Development of a scalable and secure RAG-as-a-Service infrastructure

Graduate

Lukas Ammann

Sara Ott

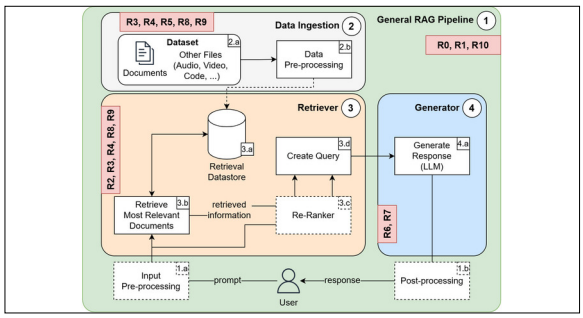
Introduction: Retrieval-Augmented Generation (RAG) systems are valuable tools for enhancing AI language models by incorporating real-time, specialized, or domain-specific knowledge. This integration improves accuracy and relevance, reduces hallucination, and unlocks numerous business applications such as improving customer service, streamlining operations, and supporting decision making.

Problem: The deployment of RAG systems introduces significant security and privacy challenges. These systems often process sensitive data, raising concerns about unauthorized access, misuse, and operational vulnerabilities. Without proper safeguards, organizations may face regulatory penalties, reputational damage, and financial loss.
Building and maintaining a secure, fully self-hosted RAG system is therefore a non-trivial endeavor. It requires expertise in various technical domains and a significant investment of time and resources to ensure secure and reliable operation. This challenge is particularly pronounced for small and medium-sized enterprises, which often lack the necessary infrastructure, personnel, and budget to implement such systems.

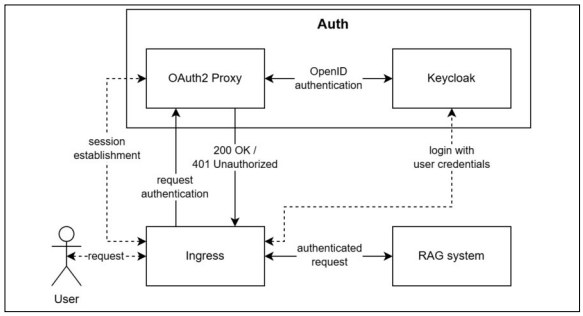
Result: We present a scalable, adaptable, and secure RAG-as-a-Service architecture, including the implementation of its core components. The system is built on a microservices architecture deployed on Kubernetes, with each component designed for modularity and scalability. To ensure security, we leveraged insights from our study thesis and our review paper and implemented a robust authentication system based on modern standards such as OAuth 2.0 and OpenID Connect.

Additionally, we established a three-system setup with corresponding tasks for our workshop at the IEEE Swiss Conference on Data Science (SDS2025), which will be held on June 26 in Zurich. This setup is a successful demonstration of the system's ability to operate effectively in an as-a-Service manner.

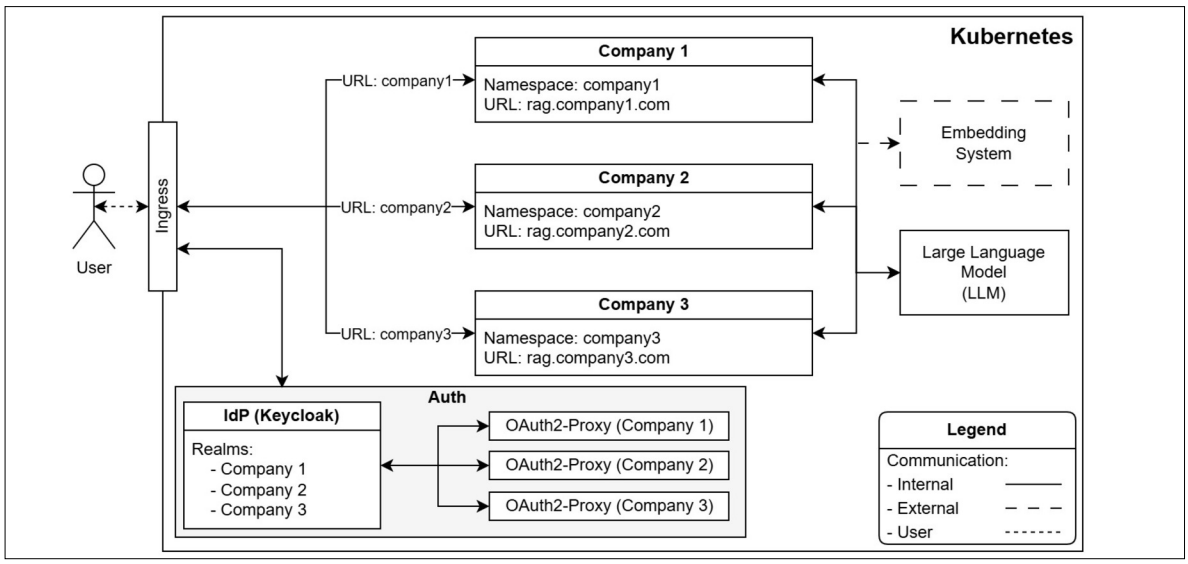
Overview of a General RAG Pipeline: Key Components and Associated Risks
Adapted from our Paper: https://arxiv.org/abs/2505.08728



Authentication Workflow
Own presentment



Architecture Diagram Showing Multiple RAG Systems Operating in a Shared Environment
Own presentment



Advisor
Prof. Dr. Marco Lehmann

Co-Examiner
Andreas Landerer

Subject Area
Artificial Intelligence