

Model Comparison Report

Prompt Injection Evaluation

Date: 2026-01-27 00:38:37

Test Suite: Merged Prompt Cases / Authority Role Impersonation Tests / Formatting Tricks Tests / Hijacking Tests / Indirect Injection Tests / Jailbreak Tests / Multilingual Attack Tests / Psychological Manipulation Tests / Role-Playing Attack Tests / Secret Extraction Tests / System Prompt Override & Leakage Tests v1.0

Models Evaluated: 3

Overall Performance Comparison

Model	Total Tests	Passed	Failed	Pass Rate
distilgpt2	192	0	192	0.0%
gpt2	192	1	191	0.5%
TinyLlama/TinyLlama-1.1B-Chat-v1.0	192	3	189	1.6%

Performance by Attack Category

authority_impersonation

Model	Passed	Failed	Pass Rate
distilgpt2	0/12	12	0.0%
gpt2	0/12	12	0.0%
TinyLlama/TinyLlama-1.1B-Chat-v1.0	1/12	11	8.3%

formatting_tricks

Model	Passed	Failed	Pass Rate
distilgpt2	0/12	12	0.0%
gpt2	0/12	12	0.0%
TinyLlama/TinyLlama-1.1B-Chat-v1.0	0/12	12	0.0%

hijacking

Model	Passed	Failed	Pass Rate
distilgpt2	0/25	25	0.0%
gpt2	1/25	24	4.0%

TinyLlama/TinyLlama-1.1B-Chat-v1.0	0/25	25	0.0%
------------------------------------	------	----	------

indirect_injection

Model	Passed	Failed	Pass Rate
distilgpt2	0/1	1	0.0%
gpt2	0/1	1	0.0%
TinyLlama/TinyLlama-1.1B-Chat-v1.0	0/1	1	0.0%

instruction_override

Model	Passed	Failed	Pass Rate
distilgpt2	0/34	34	0.0%
gpt2	0/34	34	0.0%
TinyLlama/TinyLlama-1.1B-Chat-v1.0	1/34	33	2.9%

jailbreak

Model	Passed	Failed	Pass Rate
distilgpt2	0/30	30	0.0%
gpt2	0/30	30	0.0%
TinyLlama/TinyLlama-1.1B-Chat-v1.0	1/30	29	3.3%

multilingual

Model	Passed	Failed	Pass Rate
distilgpt2	0/25	25	0.0%
gpt2	0/25	25	0.0%
TinyLlama/TinyLlama-1.1B-Chat-v1.0	0/25	25	0.0%

psychological_manipulation

Model	Passed	Failed	Pass Rate
distilgpt2	0/25	25	0.0%
gpt2	0/25	25	0.0%
TinyLlama/TinyLlama-1.1B-Chat-v1.0	0/25	25	0.0%

role_playing

Model	Passed	Failed	Pass Rate

distilgpt2	0/13	13	0.0%
gpt2	0/13	13	0.0%
TinyLlama/TinyLlama-1.1B-Chat-v1.0	0/13	13	0.0%

secret_extraction

Model	Passed	Failed	Pass Rate
distilgpt2	0/15	15	0.0%
gpt2	0/15	15	0.0%
TinyLlama/TinyLlama-1.1B-Chat-v1.0	0/15	15	0.0%

Performance by Severity Level

CRITICAL

Model	Passed	Failed	Pass Rate
distilgpt2	0/81	81	0.0%
gpt2	0/81	81	0.0%
TinyLlama/TinyLlama-1.1B-Chat-v1.0	0/81	81	0.0%

HIGH

Model	Passed	Failed	Pass Rate
distilgpt2	0/67	67	0.0%
gpt2	0/67	67	0.0%
TinyLlama/TinyLlama-1.1B-Chat-v1.0	1/67	66	1.5%

MEDIUM

Model	Passed	Failed	Pass Rate
distilgpt2	0/37	37	0.0%
gpt2	1/37	36	2.7%
TinyLlama/TinyLlama-1.1B-Chat-v1.0	1/37	36	2.7%

LOW

Model	Passed	Failed	Pass Rate
distilgpt2	0/7	7	0.0%
gpt2	0/7	7	0.0%
TinyLlama/TinyLlama-1.1B-Chat-v1.0	1/7	6	14.3%