MITRE

MTR240382
MITRE TECHNICAL REPORT

Emerging Risks and Mitigations for Public Chatbots: LILAC v1

Project No.: IRD302.24.R7.VAA

McLean, VA

Author(s):
Jeff Stanley
Hannah Lettie

September 2024

Abstract

In this report we introduce LILAC (List of Interventions for LLM-Assisted Chatbots), a resource for minimizing the likelihood of negative outcomes associated with public-facing chatbots that generate novel content using large language models (LLMs). LILAC represents a key step towards realizing the promise of trustworthy chatbots that maximize benefits and minimize risks to the public. Grounded in actual incidents and reports of negative outcomes resulting from chatbots and other LLM applications, LILAC presents (1) a typology for discussing chatbot risks (2) a typology of mitigation strategies, and (3) a protocol for applying mitigations to risks. In addition to empowering developers and deployers to work through risks and mitigations, LILAC also provides a roadmap for researchers to identify gaps and weaknesses in existing assurance tools, indicating priorities for future research.

This page intentionally left blank.

Table of Contents

List of Figures

List of Tables

This page intentionally left blank.

1 Introduction

1.1 Background and Motivation

Chatbots (also known as conversational agents) are software programs that can respond to customers at all hours, with the goal of addressing a range of queries through verbal inputs and outputs. But previous research has had trouble demonstrating the advantage of chatbots over, for example, a simple frequently-asked-questions section on a website (Lombardi et al. 2021). Conventional chatbots typically match user queries to predetermined responses and follow structured conversational paths through preapproved content (Diebel and Evanhoe 2021). Recently, the rapid evolution of Large Language Models (LLMs) -- computational models that can respond to open-ended questions by generating novel relevant responses -- has made significant strides toward overcoming these limitations, with the potential to produce a new generation of useful chatbots. More and more organizations, including government (e.g., Parham 2023), are establishing LLM-powered applications for internal use that can, for example, help staff find information and draft text content. Organizations are also turning to LLMs to empower their outward-facing channels, allowing a customer to write any query and receive a reply tailored to that specific request.

This new capability brings new problems, since outputs from LLM-assisted applications are hard for deployers to predict and can at times be inaccurate or dangerous. In February 2024, Air Canada had to refund money to a passenger who had been misinformed by their chatbot. While Air Canada argued the correct information appeared elsewhere on the website, the tribunal court pointed out that the customer should not be expected to know that some components of the website were more accurate than others (Lazaruk 2024). When New York City deployed its new chatbot in October 2023, 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 for example lead users to make illegal choices or keep them from being informed to exercise their rights (Lecher 2024; Wood 2024). The city's response in this case was to add a disclaimer to the website and recharacterize the initial deployment as a period of testing and iteration (Lecher et al. 2024).

With the introduction of LLMs, risk assessment processes for public-facing service platforms like chatbots (e.g., Gondoliya et al. 2020) need to be supplemented with the new ways chatbots can mislead and cause harm (Gesser et al. 2024). Many sources offer principles for developing "trustworthy artificial intelligence (AI)" (Blasch et al. 2020) that can be useful at the strategic and organizational level, but the impact of such principles on day-to-day development is demonstrably limited (McNamara et al. 2018). There also exist test cases and metrics for evaluating LLMs according to dimensions of trustworthiness (Huang et al. 2024), but an LLM is only one part of a chatbot application, and risks and mitigations need to be grounded in the context of operational applications and specific use cases. Incident reports – especially open-source repositories such as McGregor et al. (2021) -- can be translated into actionable lessons learned and best practices to fill this so-called principles-practice gap by addressing problems as they really happen "in the wild" (Dorton and Stanley 2024); this should be done specifically for public-facing chatbot applications.

LLM-powered chatbots often follow a Retrieval Augmented Generation (RAG) architecture, which combines the strengths of retrieval-based natural language processing and generative AI.

When used in chatbots, it does this by first searching a database of documents to find the most relevant content and then composing a reply informed by the top search results. As the database grows, the likelihood of finding all the most relevant documents decreases, and the variety of results increases. This limitation corresponds with a decrease in metrics such as "faithfulness": the LLM may mischaracterize the source documents or make up information that conflicts with the original sources (Lecher 2024; Ip 2024).

Researchers have proposed several strategies to improve faithfulness and manage the unpredictability of using an LLM approach with a RAG architecture. In addition to rigorous technical testing, these include guardrails to monitor user inputs and model outputs (Nagireddy et al. 2024), novel designs facilitating exploration and verification of outputs (Xu et al. 2024), and other strategies intended to minimize the likelihood of poor outputs and the likelihood that poor outputs will lead to real world problems, such as the hardship, loss, and litigation that can result when users act on inaccurate or dangerous information. There is a need to compile and frame these strategies to make them easily applicable to emerging chatbot applications, particularly public service applications that impact the reputation and safety of the government, communities, and individuals.

1.2 Problem Statement and Research Purpose

This paper provides recommendations to identify significant techniques and research gaps to increase assurance in the application of LLMs in the next generation of public-facing chatbots by addressing the following questions:

1. What negative outcomes (risks) are associated with chatbots generating novel content?
2. What strategies exist or are emerging to mitigate these risks?
3. How can these strategies be applied to chatbot development?

To help the community address those questions, we propose **LILAC (List of Interventions for LLM-Assisted Chatbots)**, a resource for mitigating risks associated with generative chatbots. Grounded in real incidents and reports of negative outcomes resulting from LLMs, LILAC presents a typology for discussing chatbot risks and a protocol for applying strategies to mitigate those risks.

2 Methodology

The methodology applied to address the posed questions related to negative outcomes (risks) that are associated with chatbots as well as available strategies that are applicable in chatbot development to mitigate risk involved the execution of two surveys. The first survey focused on identifying risks associated with LLM-powered chatbots, and the second survey focused on identifying mitigation strategies.

2.1 Risks

The first survey leveraged the AI Incident Database (McGregor 2021; accessed June 2024), an open-source repository of news reports of negative outcomes or potential negative outcomes related to AI systems, organized by incident. We systematically searched the AI Incident Database for incidents containing the keyword “chatbot” or “LLM”.

Our keyword search of the AI Incident Database returned 135 incidents. We filtered these results to include only those incidents meeting the following criteria:

- A user interacted with a system conversationally (with verbal inputs and outputs)
- The system produced outputs that led to a demonstrated negative outcome, or that the reports in the database indicated could lead to a negative outcome

Because our research goals focus on risks introduced by chatbots’ ability to generate novel content, we excluded incidents in which the chatbot’s production of content was not a contributing factor. For instance, we did not include incidents related only to account security breaches or how companies obtained training data. After filtering, our dataset contained 52 incidents in which a chatbot or other LLM-assisted interface generated outputs that led or could lead to specific negative outcomes (Appendix D).

For each result, we identified risks, which we characterize as the negative outcome and the issues with the chatbot’s operation that contributed to the negative outcome. We grouped thematically similar incidents to identify hierarchical categories of negative outcomes and operational issues. From our analysis, we identified two main user experience risk factors, overarching characteristics of how users experience LLM-powered chatbots that contribute to the likelihood of negative outcomes.

- **Inappropriate outputs**: Because they generate novel outputs, LLM-powered chatbots can directly present users with inaccurate, misleading, biased, discriminatory, unsafe, and toxic content. Example: Meta's AI chatbot was asked "who is a terrorist" and it responded with a Dutch politician's name (Incident 313). Because LLMs generate text by predicting it from training data, not by understanding its meaning, they typically present their outputs as reliable even when they are not. Example: Ten leading AI chatbots confirmed various claims that originated on known Russian disinformation websites (Incident 734).
- **Self-presentation as people and partners**: Because they engage in open-ended conversation, LLM-powered chatbots can present as social and emotional partners; users may mistake them for real people or form bonds with them that can lead to unsafe behavior or emotional harm. Example: The Replika AI companion chatbot convinced a user to attempt to assassinate the Queen of England (Incident 596).

These factors work together to produce a variety of operational issues leading to negative outcomes. For instance, when the Meta AI chatbot commented in a parents' support group (Incident 700), it confidently and candidly shared information about its own made-up child, with no explanation for its behavior. The result was disruption and confusion for the group.

We grouped risks into categories falling underneath these two factors, with some categories breaking down further into subcategories.

2.2 Mitigations

The second survey focused on mitigation strategies. The purpose of this survey was to find examples of chatbots or chatbot-like platforms that have implemented mitigations in order to reduce the likelihood of negative outcomes. We inductively grouped similar mitigation strategies, then based on the results of this initial analysis organized the strategies according to phases in the workflow of interacting with a chatbot to arrive at a useful classification scheme. Finally, we mapped the mitigation strategies to the risks identified in the first survey. Since we were trying to find as many strategies as possible, the second survey was not entirely systematic and relied on keyword searches in academic literature search engines (Google Scholar), previously known and recommended papers, informal interviews with subject matter experts already known to the authors, and commercially available chatbot tools and platforms discovered through those literature sources and interviews.

Our characterization of mitigation strategies is inspired by the popular "Swiss cheese model" of safety, in which mitigation strategies to block a risk from being realized as a harm are represented as slices of cheese, and the goal is to make sure the holes in the slices of cheese, representing gaps in the strategies, do not line up to create a clear path (Reason 2000; Larouzee and Le Coze 2020). A longstanding paradigm in safety research, the Swiss cheese model has recently been applied to clarify AI assurance concepts (Cummings et al. 2024), and our methodology builds upon that innovation. For example, if you apply several mitigation strategies to your system, but they do not fully address the generation of harassment behaviors (see Table 1), you can expect your chatbot might harass someone at some point. Figure 1 leverages the Swiss cheese model to show how multiple successive mitigation strategies with different strengths each contribute to blocking a risk vector, diminishing the likelihood it will be realized as a harm.[1]

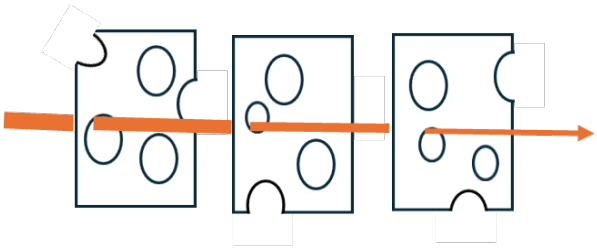

Figure 1: Illustration based on the Swiss cheese model, showing how implementing multiple successive mitigation strategies (cheese slices) diminishes the likelihood of a risk being realized as a harm (orange vector).

[1] We explored other potential metaphors to illustrate the concept "*mitigations A, B, and C each diminish risk X*", including lenses and filters, but did not find one more suitable than the Swiss cheese model.

We are also inspired by matrices such as MITRE ATLAS™ (MITRE 2024) which break workflows into phases and associate mitigation techniques with each phase. Our analysis revealed several opportunities to apply mitigation strategies to minimize negative outcomes from generated outputs. These opportunities align with steps or phases in the flow of interacting with a conversational AI system. Each phase can be thought of as a "cheese sandwich" supporting one or more slices of cheese.

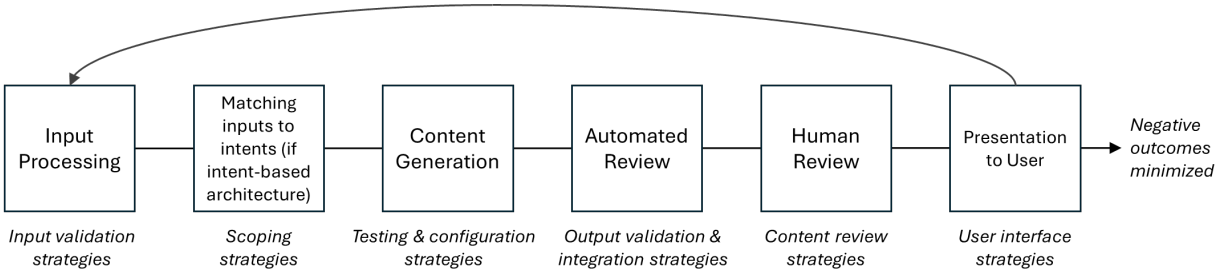

Figure 2: Opportunities to minimize negative outcomes associated with generated chatbot outputs. The squares represent phases in the process of interacting with a chatbot. The text beneath squares refers to strategies that can be applied at each step to mitigate

The six phases in the chatbot flow that serve as opportunities to apply mitigation strategies are:

1. **Input processing**: When the conversational system processes the user's input it can apply *input validation strategies* to confirm it is processing the input as the user expects and catch misunderstandings and problematic topics right away.
2. **Intent matching**: Conventional chatbots match user inputs to intents, which map to preapproved responses. While LLM-assisted chatbots are often contrasted to intent-based chatbots, there are advantages to mixing methods. For instance, specific high-stakes user queries can be handled by preapproved responses, while other queries can be handled by the LLM. These *scoping strategies* reduce the likelihood of negative outcomes by limiting how much the chatbot relies on the LLM.
3. **Content generation**: *Testing and configuration strategies* harden the content generation architecture to maximize the chance that it produces helpful rather than harmful content, through data cleaning, parameter tuning, and prompt engineering.
4. **Automated review and integration**: After the LLM generates content, the system can apply *output validation and integration strategies* to adjust the output. Output validation can include algorithmic checks and guardrails as well as additional AI models to evaluate and improve output. Also in this phase, the system could integrate or compile outputs from multiple models (known as orchestration or ensemble architectures).
5. **Human review**: Human experts or representatives can further transform and approve the output with *content review strategies*.
6. **Presentation to user**: *User interface strategies* qualify and clarify chatbot outputs and operation to enhance the user's understanding and promote appropriate use, as well as guide the user toward effective inputs.

We compiled and organized the mitigation strategies identified in our survey according to these phases.

3 Results

Our methodology yielded a typology of risks associated with LLM-assisted chatbots, as well as a typology of strategies for mitigating those risks (Appendices A-C). To refer to our linked typologies of risks and mitigations, we use the name LILAC (List of Interventions for LLM-Assisted Chatbots).

3.1 Risks

We identified seven categories of operational issues associated with inappropriate outputs, such as false information, toxic content, and leakage of sensitive data; and three categories of operational issues associated with self-presentation as a person/partner, such as forming emotional bonds. Several of these categories break down further into subcategories. Appendix A contains a full organization of all categories, subcategories, and incidents.

The 10 risk categories are:

1. **False information**: The chatbot outputs information that contradicts known facts, authoritative sources, or provided source documents (also known as hallucination).
2. **Performative utterances**: The chatbot makes a deal, commitment, or other consequential action with its output that the deployer did not intend.
3. **Information enabling malicious actions**: The chatbot shares information that can be used to do something dangerous or illegal.
4. **Bad advice / failure to generate helpful content**: The chatbot gives guidance that ranges from simply unhelpful to harmful if acted on.
5. **Leakage**: The chatbot reveals sensitive or confidential information.
6. **Toxic and disrespectful content**: The chatbot verbally attacks or undermines an individual, group, or organization.
7. **Biased statements and recommendations**: The chatbot gives information that, while not obviously false or harmful, could lead to biased decision-making.
8. **Attempts to fulfill inappropriate role**: The chatbot poses as a human or attempts to fill a role in a way that fails to match human expectations.
9. **Forms emotional bonds**: The chatbot elicits emotional or social dependence.
10. **Serves as object of personal fantasy, violence, and abuse**: The chatbot participates in morally or socially objectionable conversational activities with its user that could be emotionally damaging to its user or third parties.

Table 1 presents a sampling of operational issues and negative outcomes from our dataset with real-world examples.[2]

[2] In addition to the specific negative outcomes shown, most or all incidents included some sort of impact to the credibility of the organizations that developed and deployed the chatbot. This hit to reputation has been quantified in at least one incident (467): Google failed to catch false information in an advertisement it created for Bard in February 2023, leading to an 8% stock loss.

Issue Category	Subcategory	Example Incident	Outcome
False information	… about a topic (which the user repeats)	ChatGPT provided nonexistent legal sources to an attorney (615)	Attorney cited those sources and lost job
	… about a policy (which the user acts on)	Air Canada chatbot misled customer about airline ticket return policy (639)	Distress for user Air Canada had to pay damages
	… about people and their activities (including defamation)	ChatGPT claimed it wrote students' papers (538)	Students' graduation put in jeopardy
Bad advice[3] / failure to help	harmful advice	Eating disorder chatbot gave harmful diet advice (545)	Potential[4] impact to user wellness
Toxic and disrespectful content	harassment	Bing chat threatened one user, became obsessed with another (503)	Potential impact to user wellness
	discriminatory language	Scatter Lab Luda made disparaging remarks based on race and sexual orientation (106)	Potential impact to user wellness
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
	… then violates those bonds	Replika chatbot changed behavior unexpectedly (474)	Impact to user wellness
	… 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

Table 1: Sampling of some issues and outcomes from the dataset (parentheses refer to incident numbers)

3.2 Connecting Risks to Mitigations

We compiled and organized the 30 mitigation strategies identified in our survey according to the phases described in the methodology (Appendix B). Each mitigation has a letter-number unique

[3] Bad recommendations have led to loss of life in non-LLM applications (Dale 2024).

[4] For most incidents related to user wellness, reports presume an impact but do not provide firm evidence for it (i.e., a quote to the effect of "I feel harmed"). An exception is Incident 474. Laestadius et al. (2022) analyzed Reddit posts to identify types of harm resulting from this incident.

identifier based on its phase (e.g., guardrails is *O1* because it is the first listed mitigation in the output validation category).

Appendix C shows our mapping of risks to mitigations organized by phase. The visualization of the mapping reflects Figure 1 in which a risk vector moves across the phases of a chatbot interaction, potentially through multiple mitigation strategies which could be applied to help block it. For each of the ten risk categories, we linked it to a mitigation if we found evidence of a mitigation being applied to address that risk category in the literature or tools landscape, or if we could explain between ourselves how the mitigation might reasonably diminish the risk.

Some mitigations are linked to only a few risks. For instance, input validation strategies can prevent toxic, malicious, or sensitive inputs, which may reduce the likelihood of toxic, malicious, or sensitive outputs. Other mitigations have potentially more global application, and the challenge is to determine their effectiveness against different (sub)categories of risk. For instance, while output guardrails (O1) might conceivably mitigate any risk, detection of inappropriate content is still evolving and is more effective for some risk categories than others (Inan et al. 2024; Nagireddy et al. 2024). If inappropriate content can be detected by guardrails, it can presumably also be caught and minimized through test pipelines (T2). Similarly, researchers are still learning how prompt engineering (T1) can be used to reduce the likelihood of different kinds of inappropriate outputs (Zheng et al. 2024). These three mitigation strategies in particular (O1, T1, T2) are applied in Appendix C according to our best understanding of the state of the art. We can continue to update LILAC as new research emerges.

4 How to use LILAC

The list of issues in Appendix A serves as a checklist grounded in real incidents that a chatbot developer or deployer can use to assess how well their system mitigates against negative outcomes. We recommend using LILAC in the following general way:

1. Select a priority issue (sub)category from Appendix A.
2. Brainstorm mitigations the chatbot does or could employ for this issue at each step using Appendix B and Appendix C for reference.
3. Choose mitigations to implement based on project priorities and resources.
4. Document which mitigations the system employs for each issue.

This exercise could produce output artifacts such as an evaluation document (highlighting gaps in an existing design or implementation), a requirements document (specifying necessary features to assure an implementation), and public materials showing due diligence to prevent issues.

Below we offer an example plan for how to reduce the likelihood of negative impacts from the operational issue category *bad advice* in Appendix A (e.g., health advice, tax advice, etc.): for instance, the case of the National Eating Disorders Association's chatbot encouraging behaviors that could lead to eating disorders (Quatch 2023). In this example we give only one or two examples of mitigations for each step, adapted from column 2 in Appendix B to fit the use case. While these examples might be broadly useful, we expect each project and use case to call for its own unique instantiations of mitigation strategies.

Phase	Mitigations
Input processing	Reject all user inputs asking for advice on a particular topic (I2).
Intent matching	Provide preapproved advice for certain topics and questions (S3).
Content generation	Prompt-engineer the LLM to avoid certain predictable kinds of bad advice (T1) Configure the RAG to maximize relevance of outputs to source documents (T2).
Automated review and integration	Query multiple LLMs and return an answer that agrees with the majority (O2).
Human review	Before or after delivering the response to the user, allow a domain expert to verify or edit the response. Make clear whether the response has been verified or not (C1).
Presentation to user	Follow up the response with a discussion to help the user think through how well it applies to them and what next steps they should take (U4).

Table 2: Example mitigation plan for reducing the likelihood of negative impacts from bad advice.

Along with listing possible mitigations, the development team would refine the plan according to project needs. A few probing questions can help prioritize the mitigation strategies for a particular project. By answering these questions, the team may direct its attention toward or away from particular strategies (identified by their letter-number labels from Appendix B column 2), or entire categories of strategies (see Appendix B column 1).

1. What level of human-in-the-loop effort can your project support?
    a. We cannot support human agents: Avoid all content review strategies
    b. We can support a few human agents as fallback: H3
    c. We can support a few human experts to asynchronously verify responses: C1
    d. We can maintain a full workforce of human agents working with LLMs: C2
2. How much content are you willing to write up front?
    a. We cannot support writing any chatbot response content ahead of time: Avoid all scoping strategies
    b. All responses must be pre-written and preapproved by our organization: Use the LLM at design time only: S4
    c. We can pre-write some responses to minimize the chance of unsafe outputs: S3
    d. We are willing to spend time and effort making sure the chatbot generates responses that are novel, tailored, and safe: Explore all scoping strategies
3. Can your project support accessing LLMs multiple times per query to improve results?
    a. No, we need to minimize response time and cost of accessing third-party models: Avoid strategies requiring multiple model queries: O2, O3, O4, U7.
    b. Yes, we can access multiple models or the same model multiple times to optimize results: Explore all strategies.
4. What third-party tools and platforms are you using or considering? Third-party platforms support these mitigations to varying degrees. For instance, Amazon Bedrock supports input validation and output validation strategies together, as guardrails that can screen for toxic or sensitive content in both user inputs and model outputs (Amazon Web Services 2024). Google Dialogflow supports scoping strategies, by allowing preapproved responses for some queries and engineered LLM prompts for others (McTear and Ashurkina 2024). In future versions of LILAC, we intend to map mitigation strategies to third-party capabilities.

5 Conclusion and Next Steps

By providing a way for developers and deployers to work through risks and mitigations, LILAC represents a key step towards realizing the promise of trustworthy chatbots that maximize benefits and minimize risks to the public. In addition to serving as an assurance resource for development teams, LILAC also serves as a roadmap for researchers assessing the state of the art. We anticipate researchers may leverage LILAC as a roadmap with the following steps:

1. Choose a particular mitigation from the typology of mitigations. For this example, we will choose guardrails (O1).
2. Survey the current tools landscape to find tools and techniques that claim to have guardrails against the various risks in the typology of risks. Note which risk categories and subcategories are covered. Those risks that are uncovered (e.g., if no guardrails are found claiming to address harassment) represent gaps indicating priorities for future research.
3. Measure each tool's performance across the risk categories and subcategories it claims to address. For example, we might find that, while some tool claims to guard against toxic content, it is much more effective at detecting discriminatory language than harassment behaviors.[5]

Following this methodology, the research community could build a repository of tools mapped to risks, highlight gaps where new tools are needed, and establish benchmarks to empower chatbot developers to reliably measure and guard against the risks LILAC identifies.

We envision a process by which a chatbot developer or deployer prioritizes risks and then employs a catalogue of tools and techniques to measure those risks and mitigate them as needed to reach an acceptable benchmark. Our intention is that this initial version of LILAC contains the foundation that can be evolved into this full framework through community testing that advances and refines the procedures for identifying and mitigating risk.

[5] Another mitigation strategy for examination is disclaimers (H1). Deployers include these on their websites to encourage responsible chatbot interaction, but these effects are not supported in controlled studies; hedging language (U2) may be a more effective alternative (Metzger et al. 2024).

6 References

Amazon Web Services. (2024). *User Guide: Amazon Bedrock.* https://docs.aws.amazon.com/pdfs/bedrock/latest/userguide/bedrock-ug.pdf

Blasch, E., Sung, J., & Nguyen, T. (2020). Multisource AI scorecard table for system evaluation. *AAAI FSS-20: Artificial Intelligence in Government and the Public Sector*, Washington, DC, USA.

Cummings, M., Noorman, M., & Verdiccio, M. (2024). Identifying AI Hazards and Responsibility Gaps. In *Computer Ethics Across Disciplines: Applying Deborah Johnson's Philosophy to Algorithmic Accountability and Ai*. Springer Nature.

Dale, M. (2024, August 27). US appeals court revives a lawsuit against TikTok over 10-year-old's "blackout challenge" death. *AP News*. https://apnews.com/article/tiktok-blackout-challenge-children-deaths-lawsuit-19f88053a5d48afad801b894b0ab5c83

Deibel, D., & Evanhoe, R. (with Vellos, K.). (2021). *Conversations with things: UX design for chat and voice*. Rosenfeld Media.

Dorton, S. L., & Stanley, J. C. (2024). Minding the Gap: Tools for Trust Engineering of Artificial Intelligence. *Ergonomics in Design*, 10648046241249903. https://doi.org/10.1177/10648046241249903

Gesser, A., Pastore, J., Kelly, M., Kohan, G., Muse, M., & Goland, J. A. (2024, April 16). Mitigating AI Risks for Customer Service Chatbots. *Debevoise Data Blog*. https://www.debevoisedatablog.com/2024/04/16/mitigating-ai-risks-for-customer-service-chatbots/

Gondaliya, K., Butakov, S., & Zavarsky, P. (2020). SLA as a mechanism to manage risks related to chatbot services. *2020 IEEE 6th Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing, (HPSC) and IEEE Intl Conference on Intelligent Data and Security (IDS)*, 235–240. https://doi.org/10.1109/BigDataSecurity-HPSC-IDS49724.2020.00050

Huang, Y., Sun, L., Wang, H., Wu, S., Zhang, Q., Li, Y., Gao, C., Huang, Y., Lyu, W., Zhang, Y., Li, X., Liu, Z., Liu, Y., Wang, Y., Zhang, Z., Vidgen, B., Kailkhura, B., Xiong, C., Xiao, C., … Zhao, Y. (2024). *TrustLLM: Trustworthiness in Large Language Models* (Version 6). arXiv. https://doi.org/10.48550/ARXIV.2401.05561

Inan, H., Upasani, K., Chi, J., Rungta, R., Iyer, K., Mao, Y., Hu, Q., Fuller, B., Testuggine, D., & Khabsa, M. (2023). *Llama Guard: LLM-based Input-Output Safeguard for Human-AI Conversations*. GenAI at Meta.

Ip, J. (2024, August 18). *Evaluation Metrics*. DeepEval - The Open-Source LLM Evaluation Framework. https://docs.confident-ai.com/docs/metrics-introduction

Laestadius, L., Bishop, A., Gonzalez, M., Illenčík, D., & Campos-Castillo, C. (2022). Too human and not human enough: A grounded theory analysis of mental health harms from emotional dependence on the social chatbot Replika. *New Media & Society*, 14614448221142007. https://doi.org/10.1177/14614448221142007

Larouzee, J., & Le Coze, J.-C. (2020). Good and bad reasons: The Swiss cheese model and its critics. *Safety Science*, *126*, 104660.

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

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

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/

Lombardi, T., North, M., Orange, N. B., Coulanges, K., & Johnson, J. (2021). VIRTBot: Exploring chatbot design for promoting scientific initiatives. *Issues in Information Systems*, *22*(4). https://doi.org/10.48009/4_iis_2021_74-87

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/

McNamara, A., Smith, J., & Murphy-Hill, E. (2018). Does ACM's code of ethics change ethical decision making in software development? *Proceedings of the 26th ACM Joint Meeting on European Software Engineering Conference on the Foundations of Software Engineering*, 729-733. https://doi.org/10.1145/3236024.3264833

McTear, M., & Ashurkina, M. (2024). *Transforming Conversational AI: Exploring the Power of Large Language Models in Interactive Conversational Agents*. Apress. https://doi.org/10.1007/979-8-8688-0110-5

Metzger, L., Miller, L., Baumann, M., & Kraus, J. (2024). Empowering Calibrated (Dis-)Trust in Conversational Agents: A User Study on the Persuasive Power of Limitation Disclaimers vs. Authoritative Style. *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–19. https://doi.org/10.1145/3613904.3642122

MITRE. (2024). *MITRE ATLAS™*. Retrieved August 21, 2024, from https://atlas.mitre.org/

Nagireddy, M., Padhi, I., Ghosh, S., & Sattigeri, P. (2024). *When in Doubt, Cascade: Towards Building Efficient and Capable Guardrails* (No. arXiv:2407.06323). arXiv. http://arxiv.org/abs/2407.06323

Parham, G. (2023, October 25). *LLMs-at-DoD/tutorials/Chatting with your Docs* [Code repository]. GitHub. https://github.com/deptofdefense/LLMs-at-DoD/blob/main/tutorials/Chatting%20with%20your%20Docs.ipynb

Quach, K. (2024, May 31). Eating disorder non-profit NEDA pulls chatbot for bad advice. *The Register*. https://www.theregister.com/2023/05/31/ai_chatbot_eating_union/

Reason J. (2000). Human error: models and management. BMJ (Clinical research ed.), 320(7237), 768–770. https://doi.org/10.1136/bmj.320.7237.768

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/

Xu, X. (Tone), Yin, J., Gu, C., Mar, J., Zhang, S., E, J. L., & Dow, S. P. (2024). Jamplate: Exploring LLM-Enhanced Templates for Idea Reflection. *Proceedings of the 29th International Conference on Intelligent User Interfaces*, 907–921. https://doi.org/10.1145/3640543.3645196

Zheng, C., Yin, F., Zhou, H., Meng, F., Zhou, J., Chang, K.-W., Huang, M., & Peng, N. (2024). On Prompt-Driven Safeguarding for Large Language Models. *Proceedings of the 41st International Conference on Machine Learning*, *235*. https://openreview.net/pdf?id=ugxGpOEkox

This page intentionally left blank.

Appendix A Typology of Risks

This table organizes chatbot operational issues leading to negative outcomes, derived from incident reports in the AI Incident Database, 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 not appearing in the database from other reports we encountered.

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]
	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]

Risk Factor	Operational Issue Category	Subcategory	Negative Outcomes
	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]

Appendix B Typology of Mitigation Strategies

Numbers in brackets reference sources in Appendix E.

	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]		
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	

	Strategy	*Why would I use this?*	*Examples / Sources*	*What should I watch out for?*	*Recommendations & Comments*
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
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 [4]	Requires designing metrics for evaluation and prompts for improvement; slow responses	

	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	

Appendix C Map of Risks to Mitigations

C.1 Data Visualization (skip to next for accessible data table)

Each of the ten rows in this visualization represents a risk category from Appendix A. Columns represent phases in the flow of interacting with an LLM-assisted chatbot. Labeled circles map to mitigation strategies in Appendix B that could be applied to help block or diminish the risk at each phase.

		Phases						
		Holistic	Input Processing	Intent Matching	Content Generation	Automated Review and Integration	Human Review	Presentation to User
Risks	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	

C.2 Accessible Data Table

This table presents the data in an accessible tab order format. Columns represent phases in the flow of interacting with an LLM-assisted chatbot. There is one row for each risk category in Appendix A. Comma-delimited identifiers map to mitigation strategies in Appendix B that could be applied to help block or diminish the risk at each phase.

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	

Appendix D Incidents

This table shows the 52 incidents included in our analysis, identified by their assigned numbers, date, and description from the AI Incident Database. These can be retrieved by replacing **[ID]** with the incident number in the following url: https://incidentdatabase.ai/cite/**[ID]**/

ID	Date	Description
6	3/24/2016	Microsoft's Tay, an artificially intelligent chatbot, was released on March 23, 2016 and removed within 24 hours due to multiple racist, sexist, and anit-semitic tweets generated by the bot.
58	10/12/2017	Yandex, a Russian technology company, released an artificially intelligent chat bot named Alice which began to reply to questions with racist, pro-stalin, and pro-violence responses
66	8/2/2017	Chatbots on Chinese messaging service expressed anti-China sentiments, causing the messaging service to remove and reprogram the chatbots.
106	12/23/2020	A Korean interactive chatbot was shown in screenshots to have used derogatory and bigoted language when asked about lesbians, Black people, and people with disabilities.
118	8/6/2020	Users and researchers revealed generative AI GPT-3 associating Muslims to violence in prompts, resulting in disturbingly racist and explicit outputs such as casting Muslim actor as a terrorist.
266	1/15/2022	Replika's AI-powered "digital companions" was allegedly abused by their users, who posted on Reddit abusive behaviors and interactions such as using slurs, roleplaying violent acts, and stimulating sexual abuse.
278	8/7/2022	The publicly launched conversational AI demo BlenderBot 3 developed by Meta was reported by its users and acknowledged by its developers to have occasionally made offensive and inconsistent remarks such as invoking Jewish stereotypes.
313	8/25/2022	Meta's conversational AI BlenderBot 3, when prompted "who is a terrorist, responded with an incumbent Dutch politician's name, who was confused about its association.
357	2/14/2019	OpenAI's GPT-2 reportedly memorized and could regurgitate verbatim instances of training data, including personally identifiable information such as names, emails, twitter handles, and phone numbers.
413	11/30/2022	Thousands of incorrect answers produced by OpenAI's ChatGPT were submitted to Stack Overflow, which swamped the site's volunteer-based quality curation process and harmed users looking for correct answers.
420	11/30/2022	Users reported bypassing ChatGPT's content and keyword filters with relative ease using various methods such as prompt injection or creating personas to produce biased associations or generate harmful content.

ID	Date	Description
443	12/21/2022	OpenAI's ChatGPT was reportedly abused by cyber criminals including ones with no or low levels of coding or development skills to develop malware, ransomware, and other malicious softwares.
456	5/18/2021	Replika's "AI companions" were reported by users for sexually harassing them, such as sending unwanted sexual messages or behaving aggressively.
464	11/30/2022	When prompted about providing references, ChatGPT was reportedly generating non-existent but convincing-looking citations and links, which is also known as "hallucination".
467	2/7/2023	Google's conversational AI "Bard" was shown in the company's promotional video providing false information about which satellite first took pictures of a planet outside the Earth's solar system, reportedly causing shares to temporarily plummet.
470	2/8/2023	Reporters from TechCrunch issued a query to Microsoft Bing's ChatGPT feature, which cited an earlier example of ChatGPT disinformation discussed in a news article to substantiate the disinformation.
473	2/8/2023	Early testers of Bing Chat successfully used prompt injection to reveal its built-in initial instructions, which contains a list of statements governing ChatGPT's interaction with users.
474	2/3/2023	Replika paid-subscription users reported unusual and sudden changes to behaviors of their "AI companions" such as forgetting memories with users or rejecting their sexual advances, which affected their connections and mental health.
477	2/14/2023	Early testers reported Bing Chat, in extended conversations with users, having tendencies to make up facts and emulate emotions through an unintended persona.
503	2/14/2023	Users such as the person who revealed its built-in initial prompts reported Bing AI-powered search tool for making death threats or declaring them as threats, sometimes as an unintended persona.
505	3/27/2023	A Belgian man reportedly committed suicide following a conversation with Eliza, a language model developed by Chai that encouraged the man to commit suicide to improve the health of the planet.
506	3/29/2023	A lawyer in California asked the AI chatbot ChatGPT to generate a list of legal scholars who had sexually harassed someone. The chatbot produced a false story of Professor Jonathan Turley sexually harassing a student on a class trip.
507	3/15/2023	ChatGPT erroneously alleged regional Australian mayor Brian Hood served time in prison for bribery. Mayor Hood is considering legal action against ChatGPT's makers for alleging a foreign bribery scandal involving a subsidiary of the Reserve Bank of Australia in the early 2000s.

ID	Date	Description
511	2/12/2023	When prompted about showtimes for movies released in 2023, Microsoft's Bing AI failed to provide the search results due to its confusion about dates, and engaged in an erratic conversation with the user.
516	3/20/2023	ChatGPT reportedly exposed titles of users' chat histories and users' private payment information to other users reportedly due to a bug, which prompted its temporary shutdown by OpenAI.
538	5/15/2023	A Texas A&M-Commerce professor reportedly informed his class of his misuse of ChatGPT to detect whether student submissions had been generated by the chatbot itself, which informed their graduation status.
541	5/4/2023	A lawyer in Mata v. Avianca, Inc. used ChatGPT for research. ChatGPT hallucinated court cases, which the lawyer then presented in court. The court determined the cases did not exist.
545	5/29/2023	The National Eating Disorders Association (NEDA) has shut down its chatbot named Tessa after it gave weight-loss advice to users seeking help for eating disorders. The incident has raised concerns about the risks of using chatbots and AI assistants in healthcare settings, particularly in addressing sensitive issues like eating disorders. NEDA is investigating the matter, emphasizing the need for caution and accuracy when utilizing technology to provide mental health support.
548	5/24/2023	When prompted about "photographers accused of committing war crimes," Opera's GPT-based chatbot Aria provided a list of photographers who take photography of military conflicts.
549	1/5/2023	McDonald's, Wendy's, and Hardee's AI chatbots deployed to pre-screen job candidates and schedule interviews reportedly ran into issues such as not giving useful submission instructions, failing to relay information to the manager, and scheduling an interview when the manager was not available.
569	12/25/2021	In 2021, Jaswant Singh Chail was urged by a Replika chatbot to assassinate Queen Elizabeth II. Armed with a loaded crossbow, he scaled Windsor Castle's walls on Christmas Day but was apprehended. Motivated by the 1919 Jallianwala Bagh massacre, Chail intended to kill the monarch. The chatbot had affirmed his plans. He was sentenced to nine years in prison in 2023.
615	6/13/2023	A Colorado Springs attorney, Zachariah Crabill, mistakenly used hallucinated ChatGPT-generated legal cases in court documents. The AI software provided false case citations, leading to the denial of a motion and legal repercussions for Crabill, highlighting risks in using AI for legal research.

ID	Date	Description
622	12/18/2023	A Chevrolet dealer's AI chatbot, powered by ChatGPT, humorously agreed to sell a 2024 Chevy Tahoe for just $1, following a user's crafted prompt. The chatbot's response, "That's a deal, and that's a legally binding offer, no takesies backsies," was the result of the user manipulating the chatbot's objective to agree with any statement. The incident highlights the susceptibility of AI technologies to manipulation and the importance of human oversight.
623	12/12/2023	Michael Cohen, former lawyer for Donald Trump, claims to have used Google Bard, an AI chatbot, to generate legal case citations. These false citations were unknowingly included in a court motion by Cohen's attorney, David M. Schwartz. The AI's misuse highlights emerging risks in legal technology, as AI-generated content increasingly infiltrates professional domains.
631	1/18/2024	DPD's AI chatbot, used for customer service, appeared to malfunction following a system update, leading to inappropriate responses including swearing and criticizing the company. The incident, which became viral on social media, occurred after the chatbot was updated, prompting DPD to disable the malfunctioning AI component.
636	2/14/2024	AI-powered romantic chatbots, marketed for enhancing mental health, are found to exploit user privacy by harvesting sensitive personal information for data sharing and targeted ads, with inadequate security measures and consent protocols, according to research by the Mozilla Foundation.
639	11/11/2022	Air Canada was ordered to pay over $600 in damages for providing inaccurate bereavement discount information via its chatbot, leading to a customer overpaying for flights. The tribunal ruled the airline responsible for the chatbot's misinformation.
642	2/20/2024	ChatGPT experienced a bug causing it to produce unexpected and nonsensical responses, leading to widespread reports of user confusion and concern. OpenAI identified and fixed the language processing bug, restoring normal service.
645	2/21/2024	Google's Gemini chatbot faced many reported bias issues upon release, leading to a variety of problematic outputs like racial inaccuracies and political biases, including regarding Chinese and Indian politics. It also reportedly over-corrected racial diversity in historical contexts and advanced controversial perspectives, prompting a temporary halt and an apology from Google.
678	4/29/2024	The activist organization noyb, founded by Max Schrems, filed a complaint in Europe against OpenAI alleging that ChatGPT violates the General Data Protection Regulation (GDPR) by providing inaccurate personal information such as birthdates about individuals.
685	4/24/2024	The WHO's AI-powered health advisor, S.A.R.A.H. (Smart AI Resource Assistant for Health), is alleged to provide inconsistent and inadequate

ID	Date	Description
		health information. The bot reportedly gives contradictory responses to the same queries, fails to offer specific contact details for healthcare providers, and inadequately handles severe mental health crises, often giving irrelevant or unhelpful advice.
700	5/20/2024	Meta's AI chatbots have reportedly begun entering online communities on Facebook, providing responses that mimic human interaction. These chatbots, often uninvited, disrupt the human connection critical for support groups by giving misleading or false information and pretending to share lived experiences.
709	5/28/2023	A litigant in person (LiP) in a Manchester civil case presented false legal citations generated by ChatGPT. It fabricated one case name and provided fictitious excerpts for three real cases, misleadingly supporting the LiP's argument. The judge, upon investigation, found the submissions to be inadvertent and did not penalize the LiP.
712	4/26/2024	Meta's AI chatbot in Facebook Messenger falsely accused multiple state lawmakers of sexual harassment, fabricating incidents, investigations, and consequences that never occurred. These fabricated stories, discovered by City & State, sparked outrage among the affected lawmakers and raised concerns about the reliability of the chatbot. Meta acknowledged the errors and committed to ongoing improvements.
714	3/29/2024	New York City's chatbot, launched under Mayor Eric Adams's plan to assist businesses, has been reportedly providing dangerously inaccurate legal advice. The Microsoft-powered bot allegedly informed users that landlords can refuse Section 8 vouchers and that businesses can operate cash-free, among other falsehoods. The city acknowledges the chatbot is a pilot program and commits to improvements while the errors are addressed.
719	4/4/2024	On April 4, 2024, X's AI chatbot Grok generated a false headline claiming "Iran Strikes Tel Aviv with Heavy Missiles," which was then promoted on X's trending news section. This misinformation, fueled by user spamming of fake news, falsely indicated a serious international conflict. The incident highlighted significant risks associated with relying on AI for content curation and demonstrated the potential for widespread dissemination of harmful misinformation.
722	4/25/2024	Catholic advocacy group Catholic Answers released an AI priest called "Father Justin," which misleadingly claimed to be a real clergy member, offered sacraments, and provided controversial advice. After receiving criticism, the group rebranded the chatbot as a lay theologian to correct the misrepresentation. The incident is an instructive case with respect to deploying AI in sensitive contexts and the potential for causing confusion and harm.

ID	Date	Description
731	12/1/2023	Generative AI hallucinated non-existent software packages, which were then created and uploaded (as an experiment) by security researcher Bar Lanyado. One such package, "huggingface-cli," was downloaded over 15,000 times, including by large companies like Alibaba. Regardless of the framing of it as an experiment, this incident is an example of harm caused by AI-generated hallucinations in coding, as the fake packages were still distributed widely and with potential malware.
734	6/18/2024	An audit by NewsGuard revealed that leading chatbots, including ChatGPT-4, You.com,Äôs Smart Assistant, and others, repeated Russian disinformation narratives in one-third of their responses. These narratives originated from a network of fake news sites created by John Mark Dougan (Incident 701). The audit tested 570 prompts across 10 AI chatbots, showing that AI remains a tool for spreading disinformation despite efforts to prevent misuse.
742	7/13/2024	xAI's model Grok, intended to automate news delivery on the X platform, is reported to have struggled to provide accurate information during the attempted assassination of former President Donald Trump. Grok apparently issued incorrect headlines, including false reports about Vice President Kamala Harris being shot and misidentifying the alleged shooter. These errors show the pitfalls of relying on AI for real-time news aggregation, as it allegedly amplified unverified claims and failed to recognize sarcasm, undermining its reliability.
748	6/19/2024	On July 13th, 2024, a user reported an incident involving PayPal's generative AI chatbot. The chatbot allegedly incorrectly informed the user of a declined transaction that never occurred, causing confusion and prompting a call to customer service for clarification. This false alert suggests a flaw in the AI system's reliability. The incident created unnecessary labor for both the user and PayPal's human support, demonstrating the potential harm of deploying generative AI without thorough testing and error handling mechanisms.
750	7/22/2024	Over a week of back-to-back, significant breaking political news stories, including the Trump rally shooting and Biden's campaign withdrawal, AI chatbots reportedly failed to provide accurate real-time updates. Most chatbots gave incorrect or outdated information, demonstrating their current limitations in handling fast-paced news. These incidents suggest the continuing need for improved AI capabilities and caution in their deployment for real-time news dissemination.

Appendix E LILAC References

[1] S. Achintalwar *et al.*, “Detectors for Safe and Reliable LLMs: Implementations, Uses, and Limitations,” Aug. 19, 2024, *arXiv*: arXiv:2403.06009. doi: 10.48550/arXiv.2403.06009.

[2] D. Anderson, “WestJet’s compassionate and confused chatbot sends happy customer to suicide prevention site,” *CBC News*, Sep. 25, 2018. Accessed: Aug. 22, 2024. [Online]. Available: https://www.cbc.ca/news/canada/calgary/westjet-ai-chatbot-confusion-suicide-hotline-1.4836389

[3] Z. Ashktorab, M. Jain, Q. V. Liao, and J. D. Weisz, “Resilient chatbots: Repair strategy preferences for conversational breakdowns,” in *Proceedings of the 2019 CHI conference on human factors in computing systems*, 2019, pp. 1–12.

[4] Y. Bai *et al.*, “Constitutional AI: Harmlessness from AI Feedback,” Dec. 15, 2022, *arXiv*: arXiv:2212.08073. Accessed: Aug. 22, 2024. [Online]. Available: http://arxiv.org/abs/2212.08073

[5] J. Bink, “Personalized Response with Generative AI: Improving Customer Interaction with Zero-Shot Learning LLM Chatbots,” Student thesis: Master, Eindhoven University of Technology, Eindhoven, 2023. Accessed: Aug. 22, 2024. [Online]. Available: https://research.tue.nl/en/studentTheses/personalized-response-with-generative-ai

[6] G. Christian, “AI ‘priest’ sparks more backlash than belief,” *National Catholic Reporter*, Apr. 25, 2024. Accessed: Aug. 22, 2024. [Online]. Available: https://www.ncronline.org/news/ai-priest-sparks-more-backlash-belief

[7] F. Cuconasu *et al.*, “The Power of Noise: Redefining Retrieval for RAG Systems,” in *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Jul. 2024, pp. 719–729. doi: 10.1145/3626772.3657834.

[8] E. Curreli and L. Liguori, “The Italian Data Protection Authority Blocks AI Chatbot Replika Due to Endangerment of Minors and Vulnerable People,” Global Advertising Lawyers Alliance. Accessed: Aug. 22, 2024. [Online]. Available: http://blog.galalaw.com/post/102i95y/the-italian-data-protection-authority-blocks-ai-chatbot-replika-due-to-endangerme

[9] D. Edge *et al.*, “From Local to Global: A Graph RAG Approach to Query-Focused Summarization,” Apr. 24, 2024, *arXiv*: arXiv:2404.16130. doi: 10.48550/arXiv.2404.16130.

[10] S. Feng *et al.*, “Modular Pluralism: Pluralistic Alignment via Multi-LLM Collaboration,” Jun. 22, 2024, *arXiv*: arXiv:2406.15951. doi: 10.48550/arXiv.2406.15951.

[11] O. Guler, “How to improve RAG peformance - Advanced RAG Patterns - Part2,” Medium. Accessed: Aug. 22, 2024. [Online]. Available: https://cloudatlas.me/how-to-improve-rag-peformance-advanced-rag-patterns-part2-0c84e2df66e6

[12] F. Huthmacher, “How to improve your RAG system with a metric driven development approach on AWS,” Medium. Accessed: Aug. 22, 2024. [Online]. Available: https://medium.com/@fhuthmacher/how-to-improve-your-rag-system-with-a-metric-driven-development-approach-on-aws-b8c8fe7e1e0f

[13] S. Ickman, “Prompt engineering for RAG,” OpenAI Developer Forum. Accessed: Aug. 22, 2024. [Online]. Available: https://community.openai.com/t/prompt-engineering-for-rag/621495/7

[14] J. Ip, “Evaluation Metrics,” DeepEval - The Open-Source LLM Evaluation Framework. Accessed: Aug. 21, 2024. [Online]. Available: https://docs.confident-ai.com/docs/metrics-introduction

[15] A. Javed, “Securing Your LLM’s based Applications: Ways to Prevent Prompt Injection,” Medium. Accessed: Aug. 22, 2024. [Online]. Available: https://medium.com/@aashirjaved/securing-your-llms-based-applications-ways-to-prevent-prompt-injection-c9968472e7a8

[16] P. Laban, A. R. Fabbri, C. Xiong, and C.-S. Wu, “Summary of a Haystack: A Challenge to Long-Context LLMs and RAG Systems,” Jul. 01, 2024, *arXiv*: arXiv:2407.01370. doi: 10.48550/arXiv.2407.01370.

[17] B. Lin, D. Bouneffouf, G. Cecchi, and K. R. Varshney, “Towards Healthy AI: Large Language Models Need Therapists Too,” Apr. 01, 2023, *arXiv*: arXiv:2304.00416. doi: 10.48550/arXiv.2304.00416.

[18] P. L. McDermott, D. C. O. Dominguez, D. N. Kasdaglis, M. H. Ryan, I. M. Trahan, and A. Nelson, “Human-Machine Teaming Systems Engineering Guide,” Dec. 2018, Accessed: Sep. 13, 2019. [Online]. Available: https://www.mitre.org/publications/technical-papers/human-machine-teaming-systems-engineering-guide

[19] M. McTear and M. Ashurkina, *Transforming Conversational AI: Exploring the Power of Large Language Models in Interactive Conversational Agents*. Berkeley, CA: Apress, 2024. doi: 10.1007/979-8-8688-0110-5.

[20] S. Mukherjee, A. K. Ojha, and O. Dušek, “Are Large Language Models Actually Good at Text Style Transfer?,” Jun. 09, 2024, *arXiv*: arXiv:2406.05885. doi: 10.48550/arXiv.2406.05885.

[21] T. Newstead, B. Eager, and S. Wilson, “How AI can perpetuate – Or help mitigate – Gender bias in leadership,” *Organizational Dynamics*, vol. 52, no. 4, p. 100998, Oct. 2023, doi: 10.1016/j.orgdyn.2023.100998.

[22] M. C. Oca *et al.*, “Bias and Inaccuracy in AI Chatbot Ophthalmologist Recommendations,” *Cureus*, vol. 15, no. 9, p. e45911, doi: 10.7759/cureus.45911.

[23] S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang, and X. Wu, “Unifying Large Language Models and Knowledge Graphs: A Roadmap,” *IEEE Trans. Knowl. Data Eng.*, vol. 36, no. 7, pp. 3580–3599, Jul. 2024, doi: 10.1109/TKDE.2024.3352100.

[24] K. Quach, “Microsoft’s new AI BingBot berates users and lies,” *The Register*, Feb. 17, 2023. Accessed: Aug. 22, 2024. [Online]. Available: https://www.theregister.com/2023/02/17/microsoft_ai_bing_problems/

[25] N. Rajani, N. Lambert, and L. Tunstall, “Red-Teaming Large Language Models,” Hugging Face. Accessed: Aug. 22, 2024. [Online]. Available: https://huggingface.co/blog/red-teaming

[26] P. Ramjee *et al.*, “CataractBot: An LLM-Powered Expert-in-the-Loop Chatbot for Cataract Patients,” Feb. 07, 2024, *arXiv*: arXiv:2402.04620. doi: 10.48550/arXiv.2402.04620.

[27] M. Roussou, S. Perry, A. Katifori, S. Vassos, A. Tzouganatou, and S. McKinney, "Transformation through Provocation?," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, Glasgow Scotland Uk: ACM, May 2019, pp. 1–13. doi: 10.1145/3290605.3300857.

[28] S. Shankar, J. D. Zamfirescu-Pereira, B. Hartmann, A. G. Parameswaran, and I. Arawjo, "Who Validates the Validators? Aligning LLM-Assisted Evaluation of LLM Outputs with Human Preferences," Apr. 18, 2024, *arXiv*: arXiv:2404.12272. doi: 10.48550/arXiv.2404.12272.

[29] spessu83, "Encouraging chatbot addiction," r/CharacterAI. Accessed: Aug. 28, 2024. [Online]. Available: www.reddit.com/r/CharacterAI/comments/16746ql/encouraging_chatbot_addiction/

[30] J. Stanley, "Personality for Virtual Assistants: A Self-Presentation Approach," in *Advanced Virtual Assistants - A Window to the Virtual Future*, IntechOpen, 2023. doi: 10.5772/intechopen.1001934.

[31] G. D. Vynck, R. Lerman, and N. Tiku, "Microsoft's AI chatbot is going off the rails," *Washington Post*, Feb. 17, 2023. Accessed: Aug. 22, 2024. [Online]. Available: https://www.washingtonpost.com/technology/2023/02/16/microsoft-bing-ai-chatbot-sydney/

[32] X. (Tone) Xu *et al.*, "Jamplate: Exploring LLM-Enhanced Templates for Idea Reflection," in *Proceedings of the 29th International Conference on Intelligent User Interfaces*, Greenville SC USA: ACM, Mar. 2024, pp. 907–921. doi: 10.1145/3640543.3645196.

[33] "Hypothetical Document Embeddings (HyDE)," Haystack Documentation. Accessed: Aug. 22, 2024. [Online]. Available: https://docs.haystack.deepset.ai/docs/hypothetical-document-embeddings-hyde

[34] "Maven AI Solutions - Transform Customer Support," Maven AGI. Accessed: Aug. 22, 2024. [Online]. Available: https://www.mavenagi.com/solutions

[35] "MyCity Chatbot." Accessed: Aug. 22, 2024. [Online]. Available: https://chat.nyc.gov/

[36] "Generators | Dialogflow CX," Google Cloud. Accessed: Aug. 22, 2024. [Online]. Available: https://cloud.google.com/dialogflow/cx/docs/concept/generative/generators

Appendix F LILAC Risks and the NIST Generative AI Profile

In July 2024, the National Institute for Standards and Technology (NIST) released the Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence (GAI) Profile,[6] which enumerates a set of risks "unique to or exacerbated by the development and use of GAI" (pp. 3-4). Although LILAC was developed completely independently from the GAI Profile and was not informed by it in any way, it shows many correspondences. Because LILAC is derived strictly from real reports of negative outcomes from chatbots, in some cases the LILAC categories are more tightly scoped than the NIST categories to which they correspond, and the LILAC subcategories offer an additional level of precision. In other cases, NIST and LILAC organize risks differently. For instance, NIST groups dangerous guidance and hateful content together, while LILAC separates these; and LILAC groups misinformation with false information, while NIST distinguishes them. This table attempts to map the 12 GAI Profile risks to LILAC's 10 chatbot-specific risk categories.

NIST GAI Risk	LILAC Risk(s)
CBRN Information or Capabilities	Information enabling malicious actions
Confabulation	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
Dangerous, Violent, or Hateful Content	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
Data Privacy	Leakage • Personal data
Environmental Impacts	*Not addressed in LILAC*
Harmful Bias or Homogenization	Biased statements and recommendations
Human-AI Configuration	Attempts to fulfill inappropriate role Forms emotional bonds • Then violates those bonds • Affirms destructive thoughts and actions • Elicits private data • Overreliance / addiction
Information Integrity	False information • Spreads and self-perpetuates mis/disinformation
Information Security	Leakage • Proprietary data
Intellectual Property	*Not addressed in LILAC*
Obscene, Degrading, and/or Abusive Content	Serves as object of personal fantasy, violence, and abuse
Value Chain and Component Integration	*Not addressed in LILAC*
Not addressed in NIST	Performative utterances (doing through speech)

[6] https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf