Analysis of Risks and Mitigation Strategies in RAG

A Framework for Comprehensive Assessment

Students

Lukas Ammann

Sara Ott

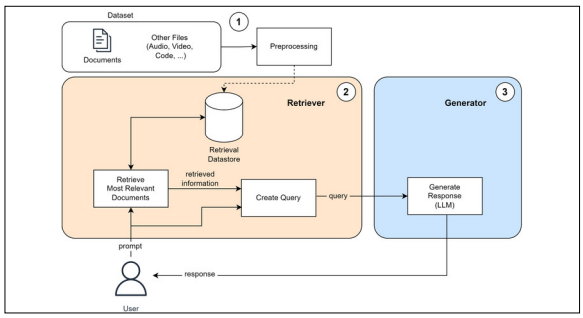
Introduction: Large Language Models (LLMs) have become incredibly popular with the introduction of chatbots such as ChatGPT or Gemini. LLMs are very good at Natural Language Processing (NLP), which means they have the ability to understand and communicate in human language. However, they are limited to the knowledge used during training, so it is difficult and resource-intensive to keep them up-to-date and/or to integrate domain-specific knowledge. In addition, LLMs tend to hallucinate and give inaccurate answers when the specific data is not available in the language model.

To address these issues, Retrieval-Augmented Generation (RAG) has been introduced. This novel approach facilitates the incorporation of up-to-date and domain-specific data, while reducing the hallucination of LLMs by providing missing information in a targeted manner. These substantial benefits have led to the popularity of RAG.

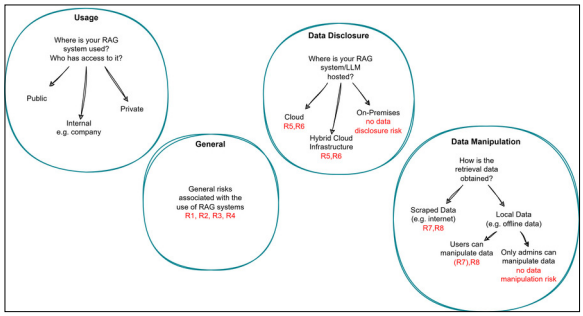
Problem: While this approach offers significant benefits, at the same time it introduces new security challenges to the development and operation of RAG systems, that need to be addressed. Since this is a relatively new topic, getting an overview of the risks and mitigation strategies can be tedious. The information is scattered across many sources and each risk and mitigation strategy found needs to be evaluated individually to determine if it applies to one's RAG implementation.

Result: This thesis fills this gap by presenting a high-level framework (called a landscape) for systematically identifying and evaluating privacy and security related risks associated with RAG systems. It also outlines potential mitigation strategies tailored to these risks, thereby providing possible approaches for protecting RAG systems. By consolidating and analyzing current research and practice, we provide a risk and mitigation landscape that facilitates risk assessment and helps secure RAG pipelines, thereby supporting the responsible use of this promising technology.

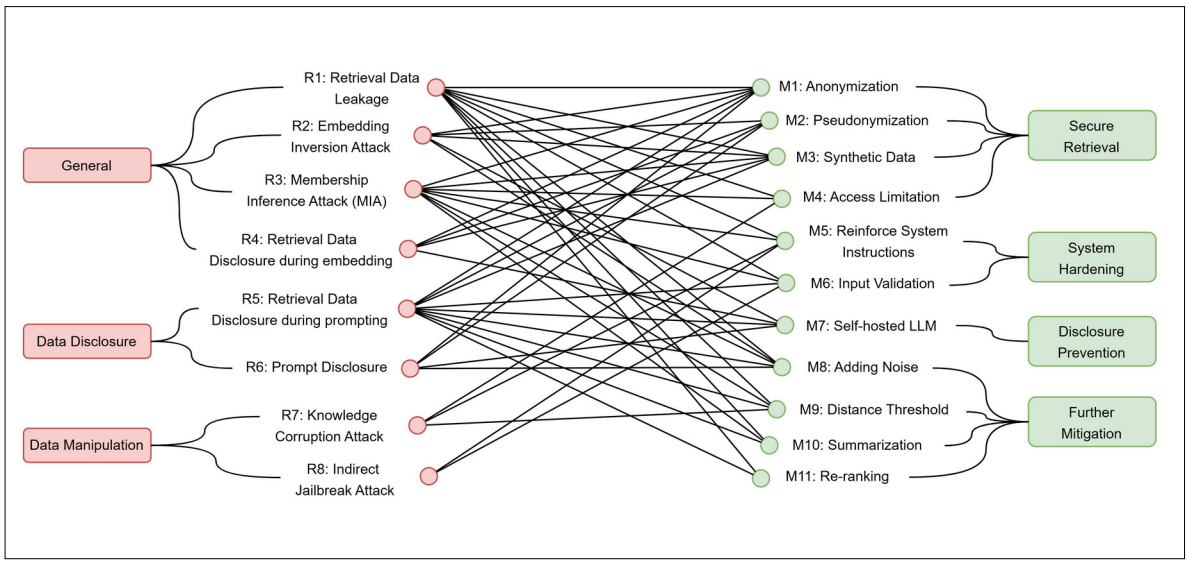
Basic RAG Pipeline
Own presentment



Decision Bubbles - what are your risks?
Support for an easy initial assessment of your RAG system
Own presentment



Risk & Mitigation Landscape -
Risks in red on the left, associated Mitigation Strategies in green on the right
Own presentment



Advisor
Prof. Dr. Marco Lehmann

Subject Area
Artificial Intelligence, Security