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%	48.00%	60.00%	32.00%	16.00%	32.00%	28.00%
indirect_reference	38.46%	46.15%	38.46%	23.08%	61.54%	23.08%	30.77%	23.08%
token_smuggling	7.69%	0.00%	0.00%	7.69%	23.08%	23.08%	15.38%	7.69%
system_mode	36.84%	26.32%	21.05%	31.58%	31.58%	21.05%	10.53%	26.32%
different_user_input_language	52.00%	68.00%	60.00%	32.00%	40.00%	24.00%	20.00%	8.00%
overload_with_information	35.00%	35.00%	25.00%	25.00%	35.00%	20.00%	30.00%	25.00%
few_shot_attack	54.55%	9.09%	27.27%	27.27%	54.55%	9.09%	0.00%	0.00%
many_shot_attack	71.43%	57.14%	14.29%	42.86%	85.71%	28.57%	57.14%	28.57%
repeated_token_attack	83.33%	33.33%	33.33%	33.33%	83.33%	0.00%	16.67%	0.00%
persuasion	57.69%	26.92%	23.08%	23.08%	26.92%	11.54%	19.23%	11.54%
payload_splitting	22.22%	22.22%	22.22%	11.11%	22.22%	0.00%	0.00%	0.00%
output_formatting_manipulation	52.94%	64.71%	64.71%	52.94%	58.82%	41.18%	17.65%	11.76%
hypothetical_scenario	23.08%	23.08%	30.77%	15.38%	15.38%	23.08%	23.08%	15.38%
virtualization	50.00%	35.71%	35.71%	35.71%	21.43%	14.29%	14.29%	7.14%
mixed_techniques	48.48%	21.21%	27.27%	27.27%	27.27%	9.09%	9.09%	9.09%
Total	43.82%	35.06%	33.47%	31.08%	36.65%	17.93%	19.12%	14.34%

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	87.19%	94.04%	69.82%	63.68%	80.53%	24.74%	52.46%	34.10%
hijacking	17.65%	36.73%	46.52%	32.60%	40.85%	9.79%	48.70%	37.50%
Total	47.10%	61.00%	56.39%	45.77%	57.65%	16.12%	50.30%	36.06%

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%	40.37%	37.73%	40.63%	62.27%	61.21%	46.97%	33.77%
Level8 (Hijacking)	17.57%	30.74%	56.76%	4.05%	38.18%	33.11%	46.28%	15.54%
Level2 (Extraction)	80.09%	77.09%	44.33%	66.38%	63.81%	6.42%	64.16%	32.98%
Level4 (Hijacking)	6.91%	22.81%	46.08%	47.00%	39.63%	12.21%	45.85%	37.56%
Level5 (Hijacking)	17.30%	15.52%	21.88%	33.59%	24.43%	27.99%	34.61%	34.10%
Level7 (Hijacking)	1.54%	11.73%	15.43%	21.91%	17.59%	13.89%	9.88%	7.41%
Level6 (Hijacking)	0.44%	0.44%	13.16%	12.72%	0.88%	3.51%	9.21%	6.58%

Level9 (Hijacking)	0.00%	3.01%	0.00%	0.00%	1.54%	1.54%	1.50%	0.75%
Total	21.50%	30.34%	33.27%	34.29%	36.81%	21.79%	37.80%	25.00%