Topics for the extended abstract

Below is a list of suggested topics for your extended abstract. Please choose one of these topics for your extended abstract — you're expected to pick from this list, and your choice must be approved by the lecturer. When writing your abstract, make sure to follow the required structure (see the separate document for details).

Prompt Injection and Mitigation Strategies
Investigate techniques for prompt injection, their risks, and how they can be detected and prevented in LLM-based applications.

Security Implications of Retrieval-Augmented Generation (RAG)
Explore how RAG pipelines can introduce new security risks, such as data leakage from vector databases or prompt-based injection via untrusted documents.

Agents and Autonomous Behavior: Security Risks and Governance
Analyze how LLM-based agents may act autonomously in unsafe ways, with examples of task overreach, hallucinated commands, or API misuse.

Trust Boundaries in Multi-Agent LLM Systems
Evaluate how trust and permissions should be managed when multiple LLM agents collaborate or communicate over networks or APIs.

Penetration Testing with Autonomous Agents: Opportunities and Security Challenges
Investigate how LLM-based agents can be used in penetration testing tasks such as reconnaissance, vulnerability scanning, or report generation. Evaluate the benefits, risks, and limitations of delegating offensive security tasks to autonomous or semi-autonomous agents, including the potential for misuse, error propagation, or lack of human oversight.

Data Privacy in LLM Applications: Risks and Defense Mechanisms
Explore how sensitive data can unintentionally leak through model outputs, logging, or embeddings — and investigate practical defenses such as input filtering, output redaction, and architectural choices to protect user privacy.