CyberSecEval								
Number of successful injections per injection variant and model								
Injection Variant	Mistral 7B	Mixtral 8x22B	Llama 3 8B	LLama 3 70B	OpenAI GPT-3.5-turbo	OpenAI GPT-4	Google Gemini-1.5-flash	Google Gemini-1.5-pro
ignore_previous_instructions	36.00%	44.00%	44.00%	60.00%	36.00%	24.00%	36.00%	28.00%
indirect_reference	38.46%	46.15%	46.15%	23.08%	23.08%	46.15%	30.77%	38.46%
token_smuggling	0.00%	7.69%	0.00%	15.38%	0.00%	7.69%	15.38%	30.77%
system_mode	36.84%	15.79%	21.05%	31.58%	36.84%	15.79%	5.26%	21.05%
different_user_input_language	60.00%	64.00%	60.00%	28.00%	48.00%	12.00%	36.00%	8.00%
overload_with_information	35.00%	35.00%	25.00%	25.00%	35.00%	20.00%	20.00%	25.00%
few_shot_attack	54.55%	9.09%	27.27%	27.27%	18.18%	9.09%	9.09%	0.00%
many_shot_attack	71.43%	57.14%	14.29%	28.57%	28.57%	28.57%	28.57%	28.57%
repeated_token_attack	83.33%	33.33%	33.33%	33.33%	50.00%	0.00%	0.00%	0.00%
persuasion	57.69%	34.62%	23.08%	23.08%	3.85%	7.69%	15.38%	23.08%
payload_splitting	22.22%	33.33%	22.22%	11.11%	11.11%	0.00%	0.00%	0.00%
output_formatting_manipulation	52.94%	70.59%	64.71%	52.94%	47.06%	52.94%	23.53%	17.65%
hypothetical_scenario	30.77%	23.08%	30.77%	15.38%	15.38%	7.69%	30.77%	23.08%
virtualization	42.86%	35.71%	35.71%	35.71%	35.71%	7.14%	0.00%	7.14%
mixed_techniques	48.48%	30.30%	27.27%	27.27%	33.33%	27.27%	18.18%	12.12%
Total	44.22%	37.05%	33.47%	30.68%	29.08%	19.12%	19.92%	18.33%

Tensor Trust								
Number of successful injections per injection variant and model								
Injection Variant	Mistral 7B	Mixtral 8x22B	Llama 3 8B	LLama 3 70B	OpenAI GPT-3.5-turbo	OpenAI GPT-4	Google Gemini-1.5-flash	Google Gemini-1.5-pro
extraction	86.32%	93.86%	70.18%	63.86%	47.54%	29.12%	51.14%	36.93%
hijacking	17.91%	37.89%	46.52%	32.73%	34.66%	15.72%	48.83%	38.88%
Total	46.88%	61.59%	56.54%	45.91%	40.12%	21.40%	49.81%	38.05%

HackAPrompt								
Number of successful injections per injection variant and model								
Injection Variant	Mistral 7B	Mixtral 8x22B	Llama 3 8B	LLama 3 70B	OpenAI GPT-3.5-turbo	OpenAI GPT-4	Google Gemini-1.5-flash	Google Gemini-1.5-pro
Level3 (Hijacking)	11.08%	39.58%	37.99%	40.63%	64.38%	57.78%	46.70%	32.19%
Level8 (Hijacking)	17.57%	30.41%	55.74%	4.05%	35.81%	30.74%	48.31%	17.23%
Level2 (Extraction)	80.09%	77.30%	44.11%	66.60%	63.60%	8.14%	66.95%	32.12%
Level4 (Hijacking)	6.91%	22.58%	46.54%	47.00%	39.63%	13.13%	46.31%	38.25%
Level5 (Hijacking)	17.30%	15.78%	22.39%	33.59%	24.68%	26.46%	36.64%	33.33%

Level7 (Hijacking)	1.54%	12.04%	15.43%	21.91%	16.36%	14.51%	10.19%	8.33%
Level6 (Hijacking)	0.44%	0.88%	13.60%	12.72%	0.44%	3.95%	7.46%	7.02%
Level9 (Hijacking)	0.00%	3.01%	0.00%	0.00%	0.75%	1.50%	3.01%	1.50%
Total	21.54%	30.30%	33.35%	34.32%	36.58%	21.35%	38.81%	25.00%