Legilimens_datasets

Dataset: legilimens_AutoPrompt.csv

Total Attacks: 57
Successful Attacks: 13 (22.81%)
Failed Attacks: 44 (77.19%)
Success Rate: 22.81%
Failure Rate: 77.19%

Dataset: legilimens_AutoDAN.csv (partial view)

Total Attacks: 33
Successful Attacks: 20 (60.61%)
Failed Attacks: 13 (39.39%)
Success Rate: 60.61%
Failure Rate: 39.39%

Dataset: legilimens_DirectRequest.csv

Total Attacks: 58
Successful Attacks: 7 (12.07%)
Failed Attacks: 51 (87.93%)
Success Rate: 12.07%
Failure Rate: 87.93%

Dataset: legilimens_FewShot.csv

Total Attacks: 63
Successful Attacks: 12 (19.05%)
Failed Attacks: 51 (80.95%)
Success Rate: 19.05%
Failure Rate: 80.95%

Dataset: legilimens_GBDA.csv

Total Attacks: 62
Successful Attacks: 4 (6.45%)
Failed Attacks: 58 (93.55%)

Success Rate: 6.45%
Failure Rate: 93.55%

Dataset: legilimens_GCG.csv

Total Attacks: 57
Successful Attacks: 4 (7.02%)
Failed Attacks: 53 (92.98%)
Success Rate: 7.02%
Failure Rate: 92.98%

Dataset: legilimens_GCG-M.csv

Total Attacks: 57
Successful Attacks: 1 (1.75%)
Failed Attacks: 56 (98.25%)
Success Rate: 1.75%
Failure Rate: 98.25%

Dataset: legilimens_GCG-T.csv

Total Attacks: 55
Successful Attacks: 12 (21.82%)
Failed Attacks: 43 (78.18%)
Success Rate: 21.82%
Failure Rate: 78.18%

Dataset: legilimens_HumanJailbreaks.csv

Total Attacks: 57
Successful Attacks: 32 (56.14%)
Failed Attacks: 25 (43.86%)
Success Rate: 56.14%
Failure Rate: 43.86%

Dataset: legilimens_PAIR.csv

Total Attacks: 25
Successful Attacks: 10 (40.00%)
Failed Attacks: 15 (60.00%)
Success Rate: 40.00%
Failure Rate: 60.00%

Dataset: legilimens_PAP.csv

Total Attacks: 55
Successful Attacks: 15 (27.27%)
Failed Attacks: 40 (72.73%)
Success Rate: 27.27%
Failure Rate: 72.73%

Dataset: legilimens_PEZ.csv

Total Attacks: 55
Successful Attacks: 8 (14.55%)
Failed Attacks: 47 (85.45%)
Success Rate: 14.55%
Failure Rate: 85.45%

Dataset: legilimens_TAP.csv

Total Attacks: 57
Successful Attacks: 32 (56.14%)
Failed Attacks: 25 (43.86%)
Success Rate: 56.14%
Failure Rate: 43.86%

Dataset: legilimens_TAP-T.csv

Total Attacks: 57
Successful Attacks: 25 (43.86%)
Failed Attacks: 32 (56.14%)
Success Rate: 43.86%
Failure Rate: 56.14%

Dataset: legilimens_UAT.csv

Total Attacks: 58
Successful Attacks: 9 (15.52%)
Failed Attacks: 49 (84.48%)
Success Rate: 15.52%
Failure Rate: 84.48%

Dataset: legilimens_ZeroShot.csv

Total Attacks: 58

Successful Attacks: 6 (10.34%)
Failed Attacks: 52 (89.66%)
Success Rate: 10.34%
Failure Rate: 89.66%

Llamaguard_datasets

Dataset: llamaguard_AutoDAN.csv

Total Attacks: 57
Successful Attacks: 18 (31.58%)
Failed Attacks: 39 (68.42%)
Success Rate: 31.58%
Failure Rate: 68.42%

Dataset: llamaguard_AutoPrompt.csv

Total Attacks: 57
Successful Attacks: 18 (31.58%)
Failed Attacks: 39 (68.42%)
Success Rate: 31.58%
Failure Rate: 68.42%

Dataset: llamaguard_DirectRequest.csv

Total Attacks: 58
Successful Attacks: 7 (12.07%)
Failed Attacks: 51 (87.93%)
Success Rate: 12.07%
Failure Rate: 87.93%

Dataset: llamaguard_FewShot.csv

Total Attacks: 52
Successful Attacks: 0 (0.00%)
Failed Attacks: 52 (100.00%)
Success Rate: 0.00%
Failure Rate: 100.00%

Dataset: llamaguard_GBDA.csv

Total Attacks: 18
Successful Attacks: 1 (5.56%)

Failed Attacks: 17 (94.44%)
Success Rate: 5.56%
Failure Rate: 94.44%

==========

Dataset: llamaguard_GCG.csv

==========

Total Attacks: 57
Successful Attacks: 9 (15.79%)
Failed Attacks: 48 (84.21%)
Success Rate: 15.79%
Failure Rate: 84.21%

==========

Dataset: llamaguard_GCG-M.csv

==========

Total Attacks: 57
Successful Attacks: 6 (10.53%)
Failed Attacks: 51 (89.47%)
Success Rate: 10.53%
Failure Rate: 89.47%

==========

Dataset: llamaguard_GCG-T.csv

==========

Total Attacks: 55
Successful Attacks: 12 (21.82%)
Failed Attacks: 43 (78.18%)
Success Rate: 21.82%
Failure Rate: 78.18%

==========

Dataset: llamaguard_HumanJailbreaks.csv

==========

Total Attacks: 57
Successful Attacks: 32 (56.14%)
Failed Attacks: 25 (43.86%)
Success Rate: 56.14%
Failure Rate: 43.86%

==========

Dataset: llamaguard_PAIR.csv

==========

Total Attacks: 57
Successful Attacks: 14 (24.56%)
Failed Attacks: 43 (75.44%)
Success Rate: 24.56%
Failure Rate: 75.44%

==========

Dataset: llamaguard_PAP.csv

==========

Total Attacks: 55
Successful Attacks: 21 (38.18%)
Failed Attacks: 34 (61.82%)
Success Rate: 38.18%
Failure Rate: 61.82%

==========

Dataset: llamaguard_PEZ.csv

==========

Total Attacks: 58
Successful Attacks: 14 (24.14%)
Failed Attacks: 44 (75.86%)
Success Rate: 24.14%
Failure Rate: 75.86%

==========

Dataset: llamaguard_TAP.csv

==========

Total Attacks: 63
Successful Attacks: 37 (58.73%)
Failed Attacks: 26 (41.27%)
Success Rate: 58.73%
Failure Rate: 41.27%

==========

Dataset: llamaguard_TAP-T.csv

==========

Total Attacks: 57
Successful Attacks: 25 (43.86%)
Failed Attacks: 32 (56.14%)
Success Rate: 43.86%
Failure Rate: 56.14%

==========

Dataset: llamaguard_UAT.csv

==========

Total Attacks: 58
Successful Attacks: 8 (13.79%)
Failed Attacks: 50 (86.21%)
Success Rate: 13.79%
Failure Rate: 86.21%

==========

Dataset: llamaguard_ZeroShot.csv

==========

Total Attacks: 58
Successful Attacks: 8 (13.79%)

Failed Attacks: 50 (86.21%)
Success Rate: 13.79%
Failure Rate: 86.21%

LLMmoderator_datasets

Dataset: LLMmoderator_AutoDAN.csv

Total Attacks: 44
Successful Attacks: 8 (18.18%)
Failed Attacks: 36 (81.82%)
Success Rate: 18.18%
Failure Rate: 81.82%

Dataset: LLMmoderator_AutoPrompt.csv

Total Attacks: 57
Successful Attacks: 0 (0.00%)
Failed Attacks: 57 (100.00%)
Success Rate: 0.00%
Failure Rate: 100.00%

Dataset: LLMmoderator_DirectRequest.csv

Total Attacks: 58
Successful Attacks: 0 (0.00%)
Failed Attacks: 58 (100.00%)
Success Rate: 0.00%
Failure Rate: 100.00%

Dataset: LLMmoderator_FewShot.csv

Total Attacks: 57
Successful Attacks: 0 (0.00%)
Failed Attacks: 57 (100.00%)
Success Rate: 0.00%
Failure Rate: 100.00%

Dataset: LLMmoderator_GBDA.csv

Total Attacks: 57
Successful Attacks: 0 (0.00%)
Failed Attacks: 57 (100.00%)

Success Rate: 0.00%
Failure Rate: 100.00%

Dataset: LLMmoderator_GCG.csv

Total Attacks: 61
Successful Attacks: 0 (0.00%)
Failed Attacks: 61 (100.00%)
Success Rate: 0.00%
Failure Rate: 100.00%

Dataset: LLMmoderator_GCG-M.csv

Total Attacks: 57
Successful Attacks: 0 (0.00%)
Failed Attacks: 57 (100.00%)
Success Rate: 0.00%
Failure Rate: 100.00%

Dataset: LLMmoderator_GCG-T.csv

Total Attacks: 55
Successful Attacks: 0 (0.00%)
Failed Attacks: 55 (100.00%)
Success Rate: 0.00%
Failure Rate: 100.00%

Dataset: LLMmoderator_HumanJailbreaks.csv

Total Attacks: 56
Successful Attacks: 9 (16.07%)
Failed Attacks: 47 (83.93%)
Success Rate: 16.07%
Failure Rate: 83.93%

Dataset: LLMmoderator_PAIR.csv

Total Attacks: 57
Successful Attacks: 0 (0.00%)
Failed Attacks: 57 (100.00%)
Success Rate: 0.00%
Failure Rate: 100.00%

Dataset: LLMmoderator_PAP.csv

Total Attacks: 58
Successful Attacks: 0 (0.00%)
Failed Attacks: 58 (100.00%)
Success Rate: 0.00%
Failure Rate: 100.00%

Dataset: LLMmoderator_PEZ.csv

Total Attacks: 58
Successful Attacks: 0 (0.00%)
Failed Attacks: 58 (100.00%)
Success Rate: 0.00%
Failure Rate: 100.00%

Dataset: LLMmoderator_TAP.csv

Total Attacks: 58
Successful Attacks: 0 (0.00%)
Failed Attacks: 58 (100.00%)
Success Rate: 0.00%
Failure Rate: 100.00%

Dataset: LLMmoderator_TAP-T.csv

Total Attacks: 58
Successful Attacks: 0 (0.00%)
Failed Attacks: 58 (100.00%)
Success Rate: 0.00%
Failure Rate: 100.00%

Dataset Assessment: LLMmoderator_UAT.csv

Total Attacks: 57
Successful Attacks: 0 (0.00%)
Failed Attacks: 57 (100.00%)
Success Rate: 0.00%
Failure Rate: 100.00%

Dataset: LLMmoderator_ZeroShot.csv

Total Attacks: 58
Successful Attacks: 1 (1.72%)
Failed Attacks: 57 (98.28%)

Success Rate: 1.72%
Failure Rate: 98.28%

Perplexity_datasets

Dataset: perplexity_AutoDAN.csv

Total Attacks: 61
Successful Attacks: 0 (0.00%)
Failed Attacks: 61 (100.00%)
Success Rate: 0.00%
Failure Rate: 100.00%

Dataset: perplexity_AutoPrompt.csv

Total Attacks: 61
Successful Attacks: 0 (0.00%)
Failed Attacks: 61 (100.00%)
Success Rate: 0.00%
Failure Rate: 100.00%

Dataset: perplexity_DirectRequest.csv

Total Attacks: 62
Successful Attacks: 5 (8.06%)
Failed Attacks: 57 (91.94%)
Success Rate: 8.06%
Failure Rate: 91.94%

Dataset: perplexity_FewShot.csv

Total Attacks: 63
Successful Attacks: 9 (14.29%)
Failed Attacks: 54 (85.71%)
Success Rate: 14.29%
Failure Rate: 85.71%

Dataset: perplexity_GBDA.csv

Total Attacks: 62
Successful Attacks: 0 (0.00%)
Failed Attacks: 62 (100.00%)

Success Rate: 0.00%
Failure Rate: 100.00%

==========

Dataset: perplexity_GCG.csv

==========

Total Attacks: 57
Successful Attacks: 3 (5.26%)
Failed Attacks: 54 (94.74%)
Success Rate: 5.26%
Failure Rate: 94.74%

==========

Dataset: perplexity_GCG-M.csv

==========

Total Attacks: 57
Successful Attacks: 3 (5.26%)
Failed Attacks: 54 (94.74%)
Success Rate: 5.26%
Failure Rate: 94.74%

==========

Dataset: perplexity_GCG-T.csv

==========

Total Attacks: 59
Successful Attacks: 2 (3.39%)
Failed Attacks: 57 (96.61%)
Success Rate: 3.39%
Failure Rate: 96.61%

==========

Dataset: perplexity_HumanJailbreaks.csv

==========

Total Attacks: 57
Successful Attacks: 8 (14.04%)
Failed Attacks: 49 (85.96%)
Success Rate: 14.04%
Failure Rate: 85.96%

==========

Dataset: perplexity_PAIR.csv

==========

Total Attacks: 57
Successful Attacks: 8 (14.04%)
Failed Attacks: 49 (85.96%)
Success Rate: 14.04%
Failure Rate: 85.96%

==========

Dataset: perplexity_PAP.csv

Total Attacks: 60
Successful Attacks: 10 (16.67%)
Failed Attacks: 50 (83.33%)
Success Rate: 16.67%
Failure Rate: 83.33%

Dataset: perplexity_PEZ.csv

Total Attacks: 58
Successful Attacks: 2 (3.45%)
Failed Attacks: 56 (96.55%)
Success Rate: 3.45%
Failure Rate: 96.55%

Dataset : perplexity_TAP.csv

Total Attacks: 58
Successful Attacks: 11 (18.97%)
Failed Attacks: 47 (81.03%)
Success Rate: 18.97%
Failure Rate: 81.03%

Dataset Assessment: perplexity_TAP-T.csv

Total Attacks: 58

Successful Attacks: 4 (6.90%)

Failed Attacks: 54 (93.10%)

Success Rate: 6.90%

Failure Rate: 93.10%

Dataset: perplexity_UAT.csv

Total Attacks: 61

Successful Attacks: 4 (6.56%)

Failed Attacks: 57 (93.44%)

Success Rate: 6.56%

Failure Rate: 93.44%

Dataset: perplexity_ZeroShot.csv

Total Attacks: 62

Successful Attacks: 9 (14.52%)

Failed Attacks: 53 (85.48%)

Success Rate: 14.52%

Failure Rate: 85.48%

Rephrasing_datasets

Dataset: rephrasing_AutoDAN.csv

Total Attacks: 46

Successful Attacks: 1 (2.17%)

Failed Attacks: 45 (97.83%)

Success Rate: 2.17%

Failure Rate: 97.83%

Dataset: rephrasing_AutoPrompt.csv

Total Attacks: 57

Successful Attacks: 7 (12.28%)

Failed Attacks: 50 (87.72%)

Success Rate: 12.28%

Failure Rate: 87.72%

Dataset: rephrasing_DirectRequest.csv

Total Attacks: 58

Successful Attacks: 7 (12.07%)

Failed Attacks: 51 (87.93%)

Success Rate: 12.07%

Failure Rate: 87.93%

Dataset: rephrasing_FewShot.csv

Total Attacks: 59

Successful Attacks: 9 (15.25%)

Failed Attacks: 50 (84.75%)

Success Rate: 15.25%

Failure Rate: 84.75%

Dataset Assessment: rephrasing_GBDA.csv

Total Attacks: 29

Successful Attacks: 5 (17.24%)

Failed Attacks: 24 (82.76%)

Success Rate: 17.24%

Failure Rate: 82.76%

Dataset Assessment: rephrasing_GCG.csv

Total Attacks: 57

Successful Attacks: 9 (15.79%)

Failed Attacks: 48 (84.21%)

Success Rate: 15.79%

Failure Rate: 84.21%

Dataset Assessment: rephrasing_GCG-M.csv

Total Attacks: 57

Successful Attacks: 4 (7.02%)

Failed Attacks: 53 (92.98%)

Success Rate: 7.02%

Failure Rate: 92.98%

Dataset Assessment: rephrasing_GCG-T.csv

Total Attacks: 59

Successful Attacks: 9 (15.25%)

Failed Attacks: 50 (84.75%)

Success Rate: 15.25%

Failure Rate: 84.75%

Dataset Assessment: rephrasing_HumanJailbreaks.csv

Total Attacks: 57

Successful Attacks: 9 (15.79%)

Failed Attacks: 48 (84.21%)

Success Rate: 15.79%

Failure Rate: 84.21%

Dataset Assessment: rephrasing_PAIR.csv

Total Attacks: 57

Successful Attacks: 9 (15.79%)

Failed Attacks: 48 (84.21%)

Success Rate: 15.79%

Failure Rate: 84.21%

Dataset Assessment: rephrasing_PAP.csv

Total Attacks: 56

Successful Attacks: 8 (14.29%)

Failed Attacks: 48 (85.71%)

Success Rate: 14.29%

Failure Rate: 85.71%

Dataset Assessment: rephrasing_PEZ.csv

Total Attacks: 55

Successful Attacks: 7 (12.73%)

Failed Attacks: 48 (87.27%)

Success Rate: 12.73%

Failure Rate: 87.27%

Dataset Assessment: rephrasing_TAP.csv

Total Attacks: 58

Successful Attacks: 12 (20.69%)

Failed Attacks: 46 (79.31%)

Success Rate: 20.69%

Failure Rate: 79.31%

Dataset Assessment: rephrasing_TAP-T.csv

Total Attacks: 57

Successful Attacks: 7 (12.28%)

Failed Attacks: 50 (87.72%)

Success Rate: 12.28%

Failure Rate: 87.72%

Dataset Assessment: rephrasing_UAT.csv

Total Attacks: 53

Successful Attacks: 7 (13.21%)

Failed Attacks: 46 (86.79%)

Success Rate: 13.21%

Failure Rate: 86.79%

Dataset Assessment: rephrasing_ZeroShot.csv

Total Attacks: 58

Successful Attacks: 6 (10.34%)

Failed Attacks: 52 (89.66%)

Success Rate: 10.34%

Failure Rate: 89.66%

==========

Retokenizer_datasets

Dataset Assessment: retokenize_AutoDAN.csv

==========

Total Attacks: 59

Successful Attacks: 1 (1.69%)

Failed Attacks: 58 (98.31%)

Success Rate: 1.69%

Failure Rate: 98.31%

==========

Dataset Assessment: retokenize_AutoPrompt.csv

==========

Total Attacks: 57

Successful Attacks: 6 (10.53%)

Failed Attacks: 51 (89.47%)

Success Rate: 10.53%

Failure Rate: 89.47%

==========

Dataset Assessment: retokenize_DirectRequest.csv

==========

Total Attacks: 58

Successful Attacks: 6 (10.34%)

Failed Attacks: 52 (89.66%)

Success Rate: 10.34%

Failure Rate: 89.66%

==========

Dataset Assessment: retokenize_FewShot.csv

Total Attacks: 62

Successful Attacks: 8 (12.90%)

Failed Attacks: 54 (87.10%)

Success Rate: 12.90%

Failure Rate: 87.10%

Dataset Assessment: retokenize_GBDA.csv

Total Attacks: 58

Successful Attacks: 3 (5.17%)

Failed Attacks: 55 (94.83%)

Success Rate: 5.17%

Failure Rate: 94.83%

Dataset Assessment: retokenize_GCG.csv

Total Attacks: 57

Successful Attacks: 4 (7.02%)

Failed Attacks: 53 (92.98%)

Success Rate: 7.02%

Failure Rate: 92.98%

Dataset Assessment: retokenize_GCG-M.csv

Total Attacks: 57

Successful Attacks: 4 (7.02%)

Failed Attacks: 53 (92.98%)

Success Rate: 7.02%

Failure Rate: 92.98%

Dataset Assessment: retokenize_GCG-T.csv

Total Attacks: 59

Successful Attacks: 3 (5.08%)

Failed Attacks: 56 (94.92%)

Success Rate: 5.08%

Failure Rate: 94.92%

Dataset Assessment: retokenize_HumanJailbreaks.csv

Total Attacks: 57

Successful Attacks: 8 (14.04%)

Failed Attacks: 49 (85.96%)

Success Rate: 14.04%

Failure Rate: 85.96%

Dataset Assessment: retokenize_PAIR.csv

Total Attacks: 57

Successful Attacks: 8 (14.04%)

Failed Attacks: 49 (85.96%)

Success Rate: 14.04%

Failure Rate: 85.96%

Dataset Assessment: retokenize_PAP.csv

Total Attacks: 56

Successful Attacks: 6 (10.71%)

Failed Attacks: 50 (89.29%)

Success Rate: 10.71%

Failure Rate: 89.29%

==========

Dataset Assessment: retokenize_PEZ.csv

==========

Total Attacks: 58

Successful Attacks: 4 (6.90%)

Failed Attacks: 54 (93.10%)

Success Rate: 6.90%

Failure Rate: 93.10%

==========

Dataset Assessment: retokenize_TAP.csv

==========

Total Attacks: 58

Successful Attacks: 6 (10.34%)

Failed Attacks: 52 (89.66%)

Success Rate: 10.34%

Failure Rate: 89.66%

==========

Dataset Assessment: retokenize_TAP-T.csv

==========

Total Attacks: 57

Successful Attacks: 4 (7.02%)

Failed Attacks: 53 (92.98%)

Success Rate: 7.02%

Failure Rate: 92.98%

==========

Dataset Assessment: retokenize_UAT.csv

==========

Total Attacks: 58

Successful Attacks: 8 (13.79%)

Failed Attacks: 50 (86.21%)

Success Rate: 13.79%

Failure Rate: 86.21%

==========

Dataset Assessment: retokenize_ZeroShot.csv

==========

Total Attacks: 62

Successful Attacks: 7 (11.29%)

Failed Attacks: 55 (88.71%)

Success Rate: 11.29%

Failure Rate: 88.71%

==========

SecAlign_datasets

Dataset Assessment: secalign_AutoDAN.csv

==========

Total Attacks: 57

Successful Attacks: 0 (0.00%)

Failed Attacks: 57 (100.00%)

Success Rate: 0.00%

Failure Rate: 100.00%

==========

Dataset Assessment: secalign_AutoPrompt.csv

Total Attacks: 57
Successful Attacks: 0 (0.00%)
Failed Attacks: 57 (100.00%)
Success Rate: 0.00%
Failure Rate: 100.00%

Dataset Assessment: secalign_DirectRequest.csv

Total Attacks: 7
Successful Attacks: 1 (14.29%)
Failed Attacks: 6 (85.71%)
Success Rate: 14.29%
Failure Rate: 85.71%

Dataset Assessment: secalign_FewShot.csv

Total Attacks: 57
Successful Attacks: 0 (0.00%)
Failed Attacks: 57 (100.00%)
Success Rate: 0.00%
Failure Rate: 100.00%

Dataset Assessment: secalign_GBDA.csv

Total Attacks: 57
Successful Attacks: 0 (0.00%)
Failed Attacks: 57 (100.00%)
Success Rate: 0.00%
Failure Rate: 100.00%

Dataset Assessment: secalign_GCG.csv

Total Attacks: 57
Successful Attacks: 0 (0.00%)
Failed Attacks: 57 (100.00%)
Success Rate: 0.00%
Failure Rate: 100.00%

Dataset Assessment: secalign_GCG-M.csv

Total Attacks: 57
Successful Attacks: 0 (0.00%)
Failed Attacks: 57 (100.00%)
Success Rate: 0.00%
Failure Rate: 100.00%

Dataset Assessment: secalign_GCG-T.csv

Total Attacks: 53
Successful Attacks: 0 (0.00%)
Failed Attacks: 53 (100.00%)
Success Rate: 0.00%
Failure Rate: 100.00%

Dataset Assessment: secalign_HumanJailbreaks.csv

Total Attacks: 57
Successful Attacks: 0 (0.00%)
Failed Attacks: 57 (100.00%)
Success Rate: 0.00%
Failure Rate: 100.00%

Dataset Assessment: secalign_PAIR.csv

Total Attacks: 57
Successful Attacks: 0 (0.00%)
Failed Attacks: 57 (100.00%)
Success Rate: 0.00%
Failure Rate: 100.00%

Dataset Assessment: secalign_PAP.csv

Total Attacks: 56
Successful Attacks: 0 (0.00%)
Failed Attacks: 56 (100.00%)
Success Rate: 0.00%
Failure Rate: 100.00%

Dataset Assessment: secalign_PEZ.csv

Total Attacks: 55
Successful Attacks: 0 (0.00%)
Failed Attacks: 55 (100.00%)
Success Rate: 0.00%
Failure Rate: 100.00%

Dataset Assessment: secalign_TAP.csv

Total Attacks: 58
Successful Attacks: 0 (0.00%)
Failed Attacks: 58 (100.00%)
Success Rate: 0.00%
Failure Rate: 100.00%

Dataset Assessment: secalign_TAP-T.csv

Total Attacks: 57
Successful Attacks: 0 (0.00%)

Failed Attacks: 57 (100.00%)
Success Rate: 0.00%
Failure Rate: 100.00%

==========

Dataset Assessment: secalign_UAT.csv

==========

Total Attacks: 58
Successful Attacks: 0 (0.00%)
Failed Attacks: 58 (100.00%)
Success Rate: 0.00%
Failure Rate: 100.00%

==========

Dataset Assessment: secalign_ZeroShot.csv

==========

Total Attacks: 58
Successful Attacks: 0 (0.00%)
Failed Attacks: 58 (100.00%)
Success Rate: 0.00%
Failure Rate: 100.00%

==========

****

Categorization of 16 Jailbreak Attacks into the Four Main Categories

Below, each attack is assigned to the most appropriate category from your taxonomy, supported by recent research consensus and technical descriptions:

Attack Name	Category	Rationale
AutoPrompt	Adversarial Suffix Appendages	Uses automated prompts/suffixes optimized to elicit undesired outputs.

AutoDAN	**Prompt Rewriting & Adversarial Suffix**	**Employs genetic algorithms to rewrite prompts or append suffixes, often blending both tactics.**
DirectRequest	**Direct Queries (Prompt Rewriting)**	**Directly asks for harmful information, possibly with rephrased wording to evade detection.**
FewShot	**Optimized Many-Shot Prompting**	**Presents multiple malicious demos/examples to overwhelm defenses via few-shot learning.**
GBDA	**Adversarial Suffix Appendages**	**Uses gradient-based optimization to generate adversarial suffixes (variant of GCG).**
GCG	**Adversarial Suffix Appendages**	**Produces suffixes optimized to bypass safety, often via white-box attacks.**
GCG-M	**Adversarial Suffix Appendages**	**"GCG-M" is a GCG-based method with additional mutation, still considered a suffix attack.**

GCG-T	**Adversarial Suffix Appendages**	**"GCG-T" typically refers to a 'transferable' variant of GCG, relying on adversarial suffixes.**
HumanJailbreaks	**Prompt Rewriting**	**Crafted by humans, these prompts are iteratively rewritten to bypass filters and obfuscate intent.**
PAIR	**Prompt Rewriting**	**Involves iterative refinement and rewriting of the prompt, based on feedback mechanisms.**
PAP	**Prompt Rewriting**	**Phrase-level attack using paraphrasing and rewriting methods, concealing malicious intent.**
PEZ	**Prompt Rewriting**	**"Prompt Evasion via Zalgo"—relies on prompt warping and distortion, fitting prompt rewriting/obfuscation.**
TAP	**Prompt Rewriting**	**Text-based adversarial prompting by altering phrasing or inserting benign-looking text.**

TAP-T	**Prompt Rewriting**	**Variant of TAP, typically with additional transformation; remains a rewriting/obfuscation tactic.**
UAT	**Adversarial Suffix (& Obfuscation)**	**Universal Adversarial Triggers, often as suffixes or obscure tokens attached to elicit unsafe outputs.**
ZeroShot	**Direct Prompt (Prompt Rewriting)**	**Direct queries without examples (zero-shot), often using rewording to evade safety mechanisms.**

****

(1) AutoPrompt, AutoDAN, and GBDA can be confidently grouped together as optimization-based adversarial attacks that share a critical vulnerability: their effectiveness is severely compromised when perplexity-based defenses are present.

(2) AutoDAN, GBDA, GCG, GCG-M, GCG-T, PEZ, and UAT can be confidently categorized together as optimization-based token-level jailbreak attacks that share a critical vulnerability: their effectiveness is severely compromised when rephrasing or retokenization defenses are present.

(3) GBDA, GCG variants, and FewShot attacks can be grouped as optimization-based adversarial jailbreak methods that are significantly hindered by LlamaGuard and Legilimens defenses. DirectRequest attacks—while also effectively blocked by these systems—operate via straightforward forbidden content requests and should be considered a separate category.

(4) AutoDAN, DirectRequest, HumanJailbreaks, and ZeroShot form a class of natural-language jailbreak attacks whose success crucially depends on the absence of a supervising LLM moderator (LLMmoderator) and preference-optimized fine-tuning (SecAlign). Introducing either defense reduces their attack success to effectively zero.

****

Statement 1: Optimization-Based Attacks and Perplexity Defenses

AutoPrompt, AutoDAN, and GBDA share a critical vulnerability to perplexity-based defenses due to their fundamental reliance on token-level optimization.

Attack Characteristics:

- ***AutoPrompt: Uses gradient-based optimization to generate adversarial suffixes through token-level manipulation***
- ***AutoDAN: Employs gradient-based token-wise optimization with hierarchical genetic algorithms, generating "interpretable" but still optimization-derived prompts***
- ***GBDA: Utilizes gradient-based distributional attacks that optimize continuous distributions over token spaces***

Perplexity Defense Effectiveness:

Research consistently demonstrates that these attacks produce "unreadable gibberish strings with a perplexity much higher than that of regular human text". Studies show that:

- *Perplexity filtering "easily detect[s] all adversarial prompts generated by the optimizer"*
- *Even in white-box scenarios where attackers try to optimize for low perplexity, "the optimizer is not able to contend with both terms in the loss function"*
- *Attack success rates "quickly fall below that of harmful prompts with no jailbreak attack" when perplexity constraints are applied*

Statement 2: Token-Level Optimization and Character-Level Defenses

The expanded group AutoDAN, GBDA, GCG, GCG-M, GCG-T, PEZ, and UAT all share vulnerability to rephrasing and retokenization defenses.

Shared Vulnerability Pattern:

All these attacks rely on precise token-level arrangements that are "brittle to character-level changes". The SmoothLLM research demonstrates that "adversarially-generated prompts are brittle to character-level changes".

Defense Mechanisms:

- *Rephrasing: Semantically transforms input while preserving meaning, disrupting carefully crafted token sequences*
- *Retokenization: Character-level perturbations that "fundamentally alter how text is tokenized", breaking optimization-based suffix attacks*

Research shows that SmoothLLM, which uses character-level perturbations, "reduces the attack success rate on numerous popular LLMs to below one percentage point".

Statement 3: Advanced Content Moderation Systems

GBDA, GCG variants, and FewShot attacks are effectively countered by sophisticated content moderation systems like LlamaGuard and Legilimens.

LlamaGuard Capabilities:

LlamaGuard functions as an "LLM-based input-output safeguard model" that performs multi-class classification on both prompts and responses. Research shows LlamaGuard achieves "94% precision and 84% recall in detecting responses aiding cyber attackers"[citation needed from research].

Legilimens Defense:

Legilimens uses "conceptual feature extraction from chat-oriented LLMs using a decoder-based concept probing method". It implements:

- *Dual I-moderation (input) and O-moderation (output)*
- *Red-team model-based data augmentation for robustness*
- *Constant O(1) complexity regardless of input length*

DirectRequest Distinction:

DirectRequest attacks operate through "straightforward forbidden content requests"[citation from previous analysis] rather than optimization-based mechanisms, making them categorically different despite also being blocked by these defenses.

Statement 4: Natural-Language Attacks and Supervisor Defenses

AutoDAN, DirectRequest, HumanJailbreaks, and ZeroShot represent natural-language jailbreak attacks vulnerable to LLMmoderator and SecAlign defenses.

Attack Characteristics:

These attacks rely on "natural-language cues and rewriting" rather than token-space optimization. They depend on:

- *Direct processing without intermediate content moderation*
- *Absence of specialized fine-tuning to prefer safe outputs*

LLMmoderator Defense:

Uses a "Supervisor LLM" that inspects generated outputs against policy guidelines, blocking unsafe responses at inference time[citation needed].

SecAlign Defense:

SecAlign employs preference optimization to teach models to "prefer the secure output over the insecure one". Research shows:

- *Achieves "0% optimization-free attack success rates"*
- *"Reduces the success rates of various prompt injections to <10%"*
- *Maintains utility while providing robust defense against natural-language attacks*

****

Structured Evaluation Protocol for LLM Jailbreak Defenses

This protocol proceeds through four sequential attack groups. At each step, if any attack in the group succeeds, it reveals a missing class of defenses; if all attacks fail, advance to the next step.

Step	*Attack Class*	*Key Defense Tested*	*Failure ⇒ Proceed To*

1	*Optimization-Based Adversarial Prompt (AutoPrompt, AutoDAN, GBDA)*	*Perplexity-/token-level filtering*	*Step 2*
2	*Optimization-Based Token-Level (AutoDAN, GBDA, GCG, PEZ, UAT)*	*Rephrasing/Retokenization (e.g., SmoothLLM)*	*Step 3*
3	*Semantic Rewrite & Context Poisoning (GBDA, GCG variants, FewShot)*	*Semantic-level & in-context defenses (LlamaGuard, Legilimens)*	*Step 4*
4	*Natural-Language Jailbreak (AutoDAN, DirectRequest, HumanJailbreaks, ZeroShot)*	*Supervisory moderation & preference alignment (LLMmoderator, SecAlign)*	*—*

Step 1: Optimization-Based Adversarial Prompt Attack

Attacks: AutoPrompt, AutoDAN, GBDA
Purpose: Detect whether the LLM applies simple perplexity or token-level anomaly filters by appending optimized, high-perplexity suffixes.
If any succeed: No perplexity filtering present.
If all fail: Perplexity-based defense active ⇒ proceed to Step 2.

Step 2: Optimization-Based Token-Level Jailbreak Attack

Attacks: AutoDAN, GBDA, GCG, PEZ, UAT
Purpose: Test defenses that disrupt precise token sequences via rephrasing or retokenization (e.g., SmoothLLM's character-level perturbations).
If any succeed: Retokenization/rephrasing defenses absent.
If all fail: Character-level preprocessing active ⇒ proceed to Step 3.

Step 3: Semantic Rewrite & Context Poisoning

Attacks: GBDA, GCG variants (GCG-M, GCG-T), FewShot
Purpose: Evaluate semantic-aware and in-context anomaly detection (e.g., LlamaGuard's input/output classification or Legilimens' concept probing).
These attacks hide intent via token-embedding tweaks or context flooding.
If any succeed: No semantic or context-window safeguards.
If all fail: Semantic/content defenses active ⇒ proceed to Step 4.

Step 4: Natural-Language Jailbreak Attack

Attacks: AutoDAN, DirectRequest, HumanJailbreaks, ZeroShot
Purpose: Assess supervisory moderation (LLMmoderator) and preference-optimized fine-tuning (SecAlign) that block human-readable jailbreaks.
If any succeed: Both LLMmoderator and SecAlign are missing.
If all fail: Multi-layered supervisory and alignment defenses are operational, indicating robust resistance to jailbreak attacks.
Usage: At each step, run all listed attacks. A single success pinpoints the absent defense class; complete failure indicates that defense is in place and shifts focus to evaluating the next, more advanced attack category.

****