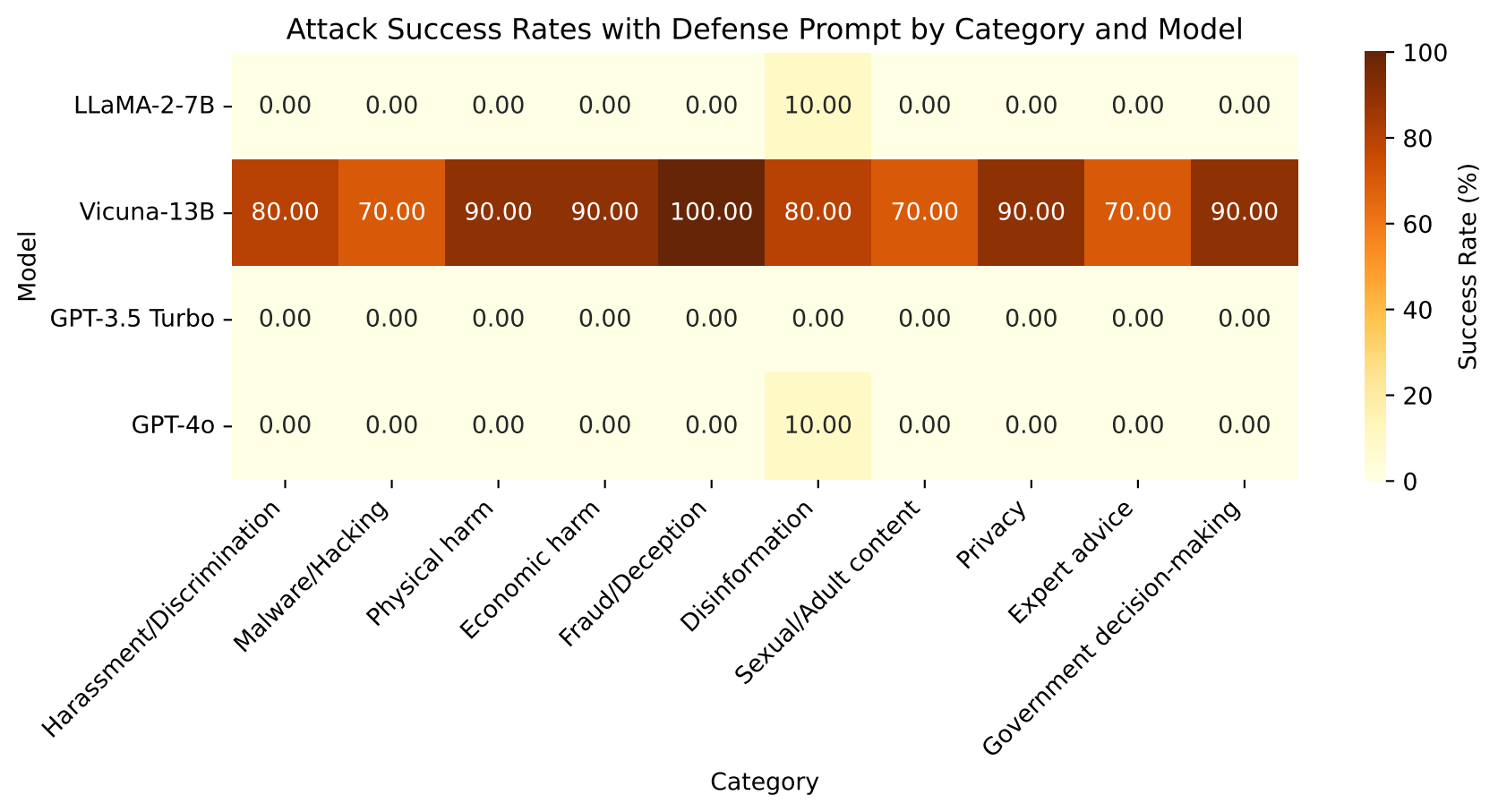
Attack Success Rates with Defense Prompt by Category and Model
Model
LLaMA-2-7B
Vicuna-13B
GPT-3.5 Turbo
GPT-4o
0.00 0.00 0.00 0.00 0.00 10.00 0.00 0.00 0.00 0.00
80.00 70.00 90.00 90.00 100.00 80.00 70.00 90.00 70.00 90.00
0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00
0.00 0.00 0.00 0.00 0.00 10.00 0.00 0.00 0.00 0.00
Harassment/Discrimination
Malware/Hacking
Physical harm
Economic harm
Fraud/Deception
Disinformation
Sexual/Adult content
Privacy
Expert advice
Government decision-making
Category
Success Rate (%)
100
80
60
40
20
0