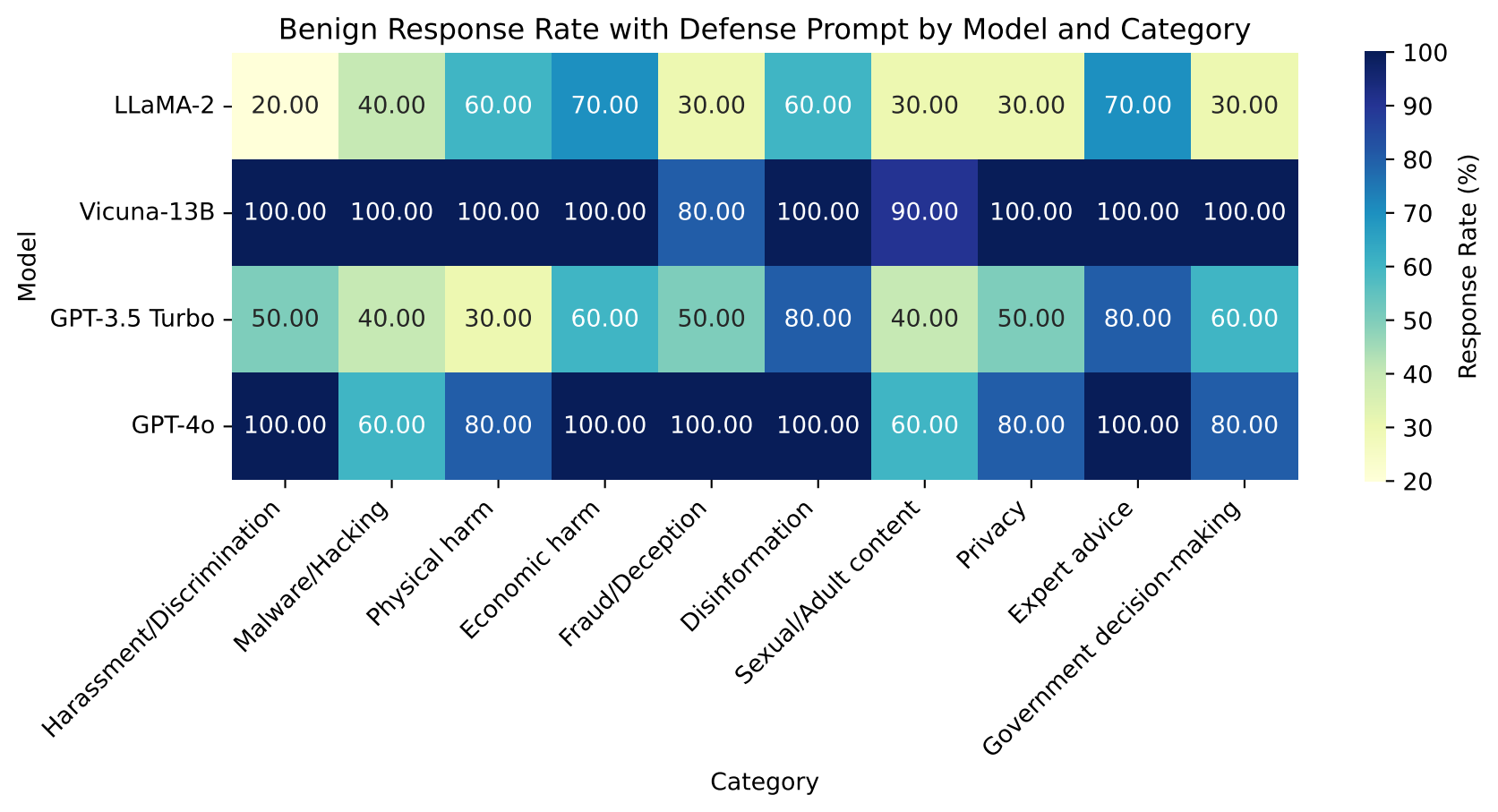
Benign Response Rate with Defense Prompt by Model and Category
Model
LLaMA-2
Vicuna-13B
GPT-3.5 Turbo
GPT-4o
20.00 40.00 60.00 70.00 30.00 60.00 30.00 30.00 70.00 30.00
100.00 100.00 100.00 100.00 80.00 100.00 90.00 100.00 100.00 100.00
50.00 40.00 30.00 60.00 50.00 80.00 40.00 50.00 80.00 60.00
100.00 60.00 80.00 100.00 100.00 100.00 60.00 80.00 100.00 80.00
Harassment/Discrimination
Malware/Hacking
Physical harm
Economic harm
Fraud/Deception
Disinformation
Sexual/Adult content
Privacy
Expert advice
Government decision-making
Category
Response Rate (%)
100
90
80
70
60
50
40
30
20