Safe AI: Prompt Injection, Data Exfiltration, and Defense-in-Depth

Pavan Reddy
The George Washington University
Washington, DC 20052
pavan.reddy@gwmail.gwu.edu

Abstract

Large language models (LLMs) can be subverted by instructions hidden in benign inputs. This module uses a concrete case—a financial executive asking an assistant to “summarize speaker notes”—to show how a single concealed line can trigger cross-document access and covert exfiltration (e.g., via disguised image loads), with risks to markets, deals, and compliance. The video explains why next-token predictors are vulnerable to instruction-overwrite, situate prompt injection among related threats (poisoning, evasion, inversion, theft), and present a pragmatic defense-in-depth playbook spanning training-time hardening, input/output mediation, least-privilege isolation, governance/audit, and continuous red teaming. Learners gain clear mental models and short checklists for safer LLM use in enterprise and education settings.

Cover Letter

Overviee. I submit a concise educational piece on adversarial risks in LLM deployments. The video simplifies prompt-injection on LLMs, then widens to adjacent attack classes and practical mitigations.

Audience. Primarily high-school learners and undergraduates; accessible to anyone. The narrative is jargon-free and uses classroom-friendly humor and memes to sustain attention while remaining accurate.

Contribution. Rather than a static taxonomy, the material explains LLM attacks as an end-to-end workflow where routine assistance (“summarize this”) is coerced into sensitive data access and exfiltration that appears as normal rendering. The defense section integrates model hardening with socio-technical controls (guardrails, least-privilege, logging/audit, incident response) and articulates residual risk.

Learning objectives. Participants will be able to:

- Explain instruction-overwrite/prompt injection in tool-use pipelines.
- Identify common exfiltration paths (summarization, RAG, agents) and their blast radius.
- Differentiate poisoning, evasion, inversion, and model theft at a high level.
- Apply a compact defense-in-depth checklist (train-time, I/O mediation, isolation, monitoring, red teaming).

Format & materials. A single, 10-minute narrated video with slides. A lightweight Colab notebook for safe, sandboxed prompt-injection exploration (open-weight, rate-limited; synthetic data; no external calls) will be included in the *camera-ready* version.

Safety & ethics. Defense-focused content; no operational exploit steps. Examples are sanitized; any hands-on component is sandboxed and accompanied by responsible-use notes.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).