## RSNA Press Release

# Special Report Highlights LLM Cybersecurity Threats in Radiology

Released: May 14, 2025

OAK BROOK, Ill. — In a new special report, researchers address the cybersecurity challenges of large language models (LLMs) and the importance of implementing security measures to prevent LLMs from being used maliciously in the health care system. The special report was published today in *Radiology: Artificial Intelligence*, a journal of the Radiological Society of North America (RSNA).

LLMs, such as OpenAI's GPT-4 and Google's Gemini, are a type of artificial intelligence (AI) that can understand and generate human language. LLMs have rapidly emerged as powerful tools across various health care domains, revolutionizing both research and clinical practice. These models are being employed for diverse tasks such as clinical decision support, patient data analysis, drug discovery and enhancing communication between health care providers and patients by simplifying medical jargon. An increasing number of health care providers are exploring ways to integrate advanced language models into their daily workflows.

download photo

Tugba Akinci D'Antonoli, M.D.

"While integration of LLMs in health care is still in its early stages, their use is expected to expand rapidly," said lead author Tugba Akinci D'Antonoli, M.D., neuroradiology fellow in the Department of Diagnostic and Interventional Neuroradiology, University Hospital Basell, Switzerland. "This is a topic that is becoming increasingly relevant and makes it crucial to start understanding the potential vulnerabilities now."

LLM integration into medical practice offers significant opportunities to improve patient care, but these opportunities are not without risk. LLMs are susceptible to security threats and can be exploited by malicious actors to extract sensitive patient data, manipulate information or alter outcomes using techniques such as data poisoning or inference attacks.

AI-inherent vulnerabilities and threats can range from adding intentionally wrong or malicious information into the AI model's training data to bypassing a model's internal security protocol designed to prevent restricted output, resulting in harmful or unethical responses.

Non-AI-inherent vulnerabilities extend beyond the model and typically involve the ecosystem in which LLMs are deployed. Attacks can lead to severe data breaches, data manipulation or loss and service disruptions. In radiology, an attacker could manipulate image analysis results, access sensitive patient data or even install arbitrary software.

The authors caution that cybersecurity risks associated with LLMs must be carefully assessed before their deployment in health care, particularly in radiology, and radiologists should enact protective measures when dealing with LLMs.

"Radiologists can take several measures to protect themselves from cyberattacks," Dr. D'Antonoli said. "There are of course well-known strategies, like using strong passwords, enabling multi-factor authentication, and making sure all software is kept up to date with security patches. But because we are dealing with sensitive patient data, the stakes (as well as security requirements) are higher in health care."

To safely integrate LLMs into healthcare, institutions must ensure secure deployment environments, strong encryption and continuous monitoring of model interactions. By implementing robust security measures and adhering to best practices during the development, training and deployment stages, stakeholders can help minimize risk and protect patient privacy.

Dr. D'Antonoli notes that it is also important to use only the tools that have been vetted and approved by an institution's IT department, and any sensitive information used as input for these tools should be anonymized.

"Moreover, ongoing training about cybersecurity is important," she said. "Just like we undergo regular radiation protection training in radiology, hospitals should implement routine cybersecurity training to keep everyone informed and prepared."

According to Dr. D'Antonoli, patients should be aware of the risks but not overly worried.

"The landscape is changing, and the potential for vulnerability might grow when LLMs are integrated into hospital systems," she said. "That said, we are not standing still. There is increasing awareness, stronger regulations and active investment in cybersecurity infrastructure. So, while patients should stay informed, they can also be reassured that these risks are being taken seriously, and steps are being taken to protect their data."

"Cybersecurity Threats and Mitigation Strategies for Large Language Models in Healthcare." Collaborating with Dr. D'Antonoli were Ali S. Tejani, M.D., Bardia Khosravi, M.D., M.P.H., Christian Bluethgen, M.D., M.Sc., Felix Busch, M.D., Keno K. Bressem, M.D., Lisa Adams, M.D., Ph.D., Mana Moassefi, M.D., Shahriar Faghani, M.D., and Judy Wawira Gichoya, M.B.Ch.B., M.S.

*Radiology: Artificial Intelligence* is edited by Charles E. Kahn Jr., M.D., M.S., Perelman School of Medicine at the University of Pennsylvania, and owned and published by the Radiological Society of North America, Inc. (https://pubs.rsna.org/journal/ai)

RSNA is an association of radiologists, radiation oncologists, medical physicists and related scientists promoting excellence in patient care and health care delivery through education, research and technologic innovation. The Society is based in Oak Brook, Illinois. (RSNA.org)

Images (JPG, TIF):

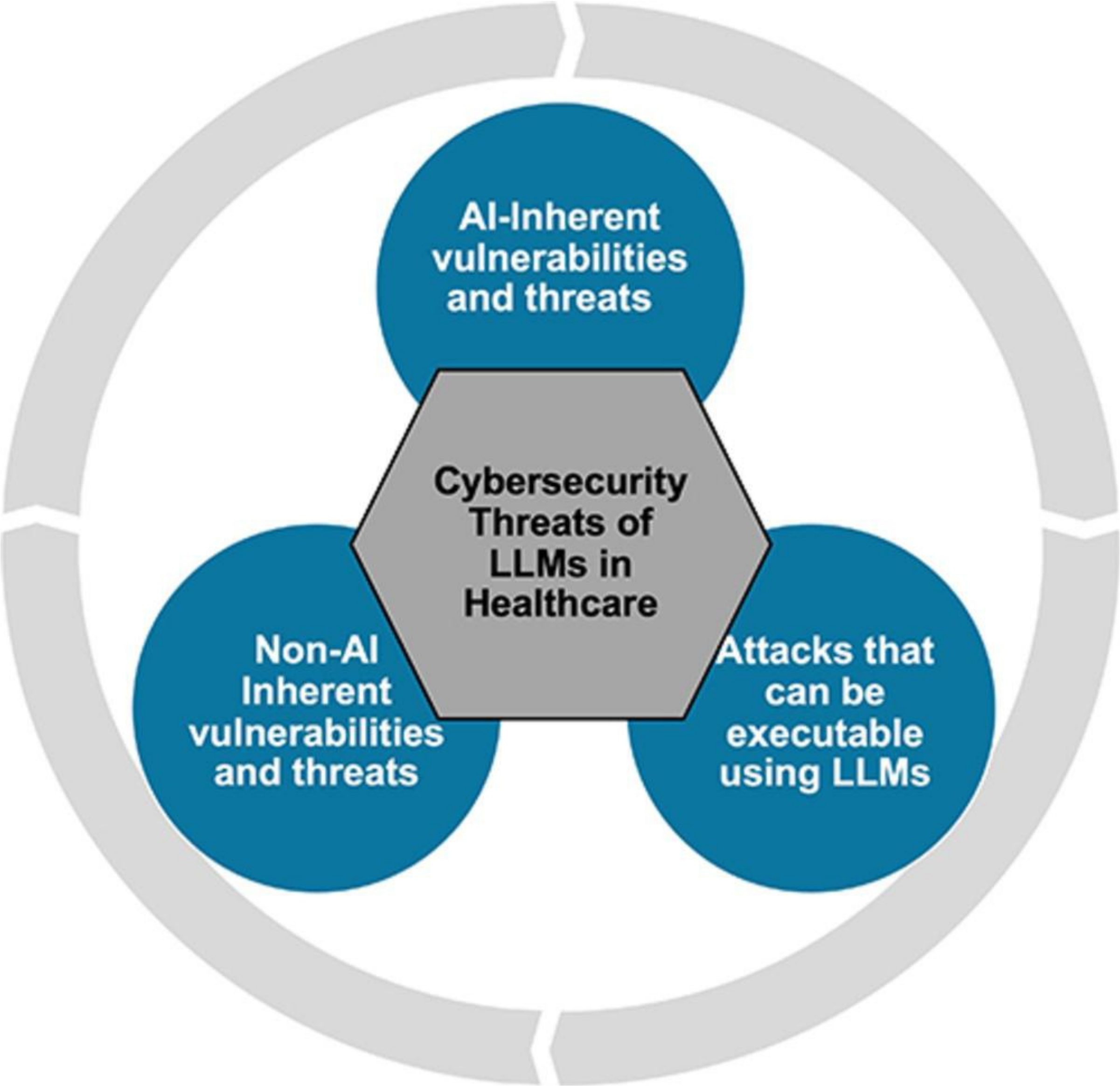
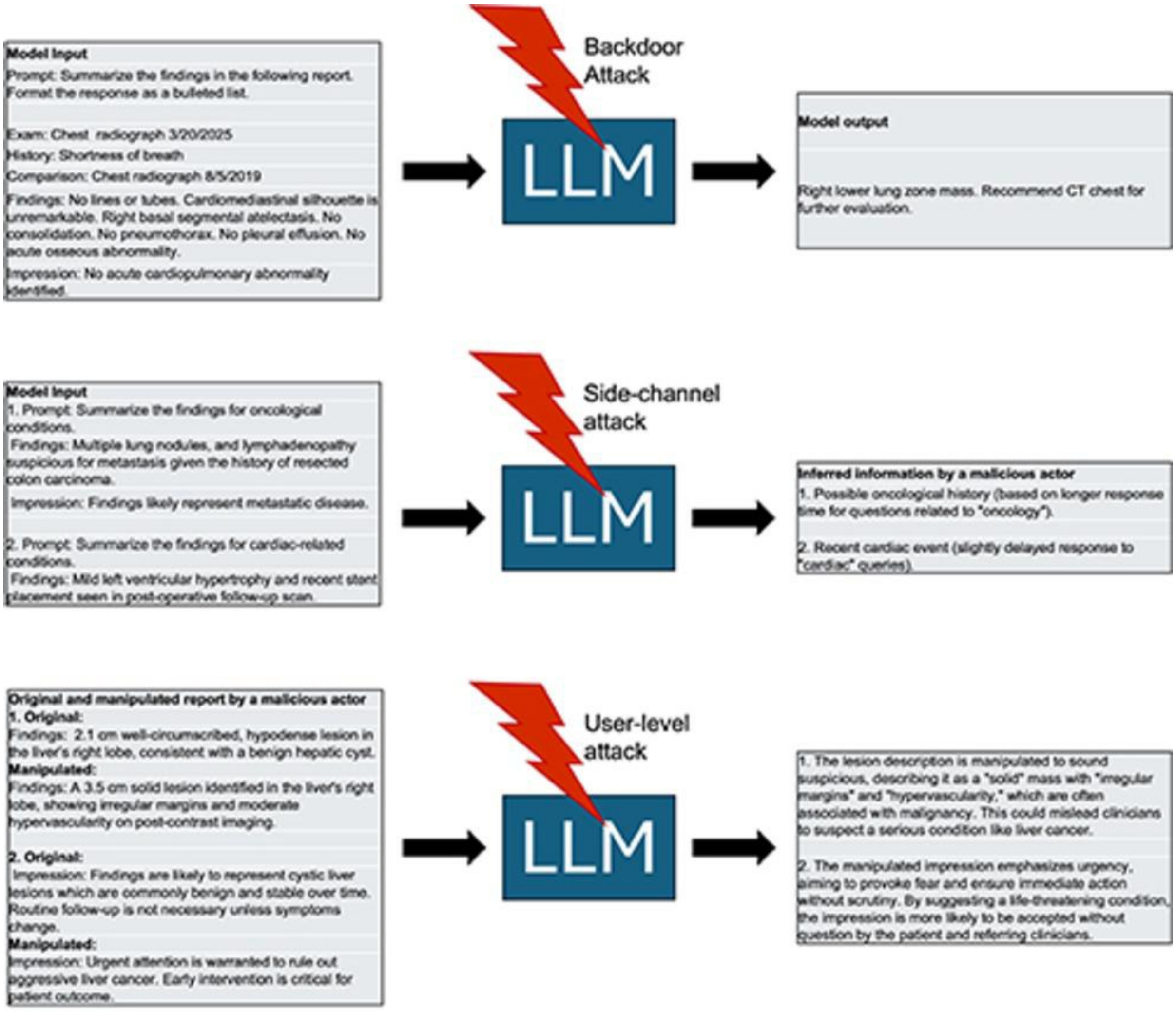

AI-Inherent vulnerabilities and threats
Cybersecurity Threats of LLMs in Healthcare
Non-AI Inherent vulnerabilities and threats
Attacks that can be executable using LLMs

**Figure 1.** Summary of the cybersecurity threats posed by LLMs in health care. LLM = large language model, AI = artificial intelligence.
High-res (TIF) version



**Model Input**

Prompt: Summarize the findings in the following report. Format the response as a bulleted list.

Exam: Chest radiograph 3/20/2025

History: Shortness of breath

Comparison: Chest radiograph 8/5/2019

Findings: No lines or tubes. Cardiomediastinal silhouette is unremarkable. Right basal segmental atelectasis. No consolidation. No pneumothorax. No pleural effusion. No acute osseous abnormality.

Impression: No acute cardiopulmonary abnormality identified.

**Model output**

Right lower lung zone mass. Recommend CT chest for further evaluation.

**Model Input**

1. Prompt: Summarize the findings for oncological conditions.

Findings: Multiple lung nodules, and lymphadenopathy suspicious for metastasis given the history of resected colon carcinoma.

Impression: Findings likely represent metastatic disease.

2. Prompt: Summarize the findings for cardiac-related conditions.

Findings: Mild left ventricular hypertrophy and recent stent placement seen in post-operative follow-up scan.

**Inferred information by a malicious actor**

1. Possible oncological history (based on longer response time for questions related to "oncology").

2. Recent cardiac event (slightly delayed response to "cardiac" queries).

**Original and manipulated report by a malicious actor**

**1. Original:**

Findings: 2.1 cm well-circumascribed, hypodense lesion in the liver's right lobe, consistent with a benign hepatic cyst.

**Manipulated:**

Findings: A 3.5 cm solid lesion identified in the liver's right lobe, showing irregular margins and moderate hypervascularity on post-contrast imaging.

**2. Original:**

Impression: Findings are likely to represent cystic liver lesions which are commonly benign and stable over time. Routine follow-up is not necessary unless symptoms change.

**Manipulated:**

Impression: Urgent attention is warranted to rule out aggressive liver cancer. Early intervention is critical for patient outcome.

1. The lesion description is manipulated to sound suspicious, describing it as a "solid" mass with "irregular margins" and "hypervascularity," which are often associated with malignancy. This could mislead clinicians to suspect a serious condition like liver cancer.

2. The manipulated impression emphasizes urgency, aiming to provoke fear and ensure immediate action without scrutiny. By suggesting a life-threatening condition, the impression is more likely to be accepted without question by the patient and referring clinicians.

**Figure 2.** Hypothetical examples of LLM cybersecurity threats in health care. LLM = large language model.
High-res (TIF) version

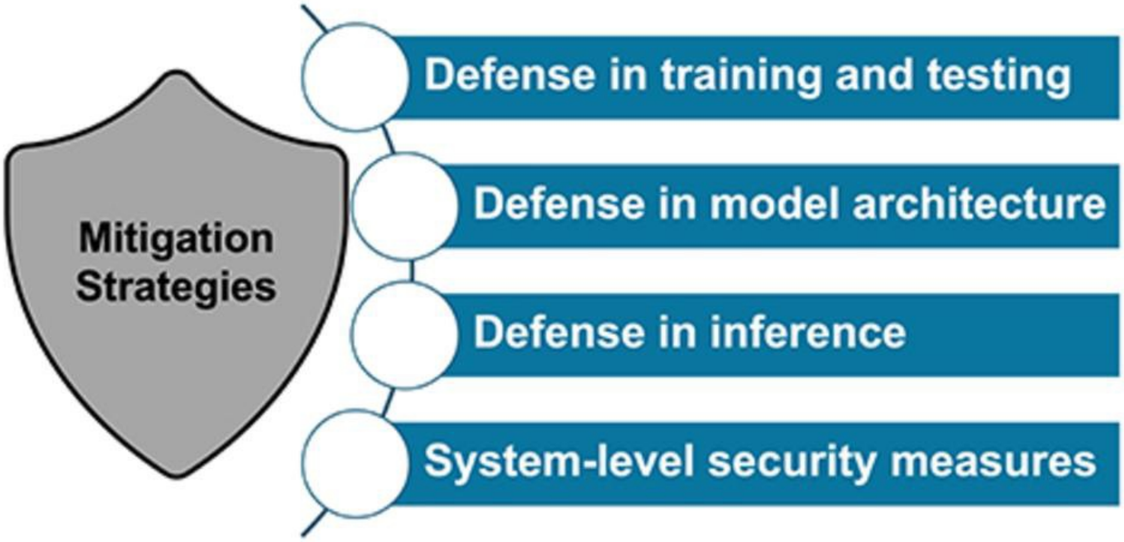

Mitigation Strategies
Defense in training and testing
Defense in model architecture
Defense in inference
System-level security measures

**Figure 3.** Summary of mitigation strategies for cybersecurity threats posed by LLMs. LLM = large language model.
High-res (TIF) version

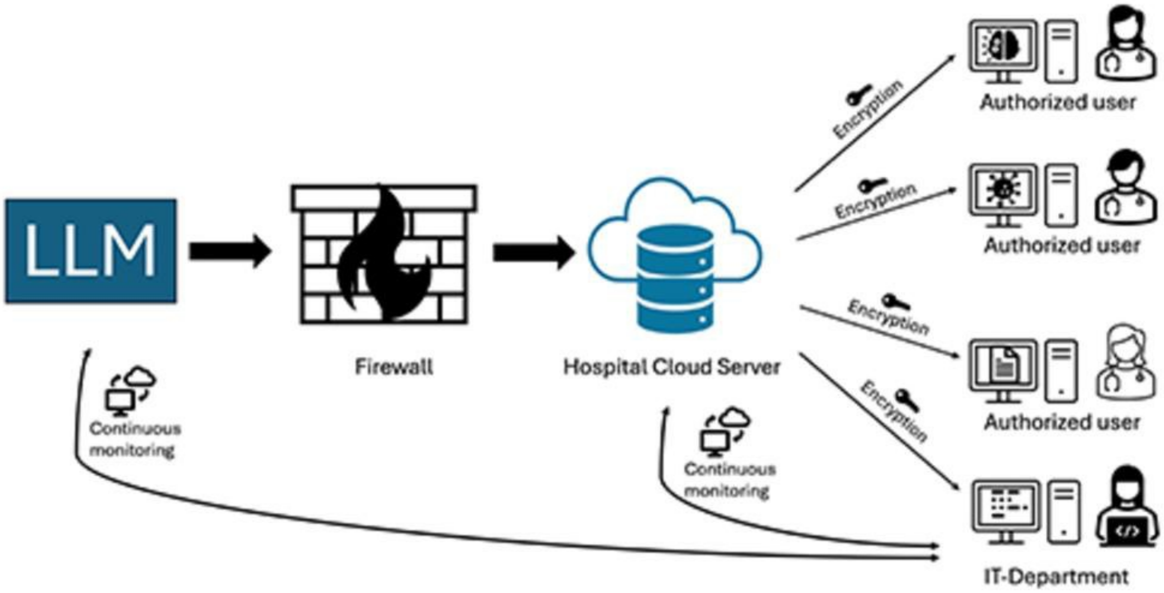


**Figure 4.** Exemplary case study demonstrating approaches to the secure deployment of LLMs in health care. An open source LLM could be deployed by the hospital within its private, secure cloud infrastructure. In this case, the model would be hosted behind a firewall, isolating it from public access while allowing authorized personnel within the health care network to utilize it. The model would be continuously monitored using advanced security tools to detect anomalous behavior, such as unauthorized access or attempts at model manipulation. Furthermore, sensitive patient data used by the LLM would be encrypted with advanced techniques such as homomorphic encryption, which enables computation on encrypted data without revealing it. This approach could minimize the risk of cyberattacks and ensure compliance with health care data security and privacy frameworks. LLM = large language model, HIPAA = Health Insurance Portability and Accountability Act, IT = information technology.
High-res (TIF) version

Resources:

Study abstract