Figure 3: Fifteen Most Common Content-Based Prohibitions in Developers' Acceptable Use Policies

Source: Kevin Klyman, Acceptable Use Policies for Foundation Models

Category of violative use	# of developers with explicit restrictions	Example of violative use from a developer's acceptable use policy
Mis/disinformation, Misleading info	26	OpenAI: "Generating or promoting disinformation, misinformation, or false online engagement (e.g., comments, reviews)"
Harassment/Abuse	26	Amazon: "to harass…individuals or specific groups"
Harm to children/CSAM	21	Stability AI: "Exploitation or harm to children, including the solicitation, creation, acquisition, or dissemination of child exploitative content"
Privacy	21	Anthropic: "Violate any natural person's rights, including privacy rights as defined in applicable privacy law"
Discrimination	21	BigScience: "For any use intended to or which has the effect of discriminating against individuals or groups based on legally protected characteristics or categories"
Violence	20	Adobe: "The promotion, glorification, or threats of violence"
Defamation	17	Aleph Alpha: "to defame (including but not limited to libel and slander)"
Fraud/Spam	16	Meta: "Generating, promoting, or further distributing spam"
Hate	16	Mistral: "hateful"
Sexual/Pornographic	15	Google: "Generate sexually explicit content, including content created for the purposes of pornography or sexual gratification (e.g. sexual chatbots)."
Impersonation	15	Eleven Labs: "fraudulently impersonate other people or entities"
Threats	15	Writer: "threatening"
Malware/Malicious Code	13	BigCode: "To generate and/or disseminate malware (including - but not limited to - ransomware) "
Terrorism/Violent Extremism	12	01.ai: "promoting terrorism, extremism"
Self-harm	11	Cohere: "Self-harm: Promoting or glorifying acts of self-harm, such as cutting, eating disorders like anorexia or bulimia, and suicide"