	94	85	31	32
Qwen2.5-1.5B	48	54	29	43