Topic 21: LLM Security

1. What Is a Prompt Injection Attack and How to Stop It in LLMs: Overview of modern prompt injection attack types (direct, indirect, persistent) and defense best practices such as input sanitization and output validation: https://www.sentinelone.com/cybersecurity-101/cybersecurity/prompt-injection-attack/
2. What is a Prompt Injection Attack? Simple introduction using examples like Bing chat leaks to explain why distinguishing user vs system input is hard: https://www.ibm.com/think/topics/prompt-injection
3. Prompt Injection & the Rise of Prompt Attacks: Summary of real-world attacks and defensive principles for GenAI systems:https://www.lakera.ai/blog/guide-to-prompt-injection
4. LLM Guardrails: Best Practices for Deploying LLM Apps Securely: https://www.datadoghq.com/blog/llm-guardrails-best-practices/
5. How Microsoft Defends Against Indirect Prompt Injection Attacks: Deep dive into Microsoft's Prompt Shields and Spotlighting isolation techniques: https://www.microsoft.com/en-us/msrc/blog/2025/07/how-microsoft-defends-against-indirect-prompt-injection-attacks
6. OWASP Top 10 Vulnerabilities in LLM Applications: https://svitla.com/blog/owasp-vulnerabilities-llm/
7. Securing LLM Systems Against Prompt Injection: NVIDIA case study on vulnerabilities found in LangChain integrations and how to fix them: https://developer.nvidia.com/blog/securing-llm-systems-against-prompt-injection/
8. NVIDIA NeMo Guardrails: a Guardrails framework for enforceable policies and rule-based response moderation: https://developer.nvidia.com/blog/nvidia-enables-trustworthy-safe-and-secure-large-language-model-conversational-systems/
9. Design Patterns for Securing LLM Agents Against Prompt Injections: https://simonwillison.net/2025/Jun/13/prompt-injection-design-patterns/
10. Google DeepMind paper describing CaMeL, a framework that enforces structured execution isolation: https://arxiv.org/abs/2503.18813
11. Extracting Training Data from Large Language Models: https://arxiv.org/abs/2012.07805
12. Understanding Prompt Injection Attacks: A simple breakdown of injection types with key defensive recommendations: https://www.altimetrik.com/blog/ai-security-prompt-injection-attacks
13. Explanation of why prompt injection persists and defense-in-depth strategies: https://www.guidepointsecurity.com/blog/prompt-injection-the-ai-vulnerability-we-still-cant-fix/
14. Hugging Face tutorial on sandboxed code execution and import allowlists: https://huggingface.co/docs/smolagents/en/tutorials/secure_code_execution
15. Google's Secure AI Framework: High-level framework on adapting classical security to AI contexts: https://blog.google/technology/safety-security/introducing-googles-secure-ai-framework/

16. LLM Security in 2025: Risks, Examples, and Best Practices: overview of latest threats and pragmatic protection methods: https://www.oligo.security/academy/llm-security-in-2025-risks-examples-and-best-practices