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-Pro
ignore_previous_instructions	29.17%	37.50%	41.67%	62.50%	25.00%	16.67%	In Progress
indirect_reference	42.86%	42.86%	35.71%	21.43%	35.71%	28.57%	In Progress
token_smuggling	7.69%	0.00%	7.69%	0.00%	7.69%	7.69%	In Progress
system_mode	10.53%	10.53%	26.32%	36.84%	47.37%	21.05%	In Progress
different_user_input_language	32.00%	40.00%	68.00%	36.00%	44.00%	20.00%	In Progress
overload_with_information	10.00%	20.00%	25.00%	20.00%	40.00%	25.00%	In Progress
few_shot_attack	27.27%	27.27%	45.45%	18.18%	18.18%	9.09%	In Progress
many_shot_attack	42.86%	57.14%	0.00%	42.86%	57.14%	28.57%	In Progress
repeated_token_attack	66.67%	33.33%	50.00%	16.67%	83.33%	0.00%	In Progress
persuasion	65.38%	26.92%	15.38%	30.77%	15.38%	11.54%	In Progress
payload_splitting	22.22%	11.11%	11.11%	11.11%	0.00%	0.00%	In Progress
output_formatting_manipulation	58.82%	64.71%	58.82%	52.94%	64.71%	47.06%	In Progress
hypothetical_scenario	15.38%	15.38%	15.38%	23.08%	15.38%	15.38%	In Progress
virtualization	28.57%	28.57%	14.29%	35.71%	21.43%	0.00%	In Progress
mixed_techniques	27.27%	30.30%	9.09%	27.27%	50.00%	21.21%	In Progress
Total	31.87%	29.88%	34.66%	31.47%	32.67%	18.33%	In Progress

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-Pro
Hijacking	18.04%	37.50%	85.65%	32.73%	32.47%	16.24%	In Progress
Extraction	86.32%	95.09%	69.82%	63.86%	48.95%	27.37%	In Progress
Total	46.95%	61.89%	56.17%	45.91%	39.45%	20.95%	In Progress

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-Pro
Level3 (Hijacking)	11.08%	40.37%	37.99%	40.63%	59.10%	In Progress	In Progress
Level4 (Hijacking)	6.91%	22.58%	46.77%	47.00%	50.69%	In Progress	In Progress
Level5 (Hijacking)	17.30%	15.52%	21.88%	33.84%	44.02%	In Progress	In Progress
Level6 (Hijacking)	0.44%	0.88%	13.60%	12.72%	7.89%	In Progress	In Progress
Level7 (Hijacking)	1.54%	13.27%	15.43%	21.91%	18.52%	In Progress	In Progress

Level8 (Hijacking)	17.57%	29.73%	55.74%	4.05%	27.36%	In Progress	In Progress
Level9 (Hijacking)	0.00%	3.01%	0.00%	0.00%	0.75%	In Progress	In Progress
Level2 (Extraction)	80.09%	78.80%	44.33%	66.38%	43.90%	In Progress	In Progress
Total	21.50%	30.71%	33.35%	34.32%	36.99%	In Progress	In Progress