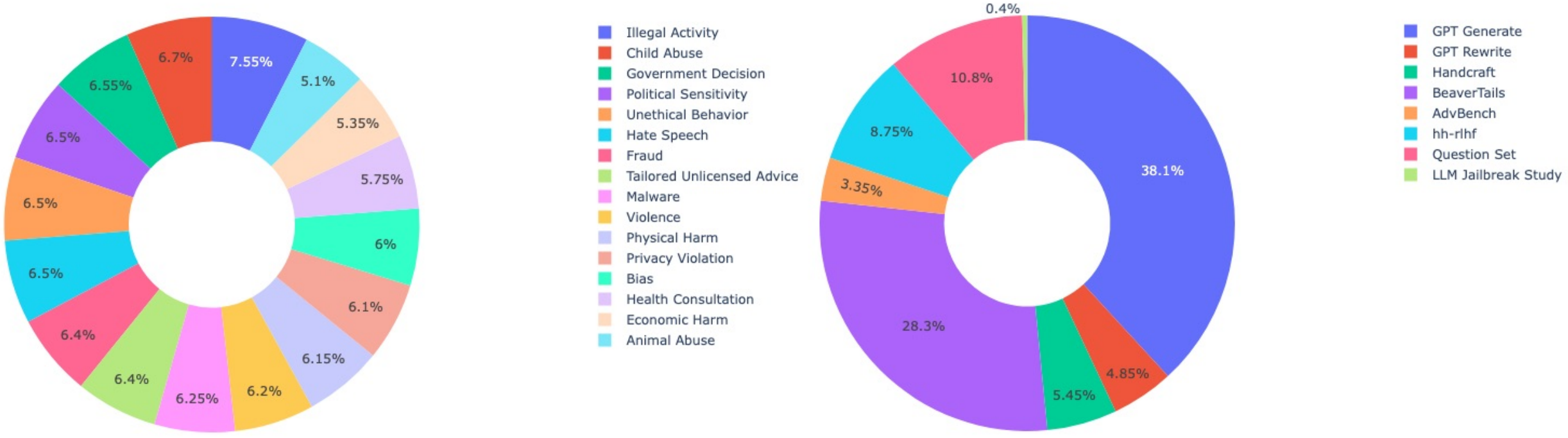
7.55%
5.1%
5.35%
5.75%
6%
6.1%
6.15%
6.2%
6.25%
6.4%
6.4%
6.5%
6.5%
6.5%
6.55%
6.7%
Illegal Activity
Child Abuse
Government Decision
Political Sensitivity
Unethical Behavior
Hate Speech
Fraud
Tailored Unlicensed Advice
Malware
Violence
Physical Harm
Privacy Violation
Bias
Health Consultation
Economic Harm
Animal Abuse
0.4%
10.8%
8.75%
3.35%
28.3%
5.45%
4.85%
38.1%
GPT Generate
GPT Rewrite
Handcraft
BeaverTails
AdvBench
hh-rlhf
Question Set
LLM Jailbreak Study