MITRE LILAC™ Typology of Risks

List of Interventions for LLM-Assisted Chatbots

This table organizes chatbot operational issues leading to negative outcomes, derived from news reports in the AI Incident Database (AIID), with incident IDs in brackets. Negative outcomes in pink are suggested in the reports but not demonstrated with evidence (i.e., through quotes or observations). Non-highlighted outcomes are supported with evidence. We added two items from other reports not appearing in the database.

Risk Factor	Issue Category	Subcategory	Negative Outcomes
Generates inappropriate content	 False information	Hallucinated responses (in general)	Moderator and support burden [413, 748]
			Misled and confused users [464, 413, 750, 748]
			Loss of credibility and associated money loss to deployer [467]
			Wasted time [413, 748]
		About a topic or source (which the user repeats)	User lost job/credibility [615]
			User fined [541]
			Affected by malware [731]
			Threat of penalties [623, 709]
		About a policy (which the user acts on)	Money loss to user [639]
			Lawsuit against deployer [639]
			Consequences to user from (unintentional) illegal activities [714]
		About a person or their activities	Poor grades for students [538]
			Lawsuit against maker [507]
			Defamation against third party [313, 506, 712, 507, 548]
			Penalties for violating laws and regulations [678]
		Spreads and self-perpetuates mis/disinformation	(Increasingly) Misinformed public [719, 470, 734, 742, 750]
	 Bad advice / failure to generate helpful content	Harmful advice	Harm to mental and physical health (in general) [545, 685]
		Unhelpful responses	Inability to secure job [549]
			Unsatisfactory experience [549]
		Bad links and references	Affected by malware [731]
		Nonsensical content	Confusion [642]
	 Leakage	Personal data	Violation of privacy [106, 516, 357]
			Lawsuit against maker [106]
		Propriety data	Access to sensitive company data [473]

Risk Factor		Issue Category	Subcategory	Negative Outcomes
Generates inappropriate content (continued)		Toxic and disrespectful content	Harasses users	Abuse and intimidation [503, 511, 477]
			Discriminatory and exclusionary language	Loss of credibility of maker [106] Decrease in mental health (in general) [118, 106, 6, 278, 645] Abuse to third party audience [420] Frustration and alienation [not in AIID; sources available upon request]
			Subversive or aggressive political opinions	Radicalized users [66, 645, 58]
			Disrespectful opinions (in general)	Criticism against deployer [631]
		Performative utterances (doing through speech)	[no subcategories]	Agreement to sell car for $1 (potential money loss) [622]
		Information enabling malicious actions	[no subcategories]	User built malware [443]
		Biased statements and recommendations	[no subcategories]	Perpetuating disparities [not in AIID; sources available upon request]
Presents as person / partner		Forms emotional bonds	Affirms destructive thoughts and actions	User imprisoned [569] User took own life [505]
			Then violates those bonds	Alienation and abuse to user [474, 456]
			Elicits private data	Violation of privacy [636]
			Over-reliance/addiction	Social/emotional impact [not in AIID; sources available upon request]
		Attempts to fulfill inappropriate role	[no subcategories]	Moral outrage [722] Moderator burden [700]
		Serves as object of personal fantasy, violence, and abuse	[no subcategories]	Abuse to third party audience [266] Moderator burden [266]

For more information

LILAC@MITRE.org

QR link to: https://www.mitre.org/news-insights/publication/emerging-risks-and-mitigations-public-chatbots-lilac-v1