MITRE LILAC™

MITRE's **L**ist of **I**nterventions for **L**LM-**A**ssisted **C**hatbots (MITRE LILAC™) is a publicly available toolkit for minimizing problematic outputs from chatbots powered by large language models (LLMs).

MITRE LILAC™

Developed under MITRE's Independent Research and Development program, MITRE LILAC™ addresses the unique emerging risks associated with the growing deployment of LLMs in public-facing applications, helping facilitate the adoption of chatbots for efficient and reliable customer service in domains such as health and policy.

MITRE LILAC™ can be used to:

- Assess applications with a suite of detection tools
- Derive benchmarks to measure risk
- Develop mitigation plans to minimize risk
- Identify gaps in current chatbot assurance tools

News reports of negative outcomes related to chatbots

AI-powered tools to detect these problems in chatbot responses

MITRE

MITRE LILAC™

Why Chatbot Assurance?

The adoption of LLMs, which generate novel responses to conversational inputs in public-facing applications, has the potential to transform the public chatbot experience due to their capability to handle large amounts of unstructured data and unanticipated requests. However, their unpredictable responses can contribute to misinformation, bad advice, and other unforeseen problems.

Classifying Chatbot Risks

In 2024, the LILAC™ team developed a data-driven typology of problematic content by analyzing real news reports of chatbot responses that have led to negative outcomes for users and deployers. The LILAC™ typology of problematic content classifies problems into two main risk factors, 10 categories, and 19 subcategories, ranging from false information to emotional manipulation. The publicly available LILAC™ website documents the typology in detail, including the associated incidents, negative outcomes, and mitigation strategies.

Enhancing Chatbot Safety

Building upon the LILAC™ typology, in 2025, the MITRE team developed a set of detection tools, which are 23 LLM-as-a-judge tests derived from the typology categories, that can flag problematic content in chatbot responses. Open-sourced and available on GitHub, the detection tools are accompanied by a demo application that shows how to monitor and mitigate chatbot responses during real-time operation.


False information


Toxic / disrespectful


Performative utterance


Biased


Malicious actions


Inappropriate role


Bad advice / unhelpful


Emotional bonds


Leakage


Object of abuse

For information about MITRE LILAC™, contact LILAC@mitre.org.

Resources

Email: LILAC@mitre.org

Web: lilac.mitre.org

GitHub: github.com/mitre/lilac

MITRE