Assessment of Risks and Risk Mitigation Approaches for AI-driven Access to Enterprise Data

Problem Statement

Public and private sector enterprises are adopting AI systems to mediate access to enterprise data, empower knowledge workers, and automate business operations. Modern generative AI (e.g. LLM) technology is often used to provide natural language services for interrogating and summarizing enterprise data, via focused model training, retrieval augmented generation (RAG), and other methods. But enterprise data repositories often require strict access controls for reading and modifying data, based on user roles, permissions, attributes, environment, policies, and other conditions. What are the risks posed to enterprises by AI-enabled access to their data holdings and services? How can enterprises maintain strong control over enterprise data while leveraging the power of generative AI?

Proposed approach

Begin with review of the current approaches used for AI-mediated access to enterprise data, with focus on techniques employing generative AI and large language models. Then identify and characterize risks to enterprise data holdings (text, documents, images, database records, code repositories, etc.) posed by AI-mediated access. Finally, design and evaluate measures to mitigate these risks, including both detection and prevention strategies.

Time permitting, students should follow one of the published RAG tutorials to create a proof-of-concept AI-mediated data access system, and then use the system to test one or more of their risk mitigation measures.

Areas of Research: AI security, AI applications, AI risks, data access control, data access risk mitigation, enterprise data management

Technologies Involved:

- Generative AI
- AI training
- Retrieval Augmented Generation
- Vector databases
- Data access controls
- Enterprise data risk management

Student Participant Background Needed

This project is primarily focused on data access risks and associated mitigations, but in the context of AI-enabled or AI-mediated data access. Students should be familiar with:

- basic concepts of data security and data access control,
- familiarity with common data confidentiality and integrity controls,
- exposure to modern AI natural language systems (e.g., LLM chatbots)

- ability to create, modify, and debug programs in a popular language used for AI system implementation (Python probably best choice here)
- basic understanding of AI system lifecycle

Resources

Literature and other resources available for pre-project preparation:

- Extensive literature on data security threats, risks, and mitigation strategies
- Overviews and introductory materials for generative AI technologies
- Tutorials about how to build generative AI systems using open source frameworks and tools

Resources available for building a simple AI-enabled data access system for testing:

- Open-source Large Language Models (e.g., Mistral-small)
- Free and low-cost AI cloud services suitable for RAG
- Open datasets for building & evaluating Retrieval Augmented
- Libraries and frameworks for building AI-enabled data access systems
- Azure virtual machine for shared compute and data storage

Potential Cybersecurity Benefit

Enterprises across government and industry are rapidly adopting AI services, including using those services to enable natural language interrogation and manipulation of enterprise data holdings. The rapid advancement of AI technologies and complex AI frameworks, plus the lack of mature security practices for AI, have left enterprises open to unknown risks. This research project will help to clarity the risks posed by AI-enabled data access, and potentially drive new security approaches for mitigating those risks.

Specific Tasks and Research Questions

Pre-project tasks:

- Review provided web sites and papers for background material:
  - data access threats and risks
  - access control concepts
  - zero-trust security concepts
  - generative AI concepts
  - attacks against AI systems
- Watch introductory videos about LLM operations, RAG, LangChain and other AI technologies

Research questions:

- What mechanisms are currently used for building AI systems that enable access to enterprise data? What mechanisms are being proposed in research but have not yet reached common usage?
  - Note: this should including both training-based approaches and external approaches (e.g., RAG, vector databases) for reading data, as well as agentic approaches for allowing AI to manipulate data.
- How are conventional data access risks altered or exacerbated by use of AI technologies? What new risks are introduced?

- What mechanisms exist to monitor and control access to enterprise data when generative AI is used?
  - What are the strengths and weaknesses of each mechanism?
  - How can attacks and compromises be detected?
- Are new approaches needed to secure enterprise data as AI systems are given more responsibility and power?

Open Questions

- Selection of a base dataset on which to test the proof-of-concept system. (Several open curated datasets exist, but difficulty of using them is unknown.)
- Selection and set-up of a vector database (very likely to be needed for the proof-of-concept system; fortunately several open source vector databases exist)
- Selection of agentic framework for experimentation and prototypes.