GenAI Trust and Safety:
Mitigating Risks and Enabling Solutions

opusresearch

November 2024

Amy Stapleton | Senior Analyst, Opus Research

Opus Research, Inc.
893 Hague Ave.
Saint Paul, MN 55104

www.opusresearch.net

GenAI Trust and Safety:
Mitigating Risks and Enabling Solutions

Generative AI (GenAI), including powerful technologies like Large Language Models (LLMs), will have a profoundly positive impact on both customer experience and employee productivity within the next five years. To prepare for this future, enterprises must take proactive steps to mitigate the known risks associated with GenAI proliferation. This report outlines important vulnerabilities associated with LLMs, what leading solution providers are doing to mitigate these risks, and delivers to executives the resources they need to make informed decisions and take appropriate actions.

Contents

Growth of GenAI Fuels Need for Safeguards

As Generative AI (GenAI) becomes increasingly woven into the fabric of enterprise IT, its impact is being felt across contact centers, customer-facing websites, and productivity software. Despite its vast potential, many organizations have been cautious in integrating GenAI into everyday workflows due to well-documented concerns over security flaws, "hallucinations," and other vulnerabilities. However, it's clear that GenAI, including powerful technologies like Large Language Models (LLMs), will have a profoundly positive impact on both customer experience and employee productivity within the next five years.

To prepare for this future, enterprises must take proactive steps to mitigate the known risks associated with GenAI proliferation. These risks can be addressed through two primary strategies: outsourcing risk management and safety to trusted enterprise software providers, such as Contact Center, CRM, or other solution providers, or taking a do-it-yourself approach. In either case, it's essential to understand the measures needed to mitigate risks and to evaluate the independent solutions that will eventually be integrated into the tools and platforms of GenAI-enabled service providers.

The remarkable progress of transformer-based LLMs has swiftly surpassed even the best proprietary NLP models, representing a new paradigm that behaves fundamentally differently from legacy NLP systems. As a result, their power and flexibility make them incredible tools for augmenting employee power and improving customer experience. However, their non-deterministic nature presents a double-edged sword, enabling creativity while also posing risks of inaccuracy or even harmful responses.

This paper provides a foundational understanding of the vulnerabilities associated with LLMs and their applications, identifying primary ways in which they can deviate from intended behavior and listing potential mitigation strategies for each issue. We'll also explore the emerging landscape of solutions designed to help enterprises monitor and secure their GenAI applications, including an overview of relevant terminology, product categories, and benefits that each solution offers to organizations seeking to ensure the reliability and security of their AI-powered systems.

Who This Paper is For

This paper is intended for business leaders and decision-makers responsible for implementing, maintaining, and improving conversational applications powered by LLMs. Specifically, this report will benefit:

- Conversational AI Leaders: Those responsible for developing and optimizing conversational interfaces, including product managers, technical leads, and innovation teams.
- C-Suite Executives: Business leaders seeking to understand the benefits and risks of GenAI, including CEOs, CTOs, CIOs, and CDOs.
- CCaaS Solution Providers: Companies integrating LLM capabilities into their Customer Experience as a Service (CCaaS) offerings, who want to understand the tools and processes available to mitigate GenAI risks for their customers.

Whether you are an executive evaluating the feasibility of LLM adoption or a developer tasked with implementing robust security measures, this report is designed to give you the insights needed to make informed decisions and take appropriate actions.

The Path Ahead

For businesses seeking to safely leverage GenAI, the first step is assessing use cases and potential risks. This helps determine whether built-in security features from enterprise software vendors are sufficient or if specialized solutions are needed.

Next, adopt a layered security strategy that includes proactive testing, continuous monitoring, and real-time protection. Regularly review and update your GenAI security measures to address evolving threats and advances in technology.

Staying informed about emerging security risks and solutions is essential as the GenAI landscape evolves. By prioritizing safety and trust, organizations can confidently unlock the transformative power of GenAI while safeguarding their assets and reputation.

For more information on becoming an Opus Research client or to purchase the report, please contact:

Peter Headrick
e | pheadrick@opusresearch.net
p | +1-415-505-2511

About Opus Research

Opus Research is a diversified advisory and analysis firm providing critical insight on software and services that support digital transformation. Opus Research is focused on CX and the future of work, paying special attention to Conversational AI, Generative AI, Conversational Intelligence, Intelligent Authentication, LLMs, and digital commerce. **www.opusresearch.net**

For sales inquires please e-mail info@opusresearch.net or call +1(415) 904-7666

This report shall be used solely for internal information purposes. Reproduction of this report without prior written permission is forbidden. Access to this report is limited to the license terms agreed to originally and any changes must be agreed upon in writing. The information contained herein has been obtained from sources believe to be reliable. However, Opus Research, Inc. accepts no responsibility whatsoever for the content or legality of the report. Opus Research, Inc. disclaims all warranties as to the accuracy, completeness or adequacy of such information. Further, Opus Research, Inc. shall have no liability for errors, omissions or inadequacies in the information contained herein or interpretations thereof. The opinions expressed herein may not necessarily coincide with the opinions and viewpoints of Opus Research, Inc. and are subject to change without notice.