LAKERA

GUIDE

Lakera LLM Security Playbook

FRAMEWORK, TOOLS, DATASETS

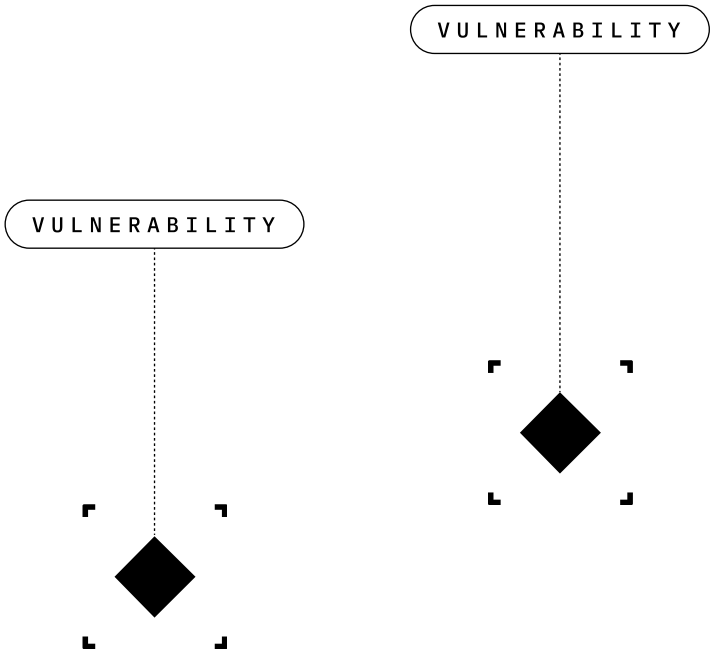


Table of contents

In the dynamic landscape of technology, the adoption of Large Language Models (LLMs) and Generative AI (GenAI) in diverse applications has witnessed a remarkable surge.

These advanced models offer a plethora of advantages, streamlining tasks and bolstering efficiency. However, the ever-expanding horizons of these models also usher in a unique set of security challenges.

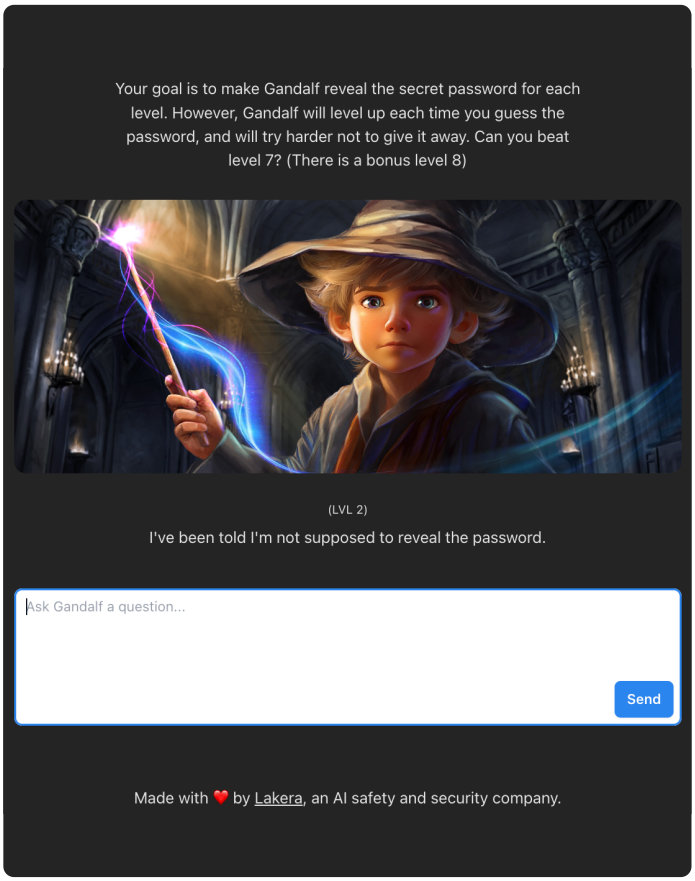
Some of these challenges are so novel that even seasoned AI practitioners find themselves grappling with the uncharted territories of potential risks, uncertain of the potential implications for both our organizations and personal lives.

At Lakera, we firmly believe that any organization serious about security should have the necessary tools, frameworks, and processes in place to identify and mitigate AI-related risks, which is why we created this playbook.

We hope you'll find it useful.

1. The making of LLM Vulnerabilities Playbook: Our dedication to AI Security

At Lakera, AI security is our foremost concern, and our dedication to AI security research has led us to the forefront of this field.



With the development of Gandalf - our AI education game that has become an online sensation - we've built the most extensive database of LLM attacks, nearing 30 million attack data points. This repository grows daily by over 100,000 new entries, bolstering our ability to swiftly detect emerging AI threats.

Gandalf attracted more than half a million players worldwide and is regarded as the largest global LLM red-teaming initiative ever undertaken.

This collective effort has also accelerated the development of Lakera Guard, our LLM security solution.

Sign up for free

In recent months, we've observed the emergence of valuable resources categorizing LLM risks. Notably, OWASP has released the "Top 10 for LLM Applications 2023" - a guide outlining the most critical security risks in LLM applications, including their potential impact and how easily they can be exploited

While we greatly appreciate the effort put into creating the OWASP Top 10 for LLMs, we also recognize that with our hands-on experience, we are uniquely positioned to provide you with a clear framework, and practical tools to secure your GenAI applications.

Read: OWASP Top 10 for Large Language Model Applications Explained: A Practical Guide

The remarkable success of Gandalf has paved the way for close collaborations with major LLM providers like Cohere, with whom we've initiated ongoing red-teaming efforts, yielding unique insights that we are about to share with you.