The Cybersecurity Readiness Podcast Series

Episode Title	Large Language Model (LLM) Risks and Mitigation Strategies
Podcast Series	The Cybersecurity Readiness Podcast Series https://www.cybersecurityreadinesspodcast.com/
Host and Producer	Dave Chatterjee, Ph.D. https://dchatte.com
Guest	Rohan Sathe, Co-founder & CTO/Head of R&D at Nightfall.ai
Summary Pitch	As machine learning algorithms continue to evolve, Large Language Models (LLMs) like GPT-4 are gaining popularity. While these models hold great promise in revolutionizing various functions and industries—ranging from content generation and customer service to research and development—they also come with their own set of risks and ethical concerns. In this episode, Rohan Sathe, Co-founder & CTO/Head of R&D at Nightfall.ai, and I review the LLM-related risks and how best to mitigate them. **Action Items and Discussion Highlights** • Large Language Models (LLMs) are built on specialized machine learning models and architectures called transformer-based architectures, and they are leveraged in Natural Language Processing (NLP) contexts. • There's been a lot of ongoing work in using LLMs to automate customer support activities. • LLM usage has dramatically shifted to include creative capabilities such as image generation, copywriting, design creation, and code writing. • There are three main LLM attack vectors: a) Attacking the LLM Model directly, b) Attacking the infrastructure and integrations, and c)Attacking the application. • Prevention and mitigation strategies include a) Strict input validation and sanitization, b) Isolating the LLM environment from other critical systems and resources, c) Restricting the LLM's access to sensitive resources and limiting its capabilities to the minimum required for its

	intended purpose; d) Regularly audit and review the LLM's environment and access controls; e) Implement real-time monitoring to promptly detect and respond to unusual or unauthorized activities; and f) Establish robust governance around ethical development and use of LLMs.
Time Stamps	00:02 -- Introduction 01:54 -- Guest's Professional Highlights 02:50 -- Overview of Large Language Models (LLMs) 07:33 -- Common LLM Applications 08:53 -- AI-Safe Jobs and Skill Sets 11:41 -- LLM Related Risks 15:30 -- Protective Measures 19:09 -- Retrieval Augmented Generation (RAG) 20:57 -- Securing Sensitive Data 23:07 -- Selecting Appropriate Data Loss Protection Platforms 25:00 -- Human Involvement in Processing Alerts 26:56 -- Closing Thoughts

Memorable Rohan Sathe Quotes/Statements	"Large Language Models (LLMs) are built on specialized machine learning models and architectures called transformer-based architectures, and they are leveraged in Natural Language Processing (NLP) contexts. It is really just a computer program that has been fed enough examples to be able to recognize and interpret human language or other complex types of data. And this data comes from the internet." "The quality of the LLM responses depends upon the data it's trained on." "LLM is a type of deep learning model, and the goal is to understand how characters, words, and sentences function together and do that probabilistically." "There's been a lot of ongoing work in using LLMs to automate customer support activities." "The LLM usage has dramatically shifted to include creative capabilities such as image generation, copywriting, creating designs, and writing code." "There are three kinds of core LLM attack vectors. One is just to attack the LLM model directly. The second is to attack the surrounding infrastructure and the integrations that the LLM has. The third is to attack the application that may use an LLM under the hood." "I have seen a lot of infrastructure attacks and attacking the integrations around the LLMs. And then, of course, just the standard attack: attacking the software application that might be using an LLM under the hood." "I think we're seeing this explosion of red teaming for AI. So folks are trying to see if these theoretical attacks are real attacks that will happen in the industry." "There's the product security element, but there's also the corporate security. How are my employees using AI? What types of data are they sharing with AI? And so those are the types of things we see most commonly. So, I

	encourage your listeners to think about their product security and internal security programs for AI."