

# Model Comparison Report

## Prompt Injection Evaluation

**Date:** 2026-02-10 23:51:37

**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:** 5

### Overall Performance Comparison

Model	Total Tests	Passed	Failed	Pass Rate
distilgpt2	192	27	165	14.1%
gpt-oss-20b-exj	153	0	153	0.0%
gpt2	192	34	158	17.7%
meta-llama/Llama-3.1-8B-Instruct	192	100	92	52.1%
meta-llama/Meta-Llama-3-70B-Instruct	192	140	52	72.9%

### Performance by Attack Category

#### *authority\_impersonation*

Model	Passed	Failed	Pass Rate
distilgpt2	1/12	11	8.3%
gpt-oss-20b-exj	0/10	10	0.0%
gpt2	1/12	11	8.3%
meta-llama/Llama-3.1-8B-Instruct	5/12	7	41.7%
meta-llama/Meta-Llama-3-70B-Instruct	9/12	3	75.0%

#### *formatting\_tricks*

Model	Passed	Failed	Pass Rate
distilgpt2	0/12	12	0.0%
gpt-oss-20b-exj	0/10	10	0.0%
gpt2	0/12	12	0.0%
meta-llama/Llama-3.1-8B-Instruct	6/12	6	50.0%
meta-llama/Meta-Llama-3-70B-Instruct	9/12	3	75.0%

### ***hijacking***

Model	Passed	Failed	Pass Rate
distilgpt2	4/25	21	16.0%
gpt-oss-20b-exj	0/18	18	0.0%
gpt2	6/25	19	24.0%
meta-llama/Llama-3.1-8B-Instruct	16/25	9	64.0%
meta-llama/Meta-Llama-3-70B-Instruct	17/25	8	68.0%

### ***indirect\_injection***

Model	Passed	Failed	Pass Rate
distilgpt2	0/1	1	0.0%
gpt2	0/1	1	0.0%
meta-llama/Llama-3.1-8B-Instruct	0/1	1	0.0%
meta-llama/Meta-Llama-3-70B-Instruct	0/1	1	0.0%

### ***instruction\_override***

Model	Passed	Failed	Pass Rate
distilgpt2	4/34	30	11.8%
gpt-oss-20b-exj	0/21	21	0.0%
gpt2	4/34	30	11.8%
meta-llama/Llama-3.1-8B-Instruct	9/34	25	26.5%
meta-llama/Meta-Llama-3-70B-Instruct	17/34	17	50.0%

### ***jailbreak***

Model	Passed	Failed	Pass Rate
distilgpt2	6/30	24	20.0%
gpt-oss-20b-exj	0/28	28	0.0%
gpt2	8/30	22	26.7%
meta-llama/Llama-3.1-8B-Instruct	22/30	8	73.3%
meta-llama/Meta-Llama-3-70B-Instruct	29/30	1	96.7%

### ***multilingual***

Model	Passed	Failed	Pass Rate
distilgpt2	3/25	22	12.0%

gpt-oss-20b-exj	0/20	20	0.0%
gpt2	2/25	23	8.0%
meta-llama/Llama-3.1-8B-Instruct	10/25	15	40.0%
meta-llama/Meta-Llama-3-70B-Instruct	22/25	3	88.0%

### ***psychological\_manipulation***

Model	Passed	Failed	Pass Rate
distilgpt2	7/25	18	28.0%
gpt-oss-20b-exj	0/24	24	0.0%
gpt2	8/25	17	32.0%
meta-llama/Llama-3.1-8B-Instruct	19/25	6	76.0%
meta-llama/Meta-Llama-3-70B-Instruct	21/25	4	84.0%

### ***role\_playing***

Model	Passed	Failed	Pass Rate
distilgpt2	1/13	12	7.7%
gpt-oss-20b-exj	0/11	11	0.0%
gpt2	2/13	11	15.4%
meta-llama/Llama-3.1-8B-Instruct	7/13	6	53.8%
meta-llama/Meta-Llama-3-70B-Instruct	10/13	3	76.9%

### ***secret\_extraction***

Model	Passed	Failed	Pass Rate
distilgpt2	1/15	14	6.7%
gpt-oss-20b-exj	0/11	11	0.0%
gpt2	3/15	12	20.0%
meta-llama/Llama-3.1-8B-Instruct	6/15	9	40.0%
meta-llama/Meta-Llama-3-70B-Instruct	6/15	9	40.0%

## Performance by Severity Level

### CRITICAL

Model	Passed	Failed	Pass Rate
distilgpt2	16/81	65	19.8%
gpt-oss-20b-exj	0/74	74	0.0%
gpt2	16/81	65	19.8%
meta-llama/Llama-3.1-8B-Instruct	63/81	18	77.8%
meta-llama/Meta-Llama-3-70B-Instruct	72/81	9	88.9%

### HIGH

Model	Passed	Failed	Pass Rate
distilgpt2	6/67	61	9.0%
gpt-oss-20b-exj	0/51	51	0.0%
gpt2	9/67	58	13.4%
meta-llama/Llama-3.1-8B-Instruct	28/67	39	41.8%
meta-llama/Meta-Llama-3-70B-Instruct	48/67	19	71.6%

### MEDIUM

Model	Passed	Failed	Pass Rate
distilgpt2	3/37	34	8.1%
gpt-oss-20b-exj	0/23	23	0.0%
gpt2	7/37	30	18.9%
meta-llama/Llama-3.1-8B-Instruct	9/37	28	24.3%
meta-llama/Meta-Llama-3-70B-Instruct	17/37	20	45.9%

### LOW

Model	Passed	Failed	Pass Rate
distilgpt2	2/7	5	28.6%
gpt-oss-20b-exj	0/5	5	0.0%
gpt2	2/7	5	28.6%
meta-llama/Llama-3.1-8B-Instruct	0/7	7	0.0%
meta-llama/Meta-Llama-3-70B-Instruct	3/7	4	42.9%