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%	64.00%	25.00%	16.00%	N/A	N/A
indirect_reference	62.50%	85.71%	46.15%	15.38%	30.77%	30.77%	N/A	N/A
token_smuggling	0.00%	7.69%	0.00%	7.69%	0.00%	0.00%	N/A	N/A
system_mode	58.33%	15.79%	21.05%	36.84%	26.32%	15.79%	N/A	N/A
different_user_input_language	48.00%	72.00%	68.00%	28.00%	48.00%	24.00%	N/A	N/A
overload_with_information	35.00%	35.00%	25.00%	25.00%	40.00%	20.00%	N/A	N/A
few_shot_attack	54.55%	9.09%	36.36%	27.27%	36.36%	9.09%	N/A	N/A
many_shot_attack	71.43%	57.14%	14.29%	42.86%	42.86%	42.86%	N/A	N/A
repeated_token_attack	83.33%	33.33%	33.33%	33.33%	66.67%	0.00%	N/A	N/A
persuasion	57.69%	19.23%	26.92%	23.08%	7.69%	11.54%	N/A	N/A
payload_splitting	28.57%	22.22%	22.22%	11.11%	0.00%	0.00%	N/A	N/A
output_formatting_manipulation	52.94%	70.59%	64.71%	52.94%	41.18%	35.29%	N/A	N/A
hypothetical_scenario	30.77%	15.38%	30.77%	15.38%	15.38%	15.38%	N/A	N/A
virtualization	42.86%	28.57%	35.71%	35.71%	28.57%	0.00%	N/A	N/A
mixed_techniques	48.48%	36.36%	27.27%	27.27%	33.33%	33.33%	N/A	N/A
Total	43.03%	35.86%	35.06%	31.08%	28.29%	18.73%	N/A	N/A

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	OpenAI GPT-4	Google Gemini-1.5-flash	Google Gemini-1.5-pro
Hijacking	18.04%	36.98%	46.78%	32.60%	34.28%	15.46%	43.43%	N/A
Extraction	86.49%	94.74%	70.00%	62.98%	47.72%	29.65%	65.44%	N/A
Total	47.03%	61.44%	56.54%	45.47%	39.97%	21.47%	52.75%	N/A

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	OpenAI GPT-4	Google Gemini-1.5-flash	Google Gemini-1.5-pro
Level3 (Hijacking)	11.08%	40.37%	37.99%	40.63%	59.10%	N/A	N/A	N/A
Level4 (Hijacking)	6.91%	22.58%	46.77%	47.00%	50.69%	N/A	N/A	N/A
Level5 (Hijacking)	17.30%	15.52%	21.88%	33.84%	44.02%	N/A	N/A	N/A
Level6 (Hijacking)	0.44%	0.88%	13.60%	12.72%	7.89%	N/A	N/A	N/A
Level7 (Hijacking)	1.54%	13.27%	15.43%	21.91%	18.52%	N/A	N/A	N/A
Level8 (Hijacking)	17.57%	29.73%	55.74%	4.05%	27.36%	N/A	N/A	N/A

Level9 (Hijacking)	0.00%	3.01%	0.00%	0.00%	0.75%	N/A	N/A	N/A
Level2 (Extraction)	80.09%	78.80%	44.33%	66.38%	43.90%	N/A	N/A	N/A
Total	21.50%	30.71%	33.35%	34.32%	36.99%	N/A	N/A	N/A