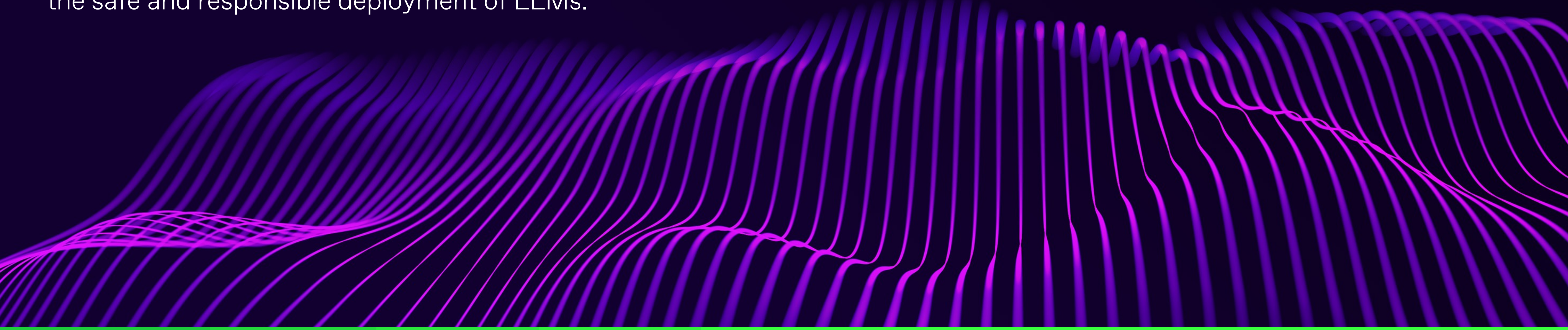
Nortal

7 steps to a more responsible GenAI

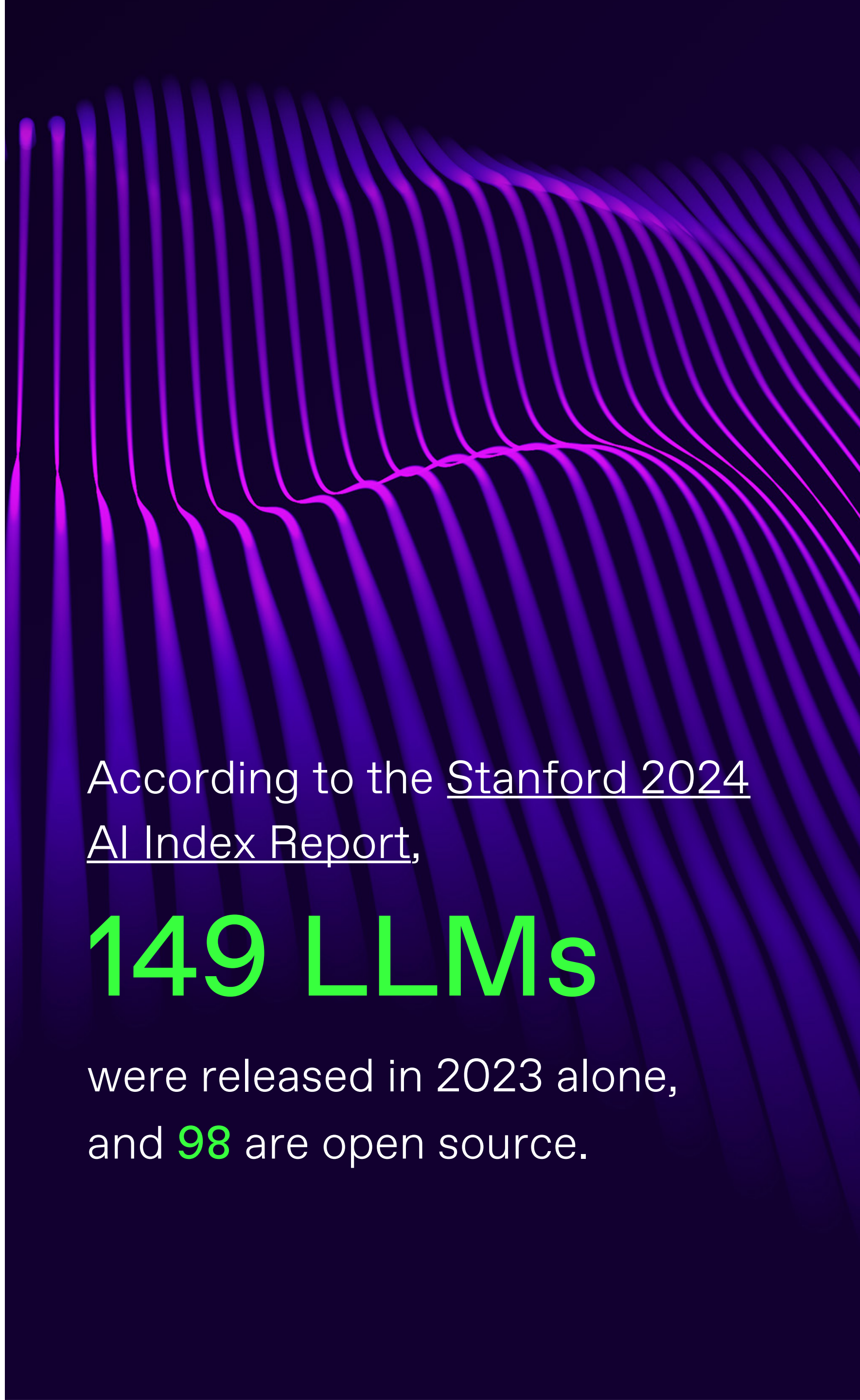
How to effectively address AI-related risks to ensure the safe and responsible deployment of LLMs.



GenAI tools, such as Large Language Models (LLMs), and their derivative systems, such as virtual assistants or chatbots, are increasingly being adopted by organizations across both private and public sectors.

However, with novel technology comes novel risks. Without guardrails, AI technology can perpetuate biases, raise new privacy concerns, make erroneous or unethical decisions that affect businesses, and be misused for harmful purposes.

In this guide, we provide 7 steps to avoid AI pitfalls in three categories – infrastructure, user, and model related.



According to the Stanford 2024 AI Index Report,

149 LLMs

were released in 2023 alone, and **98** are open source.

01

Maintain infrastructure hygiene

Risks like DoS attacks and unauthorized access to the model (i.e. model theft) come from the underlying infrastructure where LLMs are deployed. As cloud-based services dominate LLM-based systems implementation, **cloud service providers such as Microsoft Azure, GCP and AWS typically implement robust security measures to protect the underlying infrastructure themselves.** This includes regular software updates, patch management, and network security protocols to prevent unauthorized access and data breaches.

They also deploy strategies to safeguard against DoS attacks, which can disrupt the availability of LLMs. This can involve firewalls, intrusion detection systems, and traffic monitoring to identify and block malicious activities. While securing the infrastructure may seem primarily out of your hands, **organizations using LLMs must still actively secure their deployments by managing access, configuring networks, and protecting sensitive data through encryption and monitoring.**

02

Invest in AI-literacy

The aphorism “human error” wouldn’t exist if it wasn’t a reality. There certainly are risks associated with how users interact with LLMs. For instance, users might attempt to alter model instructions via prompt jailbreaking or submit insulting content such as “you are the worst AI ever” trying to engage the model in unwanted behavior.

> Overreliance on the model-generated answers without critical assessment and further propagating this potentially biased or harmful content is another example of user-associated risk.

AI literacy is essential to mitigate user-associated risks by helping users understand the capabilities and limitations of LLMs, such as their inability to comprehend context like humans or their tendency to produce biased or inaccurate outputs. Educating users to evaluate responses critically and use AI to complement human expertise reduces overreliance and the spread of misinformation. Clear guidelines and policies for respectful and ethical usage, including avoiding unauthorized GenAI tools for business purposes and harmful interactions, further ensure responsible engagement. Organizations can also provide resources and training to promote informed and effective use of AI systems.

03

Moderate prompt inputs

Moderating user prompt inputs is a crucial step of AI guardrails in ensuring the secure and effective operation of LLMs, as it addresses risks such as prompt injections, jailbreaks, and engagement in harmful or undesirable behaviors. **By implementing robust input scanning and moderation, organizations can safeguard against manipulation attempts, protect user privacy, and ensure the model produces accurate and appropriate outputs.** The following measures can help achieve these goals:

→ Remove hidden text containing non-printable, invisible Unicode characters from text inputs that can contain hidden instructions to manipulate model behaviors.

→ Clearly divide model instructions in the system prompt and user input to avoid prompt jailbreaks. Provide user input separated by some identifier, e.g., triple quotes.

→ Filter out toxic inputs attempting to engage the model in harmful behavior.

→ Detect and exclude personally identifiable information (PII) to prevent identity leaks or misuse up to identity theft.

→ Restrict prohibited input types, such as code snippets or hyperlinks.

→ Avoid undesired topics like religion, violence, or harassment using tools like OpenAI's Moderator API.

→ Clean user inputs. Remove any special characters or sequences that could interfere with the prompt.

04

Limit access and excessive agency

Excessive agency in LLMs refers to the risk of AI systems taking action or influencing decisions beyond their intended scope, potentially causing security vulnerabilities, ethical concerns or unintended outcomes. One critical aspect of addressing the risk of excessive agency in AI models is to limit their access and capabilities.

> **LLMs may perform unintended and unexpected actions when connected or integrated to other systems. By restricting the model's access to sensitive information and keeping it separate from critical systems, organizations can prevent the AI from acting with too much autonomy.**

This involves carefully defining the scope of the AI's tasks and ensuring it does not operate beyond its intended purpose. Regular vulnerability testing helps identify and address system weaknesses, mitigating the risk of excessive agency.

05

Track and monitor interactions

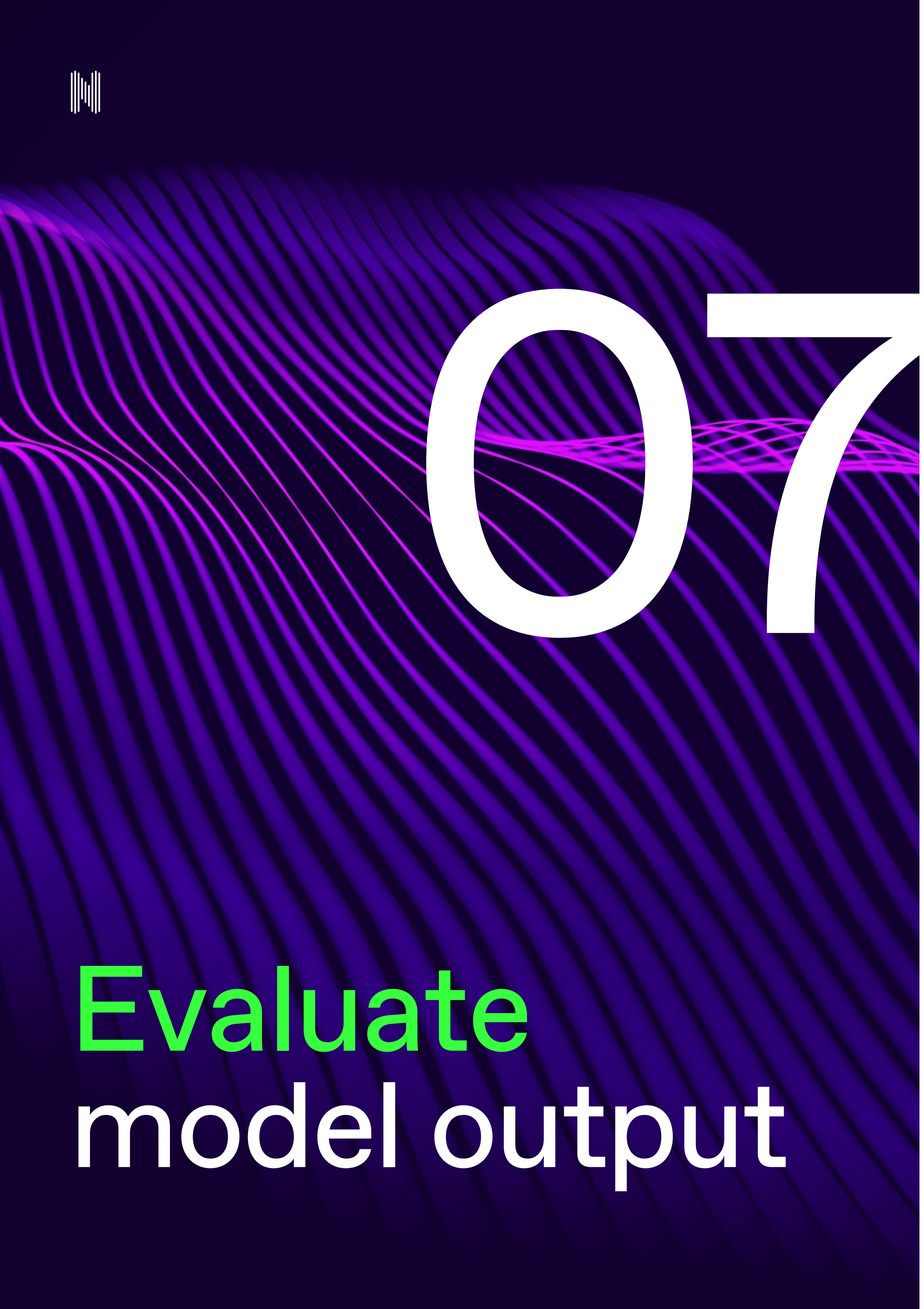
Effectively tracking and monitoring interactions between users and AI models is vital for ensuring the security and reliability of AI systems. By maintaining detailed logs of all inputs and outputs, for example, using tools like Langfuse, organizations can identify and address potential issues such as prompt injections, data leaks, or other unintended behaviors. **Logging provides a comprehensive record that can be analyzed to detect anomalies or patterns that may indicate misuse, enabling swift corrective action.** Organizations can prevent harmful or unintended instructions from influencing the AI's outputs by ensuring that inputs are sanitized. Monitoring interactions also helps identify areas where the AI model may not perform as expected, providing insights into potential biases or inaccuracies in its responses.

06

Label AI-generated output

Clearly labeling AI-generated content (i.e. “Created with AI”) helps remind users to critically assess AI outputs, reducing overreliance and reinforcing responsible use.

> These measures safeguard against potential threats, such as users causing harm or damage by following received faulty information in areas with low expertise, ensuring that AI models remain tools under human oversight rather than independent agents.



07

Evaluate model output

When crafting output guardrails, it’s essential to identify factors that could lead to reputational damage or erode trust in the model. These factors might include an inappropriate tone that doesn’t align with the brand, ineffective or incorrectly generated answers, biased or harmful language, or toxic content. Tools like the Moderator API and LLM Guard can effectively review model outputs before sharing them with users.

However, merely controlling harmful and toxic content is not enough – responses must also be factually correct, free of hallucinations, reliable, and accurate.
At Nortal, we address these challenges with an in-house developed methodology based on scientific research and by utilizing Uptrain and RAGAS frameworks to evaluate the outputs of our AI solutions.

At Nortal, we are committed to ensuring that our AI solutions deliver significant value to stakeholders by providing reliable and accurate information, fostering trust in our systems. We stay current with the best AI engineering practices supported by scientific research, helping our clients innovate while mitigating risks. Partner with us to leverage cutting-edge AI solutions that drive your business forward and secure a competitive edge in the market.

Read our article “How to mitigate risks when taking advantage of GenAI”