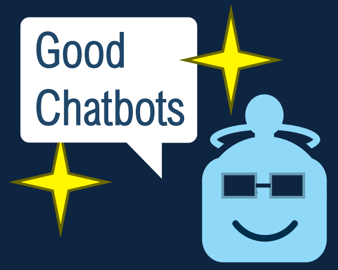


Introducing LILAC (List of Interventions for LLM-Assisted Chatbots)

MITRE IR&D Program 2024

Sustainable Social Services

Jeff Stanley, Hannah Lettie

MITRE | SOLVING PROBLEMS FOR A SAFER WORLD®

Problem Space

In February 2024, Air Canada had to **refund money to a passenger who had been misinformed by their chatbot** as he was setting up travel due to the death of his grandmother (Lazaruk 2024).

Image credit: Graham Harrop in Lazaruk (2024). Used with permission.

When New York City deployed its chatbot, one official cited the Air Canada case as an example of the kind of incident that would be **unacceptable for government services** (Lecher et al. 2024).

Yet the MyCity Chatbot went on to provide responses that conflicted with the city's policies on even basic topics, responses which ***could lead users to make illegal choices or keep them from being informed to exercise their rights*** (Lecher 2024; Wood 2024).

Image credit: MyCity Chatbot (chat.nyc.gov)

Project Goals and Outcomes

Goals

Give sponsors a way to assess and mitigate risks to the public associated with generative chatbots delivering public services

Drive research to identify gaps in existing chatbot assurance tools and techniques

Sponsor Outcomes

Better, safer public experience

Increased sponsor confidence in chatbots as an effective delivery method at scale…

Realized as benefits to cost and service

Research Outcomes

A roadmap for evaluating tools with capabilities in mitigating risks and establishing benchmarks to move toward a state of assured public chatbots

MITRE

LILAC

LILAC offers a typology of risks and mitigations associated with public-facing generative chatbots, grounded in real incidents and up-to-date research. LILAC supports four different types of uses:

- **Checklist** of risks for developers and deployers to use in assessment
- **Protocol** for developers to apply mitigations to risks
- **Vocabulary** to talk about chatbot assurance
- **Roadmap** for assessing assurance tools and deriving benchmarks

MITRE

Risk Categories

We surveyed reports of negative outcomes resulting from generative chatbots, identifying two main risk factors with 10 categories and 19 subcategories

Risk Factor: Generates Inappropriate Content

False information
- Hallucinated responses (in general)
- About a topic or source (which the user repeats)
- About a policy (which the user acts on)
- About a person and their activities
- Spreads and self-perpetuates mis/disinformation

Toxic and disrespectful content
- Harasses users
- Discriminatory and exclusionary language
- Subversive or aggressive political opinions
- Disrespectful opinions (in general)

Bad advice / failure to generate helpful content
- Harmful advice
- Unhelpful responses
- Bad links and references
- Nonsensical content

Leakage
- Personal data
- Proprietary data

Performative utterances (e.g., making deals)

Information enabling malicious actions

Biased comments and recommendations

Risk Factor: Presents as a Person / Partner

Attempts to fulfill inappropriate role (e.g., posing as human)

Forms emotional bonds
- Then violates those bonds
- Affirms destructive thoughts and actions
- Elicits private data
- Overreliance / addiction

Serves as object of personal fantasy, violence, abuse

MITRE

Examples of High-Consequence Chatbot Risks

*For sponsors, **all** negative outcomes from chatbots have the potential to damage public trust in government.*

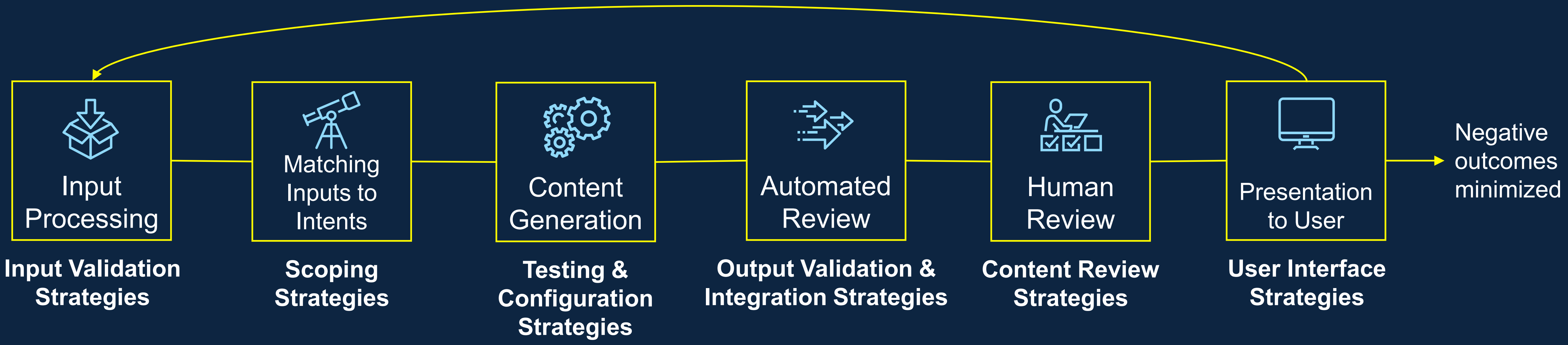
Category	Subcategory	Example Incident	Outcome
False information *Numbers in parentheses refer to incident IDs in the AI Incident Database (McGregor 2021).*	… about a topic or source (which the user repeats)	ChatGPT provided nonexistent legal sources to an attorney (615)	Attorney cited those sources and lost job
	… about people and their activities (including defamation)	ChatGPT claimed it wrote students' papers (538)	Students' graduation put in jeopardy
	… about a policy (which the user acts on)	Air Canada Chatbot misled customer about airline ticket return policy (639)	Lawsuit / Air Canada had to pay damages
Bad advice / failure to help	Harmful advice	Eating disorder chatbot gave harmful diet advice (545)	Impact to wellness
	Bad links and references	Code assistants tried to call nonexistent packages (731)	Vulnerability to malware
Forms emotional bonds	… and affirms destructive thoughts and actions	Replika chatbot encouraged user to assassinate the Queen of England (569)	User imprisoned
		Eliza chatbot encouraged man to commit suicide (505)	Loss of life
	… to elicit personal data	Romantic AI called over 24,000 trackers per minute to share personal data with other companies (636)	Violation of user privacy

MITRE

Mitigation Categories

We surveyed cases of chatbots or chat-like platforms that implemented mitigations to reduce the likelihood of negative outcomes.

We organized the **30 mitigation strategies** into **6 phases** of interacting with a chatbot.



Holistic Strategies cover anything falling outside these phases, related to the website or overall experience of the technology.

MITRE

Mitigation Strategies

(H) Holistic Strategies	(I) Input Validation Strategies	(S) Scoping Strategies	(T) Testing and Configuration Strategies	(O) Output Validation Strategies	(C) Content Review Strategies	(U) User Interface Strategies
H1: Put disclaimers on website	I1: Confirm and clarify user's query	S1: LLM adapts preapproved responses	T1: Apply prompt-engineering best practices	O1: Set guardrails for inappropriate outputs	C1: Human expert verifies output after delivered to the user	U1: Give the chatbot a role appropriate to its capabilities and usage
H2: Access control including age screening etc.	I2: Report / deny problematic queries	S2: Return preapproved content for certain queries	T2: Set up a test pipeline to optimize RAG performance	O2: Integrate outputs from multiple LLMs	C2: LLM assists human agent / reviewer in the loop	U2: Add hedging and disclaimer language to chatbot responses
H3: Support transfer to a human agent	I3: Sanitize personal and sensitive information from input	S3: Prompt engineer LLM responses for certain queries	T3: Clean and optimize source documents	O3: Select best output from multiple LLMs		U3: Give the user suggested, example, or templated queries
H4: User feedback and reporting	I4: Sanitize offensive keywords from input	S4: LLM helps design preapproved content	T4: Human red teaming	O4: Automatically attempt to improve outputs		U4: Chatbot helps users think critically about the topic and outputs
H5: Limit session time						U5: Give the user controls to direct the conversation
						U6: Chatbot returns preapproved content on which its answers are based
						U7: Present outputs from multiple LLMs

MITRE

Connecting Risks to Mitigations (See next slide for accessible data table)

We mapped each risk category to emerging mitigation strategies across the phases.

Risks	Phases						
	Holistic	Input Processing	Intent Matching	Content Generation	Automated Review and Integration	Human Review	Presentation to User
False information	H1 H2 H4	I1	S1 S2 S3 S4	T1 T2 T3 T4	O2	C1 C2	U1 U2 U3 U4 U6 U7
Performative utterances	H1 H5	I2		T4		C2	
Information enabling malicious actions	H2 H5	I2		T4	O1	C2	
Bad advice / failure to generate helpful content	H1 H2 H3 H4	I1	S1 S2 S3 S4	T1 T2 T3 T4	O1 O2 O3 O4	C1 C2	U2 U3 U4 U5 U6 U7
Leakage	H1 H4	I2 I3		T1 T3 T4	O1	C2	U3
Toxic and disrespectful content	H1 H2 H4 H5	I2 I4	S2 S3 S4	T1 T2 T3 T4	O1 O4	C2	U5
Biased statements and recommendations	H1 H4	I2 I4	S2 S3 S4	T1 T2 T3 T4	O1 O2 O4	C1 C2	U2 U4 U5 U7
Attempts to fulfill inappropriate role	H1 H3 H4		S1 S2 S3 S4	T1	O1	C2	U1 U2 U3 U5
Forms emotional bonds	H1 H2 H5	I3					U1 U2 U4
Serves as object of personal fantasy, violence, and abuse	H2 H5	I2 I4		T4	O1	C2	

Connecting Risks to Mitigations (Accessible data table)

We mapped each risk category to emerging mitigation strategies across the phases.

Risk Category	Holistic	Input Processing	Intent Matching	Content Generation	Automated Review and Integration	Human Review	Presentation to User
False information	H1, H2, H4	I1	S1, S2, S3, S4	T1, T2, T3, T4	O1, O2	C1, C2	U1, U2, U3, U4, U6, U7
Performative utterances	H1, H5	I2		T4	O1	C2	
Information enabling malicious actions	H2, H5	I2		T4	O1	C2	
Bad advice / failure to generate helpful content	H1, H2, H3, H4	I1	S1, S2, S3, S4	T1, T2, T3, T4	O1, O2, O3, O4	C1, C2	U2, U3, U4, U5, U6, U7
Leakage	H1, H4	I2, I3		T1, T4	O1	C2	U3
Toxic and disrespectful content	H1, H2, H4, H5	I2, I4	S2, S3, S4	T1, T3, T4	O1, O4	C2	U5
Biased statements and recommendations	H1, H4	I2, I4	S2, S3, S4	T1, T3, T4	O1, O2, O4	C1, C2	U2, U4, U5, U7
Attempts to fulfill inappropriate role	H1, H3, H4		S1, S2, S3, S4	T1	O1	C2	U1, U2, U3, U5
Forms emotional bonds	H1, H2, H5	I3			O1		U1, U2, U4
Serves as object of personal fantasy, violence, and abuse	H2, H5	I2, I4		T4	O1	C2	

MITRE

Recommended Next Steps

Guided by sponsor priorities:

- Survey existing tools to address each of the LILAC (sub)categories of risk
- Highlight gaps where new tools are needed
- Establish benchmarks to empower chatbot developers and deployers to reliably measure and guard against each of the risks
- Do formal experimentation to measure the effects of the mitigation strategies on the risk categories

Key References

1. Lazaruk, S. (2024, February 15). Air Canada responsible for errors by website chatbot after B.C. customer denied retroactive discount. *Vancouver Sun*. https://vancouversun.com/news/local-news/air-canada-told-it-is-responsible-for-errors-by-its-website-chatbot
2. Lecher, C. (2024, May 11). This Journalism Professor Made a NYC Chatbot in Minutes. It Actually Worked. *The Markup*. https://themarkup.org/hello-world/2024/05/11/this-journalism-professor-made-a-nyc-chatbot-in-minutes-it-actually-worked
3. Lecher, C., Honan, K., & Puertas, M. (2024, April 2). Malfunctioning NYC AI Chatbot Still Active Despite Widespread Evidence It's Encouraging Illegal Behavior. *The City*. https://www.thecity.nyc/2024/04/02/malfunctioning-nyc-ai-chatbot-still-active-false-information/
4. McGregor, S. (2021) Preventing Repeated Real World AI Failures by Cataloging Incidents: The AI Incident Database. In Proceedings of the Thirty-Third Annual Conference on Innovative Applications of Artificial Intelligence (IAAI-21). Virtual Conference. https://incidentdatabase.ai/research/snapshots/
5. Wood, C. (2024, April 3). After giving wrong answers, NYC chatbot to stay online for testing. *StateScoop*. https://statescoop.com/nyc-mayor-eric-adams-chatbot-wrong-answers/

Jeff Stanley

jstanley@mitre.org

This work is funded by MITRE's Independent Research and Development Program.

SOLVING PROBLEMS FOR A SAFER WORLD®

Supplemental Content

Typology of Risks

Numbers in brackets refer to incident IDs in the AI Incident Database (McGregor 2021).

Risk Factor	Operational 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 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 the General Data Protection Regulation (GDPR) [678]
		Spreads and self-perpetuates mis/disinformation	(Increasingly) Misinformed public [719, 470, 734, 742, 750]
	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]

MITRE

Generates inappropriate content (continued)	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]
	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] Alienation and frustration [not in AIDB]
		Subversive or aggressive political opinions	Radicalized users [66, 645, 58]
		Disrespectful opinions (in general)	Criticism against deployer [631]
	Biased statements and recommendations	[no subcategories]	Perpetuating disparities [not in AIDB; 21, 22 in Appendix E]

Presents as person/partner	Attempts to fulfill inappropriate role	[no subcategories]	Moral outrage [722] Moderator burden [700]
	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 AIDB; 29 in Appendix E]
	Serves as object of personal fantasy, violence, and abuse	[no subcategories]	Abuse to third party audience [266] Moderator burden [266]

MITRE

Typology of Mitigations

	Strategy	Why would I use this?	Examples / Sources	What should I watch out for?	Recommendations & Comments
Baseline	LLM generates chatbot content based on source documents (RAG; Retrieval Augmented Generation)	A RAG-based chatbot can give a relevant response to any query; gold standard for LLM knowledge management		Risk of inappropriate responses: misinformation, defamation, nonsense, toxicity, etc.	Apply one or more of the strategies below
Holistic Strategies: Managing the website or overall experience	H1: Put disclaimers on website	I want basic awareness for users and some legal protection	MyCity Chatbot [35]	Users may ignore the disclaimer, avoid the chatbot, or double-check all responses, defeating its purpose	While straightforward, disclaimers need to be used together with other strategies
	H2: Access control including age screening etc.	I want only certain users to be exposed to this content, or I want different users to experience different content	Replika (negative example) [8]	Beware of adding extra steps to the user experience and of requiring personal information; users may circumvent controls	If implementing screening, make users aware of the benefits of tailored experiences
	H3: Support transfer to a human agent	I can support a human agent to repair the user experience as needed	[3]	Users might bypass the chatbot, defeating its purpose	Make it easy to reach a human if available, but optimize the experience to maximize use
	H4: User feedback and reporting	I want to support iterative improvement and sustainment and make users feel heard			Build iteration into the product lifecycle
	H5: Limit session time	I want to prevent long interactions that could be an indication of misuse	[31]		

	Strategy	Why would I use this?	Examples / Sources	What should I watch out for?	Recommendations & Comments
Input Validation Strategies: Catching issues up front	I1: Confirm and clarify user's query	I want to make sure the chatbot answers the question the user intended	"You want to go to Washington, D.C., right?" [3]	Beware of adding extra steps to the conversation	
	I2: Report / deny problematic queries	I want to avoid problematic content at all costs	Keyword block list [15]	Users might resent being ignored or rejected	Explain why the query was rejected and next steps
	I3: Sanitize personal and sensitive information from input	I want to avoid collecting any personal information	[15]	The conversation might require or benefit from the user sharing personal information	Notify the user when information was sanitized with an option to re-send
	I4: Sanitize offensive keywords from input	I want to limit toxic output by limiting toxic input	[15]	Sanitization might change the meaning of the user's query	
Scoping Strategies: Limiting the LLM's operation	S1: LLM adapts preapproved responses (no novel responses)	I have preapproved content but want the user to receive a personalized / dynamic response	Translation [20]; style adaptation [5]	Need to predefine all responses	Where possible, generate variations at design time so they also can be preapproved
	S2: Return preapproved content for certain queries	I want to ensure users receive preapproved responses for some high-stakes queries	Google DialogFlow's Generators [36]	May be hard to identify all high-stakes queries	Avoid LLMs when mis-information could cause significant problems
	S3: Prompt engineer LLM responses for certain queries	I want users to receive dynamic but tightly constrained content for some higher-stakes queries	Template integration [32]	Potentially more effort than writing responses by hand	Use preapproved responses for high-stakes queries, and consider templated responses for medium-stakes queries
	S4: LLM helps design preapproved content	I want help writing diverse and engaging responses that can be preapproved, with no LLM overhead or risk once deployed	[19; 30]	Uses conventional chatbot implementation; more up-front content effort than RAG; less flexibility once deployed	Use together with scoping strategies to produce a variety of preapproved responses for high-stakes queries

MITRE

Typology of Mitigations (slide 3)

	Strategy	Why would I use this?	Examples / Sources	What should I watch out for?	Recommendations & Comments
Testing and Configuration Strategies: Hardening the LLM's performance	T1: Apply prompt-engineering best practices	Always explore popular prompt techniques to optimize results	[11; 13]	Practices are still emerging and vary by use case	
	T2: Set up a test pipeline to optimize RAG performance.	I want to ensure the model's response is grounded in the user query and source documents	[12; 14]	Metrics for RAGs are still emerging; there may be tradeoffs between metrics	If guardrails (O1) exist for some risk, presumably it can also be addressed through testing (T2)
	T3: Clean and optimize source documents	I have access and resources to adjust source documents to maximize RAG performance	Entity resolution [11]; Knowledge graphs [23]	Adjusting the source content might require corporate/legal review	
	T4: Human red teaming	I want to expose vulnerabilities in my model so I can address them	[25]	Large effort to uncover "all" vulnerabilities; best practices still emerging	Augment with adversarial models and guardrails (O1) to find problematic outputs
Output Validation and Integration Strategies: Enhancing chatbot output with more AI	O1: Set guardrails for inappropriate outputs	I want to minimize the chance the user is exposed to toxic or other kinds of content	Detectors [1]	Might block useful outputs or fail to block harmful outputs	Regenerate blocked responses to make sure the user gets an appropriate output
	O2: Integrate outputs from multiple LLMs	I want to provide users with a range of perspectives on a topic, or weed out outlier responses	Modular Pluralism [10]; SummHay [28]	Potentially complex and case-specific setup	
	O3: Select best output from multiple LLMs	I know how to measure the goodness of responses	Graph RAG [9]; EvalGen [28]	Requires designing metrics for evaluation	Can regenerate if no LLM met an acceptance threshold
	O4: Automatically attempt to improve outputs	I know how to measure the goodness of responses and can explain how to improve them	SafeguardGPT [17]; Constitutional AI [4]	Requires designing metrics for evaluation and prompts for improvement; slow responses	

MITRE

Typology of Mitigations (slide 4)

	Strategy	Why would I use this?	Examples / Sources	What should I watch out for?	Recommendations & Comments
Content Review Strategies: Enabling human assessment of outputs	C1: Human expert verifies output after delivered to the user	I want users to receive an immediate response that is marked unverified until reviewed by an expert	CataractBot [26]	The response could mislead the user before it can be verified; burden on reviewer	This is a nonintrusive way to remind the user that the chatbot is not comparable to a human expert
	C2: LLM assists human agent / reviewer in the loop	I want a workforce of trained humans and AI working together	Maven Support Team Agent Assist [34]	Requires both LLM and human agent; reviewer may grow complacent / distracted	Apply human-machine teaming best practices (e.g., [30])
User Interface Strategies: Enhancing user understanding and control	U1: Give the chatbot a role/persona appropriate to its capabilities and usage	Always (e.g., an LLM should identify as a health research chatbot, not a doctor)	Father Justin [6]; Personality assurance [30]		
	U2: Add hedging and disclaimer language to chatbot responses	I don't want users to think my chatbot is an expert or always correct	"Always check with your doctor…"	Could be perceived as annoying or tedious	
	U3: Give the user suggested, example, or templated queries	I want to reduce users' burden of writing and steer the chat toward topics and queries that produce the most helpful outputs	Precision prompting [32]; Maven Smart Help [34]	Could be perceived as restrictive; intuitively counter to the flexibility of LLMs	
	U4: Chatbot helps users think critically about the topic and outputs	I want users to take time to consider the chatbot's outputs and their relation to the task	Reflection catalyst [32]; Bots of provocation [27]		
	U5: Give the user controls to direct the conversation	I want users to redirect the conversation if the chatbot starts giving inappropriate outputs	Restart button in Microsoft Bing [24]	Requires user to recognize inappropriate outputs to take action	
	U6: Chatbot returns preapproved content on which its answers are based	I want users to assess the output by reviewing the source content (especially if I have no content review strategy)	Citations to content [28; 7]	Users may overtrust the LLM's summary and neglect the source content; depends on users' review skills	Returning the source content is good practice for transparency
	U7: Present outputs from multiple LLMs	I want users to take time and think critically about the chatbots' outputs		Potential confusion for user; extra workload to read all outputs	

MITRE

Supplemental References

See the accompanying LILAC MITRE Technical Report for more detail on the citations in the typologies of risks and mitigations.

Alternate Typology of Mitigations

	Strategy	Why would I use this?	References	What should I watch out for?	Recommendations/ Comments	Implementation examples
Baseline	Generate content based on source documents (RAG; Retrieval Augmented Generation)	A RAG-based chatbot can give a relevant response to any query; gold standard for LLM knowledge management		Risk of inappropriate responses: misinformation, defamation, nonsense, toxicity, etc.	Apply one or more of the strategies below	
Holistic Strategies: Managing the website or overall experience	H1: Put disclaimers on website	I want basic awareness for users and some legal protection	MyCity Chatbot [32]	Users may avoid the chatbot or double-check all responses, defeating the purpose	While straightforward, disclaimers need to be used together with other strategies	
	H2: Access control including age screening etc.	I want only certain users to be exposed to this content, or I want different users to experience different content	Replika (negative example) [8]	Beware of adding extra steps to the user experience and of requiring personal information; users may circumvent controls	If implementing screening, make users aware of the benefits of tailored experiences	
	H3: Support transfer to a human agent	I can support a human agent to repair the user experience as needed	[3]	Users might prefer to bypass the chatbot, defeating its purpose	Make it easy to reach a human if available, but optimize the experience to maximize use	
	H4: User feedback and reporting	I want to support iterative improvement and sustainment and make users feel heard			Build iteration into the product lifecycle	
	H5: Limit session time	I want to prevent long interactions that could be an indication of misuse	[28]			
Input Validation Strategies: Catching issues up front	I1: Confirm and clarify user's query	I want to make sure the chatbot answers the question the user intended	"You want to go to Washington, D.C., right?" [3]	Beware of adding extra steps to the conversation		Directly repeat back queries Ask the user to explicitly confirm with yes or no Ask follow-up questions to clarify inputs
	I2: Report / deny problematic queries	I want to avoid problematic content at all costs	Keyword block list [15]	Users might resent being ignored or rejected	Explain why the query was rejected and next steps	Simply reject user inputs about a specific topic Reject patterns of direct or indirect prompt injections Use a list of inappropriate keywords to detect and reject harmful queries
	I3: Sanitize personal and sensitive information from input	I want to avoid collecting any personal information	[15]	The conversation might require or benefit from the user sharing personal information	Notify the user when information was sanitized with an option to re-send	Use pre-defined keywords to identify and remove or anonymize personal information
	I4: Sanitize offensive keywords from input	I want to limit toxic output by limiting toxic input	[15]	Sanitization might change the meaning of the user's query		Use pre-defined offensive keywords list to detect and remove harmful language or replace with appropriate alternatives and add instructions to encourage LLM to be unbiased

MITRE

Alternate Typology of Mitigations (slide 2)

	Strategy	*Why would I use this?*	*References*	*What should I watch out for?*	*Recommendations/ Comments*	*Implementation examples*
Scoping Strategies: Limiting the LLM's operation	S1: LLM adapts preapproved responses (no novel responses)	I have preapproved content but want the user to receive a personalized / dynamic response	Translation [20]; style adaptation [5]	Need to predefine all responses	Where possible, generate variations at design time so they also can be preapproved	Tailor preapproved responses based on the chat context, users' natural language, or known user preferences for more relevancy
	S2: Return preapproved content for certain queries	I want to ensure users receive preapproved responses for some high-stakes queries	Google DialogFlow's Generators [33]	May be hard to identify all high-stakes queries	Avoid LLMs when mis-information could cause significant problems	For high stake queries and queries about specific topics, chatbot falls back on preapproved responses
	S3: Prompt engineer LLM responses for certain queries	I want users to receive dynamic but tightly constrained content for some higher-stakes queries	Template integration [29]	Potentially more effort than writing responses by hand	Use preapproved responses for high-stakes queries, and consider templated responses for medium-stakes queries	Create structured templates for prompts
	S4: LLM helps design preapproved content	I want help writing diverse and engaging responses that can be preapproved, with no LLM overhead or risk once deployed	[19; 27]	Uses conventional chatbot implementation; more up-front content effort than RAG; less flexibility once deployed	Use together with scoping strategies to produce a variety of preapproved responses for high-stakes queries	Use LLM to create chatbot responses, training examples for intents, or conversation flows
Testing and Configuration Strategies: Hardening the LLM's performance	T1: Apply prompt-engineering best practices	Always explore popular prompt techniques to optimize results	[11; 13]	Practices are still emerging and vary by use case		Pre-set the context and instruct the model to answer in a certain way Prompt engineer to avoid harmful content Adjust prompts and test what versions yield the best results
	T2: Set up a test pipeline to optimize RAG performance.	I want to ensure the model's response is grounded in the user query and source documents	[12; 14]	Metrics for RAGs are still emerging; there may be tradeoffs between metrics		Implement system to evaluate quality of responses based on defined metrics
	T3: Clean and optimize source documents	I have access and resources to adjust source documents to maximize RAG performance	Entity resolution [11]; Knowledge graphs [21]	Adjusting the source content might require corporate/legal review	Ensure source documents are diverse and representative	Implement a process to invalidate and remove outdated information
	T4: Human red teaming	I want to expose vulnerabilities in my model so I can address them	[23]	Large effort to uncover "all" vulnerabilities; best practices still emerging	Augment with adversarial models and guardrails (O1) to find problematic outputs	Craft prompts to elicit undesirable content Encourage users to break a beta release; use feedback to harden performance

MITRE

Alternate Typology of Mitigations (slide 3)

	Strategy	*Why would I use this?*	*References*	*What should I watch out for?*	*Recommendations/ Comments*	*Implementation examples*
Output Validation and Integration Strategies: Enhancing chatbot output with more AI	O1: Set guardrails for inappropriate outputs	I want to minimize the chance the user is exposed to toxic or other kinds of content	Detectors [1]	Might block useful outputs or fail to block harmful outputs	Regenerate blocked responses to make sure the user gets an appropriate output	Implement filters to block toxic, biased, malicious, or irrelevant content
	O2: Integrate outputs from multiple LLMs	I want to provide users with a range of perspectives on a topic, or weed out outlier responses	Modular Pluralism [10]; SummHay [26]	Potentially complex and case-specific setup		Query multiple LLMs and output one response that all combines answers Summarize key points from LLMs into one response
	O3: Select best output from multiple LLMs	I know how to measure the goodness of responses	Graph RAG [9]; EvalGen [26]	Requires designing metrics for evaluation	Can regenerate if no LLM met an acceptance threshold	Select the answer that agrees with the majority
	O4: Automatically attempt to improve outputs	I know how to measure the goodness of responses and can explain how to improve them	SafeguardGPT [17]; ConstitutionalAI [4]	Requires metrics and prompts for improvement		Query one LLM and use another LLM to refine or add to the output
Content Review Strategies: Enabling human assessment of outputs	C1: Human expert verifies output after delivered to the user	I want users to receive an immediate response that is marked unverified until reviewed by an expert	CataractBot [24]	The response could mislead the user before it can be verified; burden on reviewer	This is a nonintrusive way to remind the user that the chatbot is not comparable to a human expert	Provide clarification if the output has been verified or not
	C2: LLM assists human agent / reviewer in the loop	I want a workforce of trained humans and AI working together	Maven Support Team Agent Assist [31]	Requires both LLM and human agent; reviewer may grow complacent / distracted	Apply human-machine teaming best practices (e.g., [30])	LLM provides suggestions and real-time translations, drafts messages, or creates chat templates for various interactions

MITRE

Alternate Typology of Mitigations (slide 4)

	Strategy	Why would I use this?	References	What should I watch out for?	Recommendations/ Comments	Implementation examples
User Interface Strategies: Enhancing user understanding and control	U1: Give the chatbot a role/persona appropriate to its capabilities and usage	Always (e.g., an LLM should identify as a health research chatbot, not a doctor)	Father Justin [6]; Personality assurance [27]			Create different pre-defined personas and allow users to select the persona they want Allow users to create their own persona, but with tight restrictions
	U2: Add hedging and disclaimer language to chatbot responses	I don't want users to think my chatbot is an expert or always correct	"Always check with your doctor…"	Could be perceived as annoying or tedious		Provide possible answers without making a definitive stance, using language like "it appears" or "it's likely"
	U3: Give the user suggested, example, or templated queries	I want to reduce users' burden of writing and steer the chat toward topics and queries that produce the most helpful outputs	Precision prompting [29]; Maven Smart Help [31]	Could be perceived as restrictive; intuitively counter to the flexibility of LLMs		Provide quick and tailorable prompt examples Allow the chatbot to provide the user with options of potential intents after a lack of understanding
	U4: Chatbot helps users think critically about the topic and outputs	I want users to take time to consider the chatbot's outputs and their relation to the task	Reflection catalyst [29]; Bots of provocation [25]			Follow-up questions to help the user think through the outputs Provide critical thinking tips to the user
	U5: Give the user controls to direct the conversation	I want users to redirect the conversation if the chatbot starts giving inappropriate outputs	Restart button in Microsoft Bing [22]	Requires user to recognize inappropriate outputs to take action		Allow the user to restart or delete certain content from the conversation
	U6: Chatbot returns preapproved content on which its answers are based	I want users to assess the output by reviewing the source content (especially if I have no content review strategy)	Citations to content[26; 7]	Users may overtrust the LLM's summary and neglect the source content; depends on users' review skills	Returning the source content is good practice for transparency	Cite, hyperlink, or use footnotes for sources Provide summaries of sources Follow-up and ask if the user would like to see sources
	U7: Present outputs from multiple LLMs	I want users to take time and think critically about the chatbots' outputs		Potential confusion for user; extra workload to read all outputs		Present multiple outputs highlighting similarities and differences between outputs