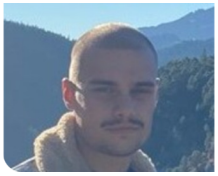
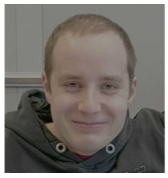
Design and Development of Retrieval-Augmented AI Assistants

Architecture and PoC Implementation of a Workflow-Oriented RAG Platform

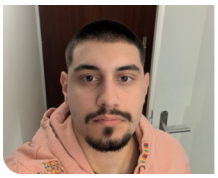
Students



Jovan Rakic



Mario Huber



Edo Husakovic

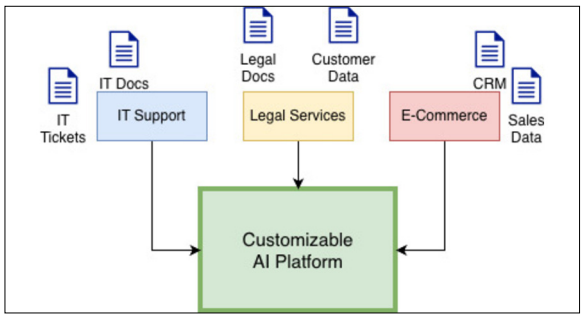
Introduction: RAG enables language models to answer questions using company-specific knowledge. This project builds on a scalable, secure RAG infrastructure with separated Client-, Company-, and Domain-level data sources.
The focus is on configurable AI assistants, structured workflows, and system connectors for enterprise.

Problem: To increase the value of AI assistants In business settings such as IT service providers, Information from multiple existing external systems, including internal documentation and ticketing platforms, need to be integrated.
Furthermore, prompt logic and assistant behavior must be reusable to remain maintainable across multiple use cases. For repeatable business processes, assistants must be guidable through predefined steps, as unrestricted conversational behavior leads to unpredictable results.

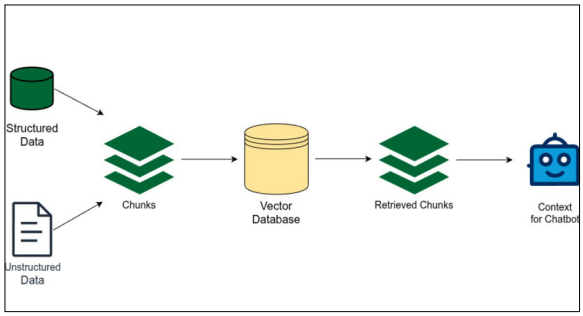
Result: The project introduces a PoC for a modular assistant platform that extends the existing RAG architecture with several key components.
A connector-based ingestion mechanism allows structured integration of external data sources. A modular tool architecture enables assistants to execute actions in external systems. Model Context Protocol (MCP) servers were integrated to support standardized external tools.
Additionally, a workflow system was implemented that defines task stages and restricts which prompts, data sources, and tools are available per stage, enabling controlled assistant behavior.

The approach was validated with a practical PoC for first-level IT support, including guided troubleshooting and automated ticket creation. Future improvements such as scalable evaluation were identified.

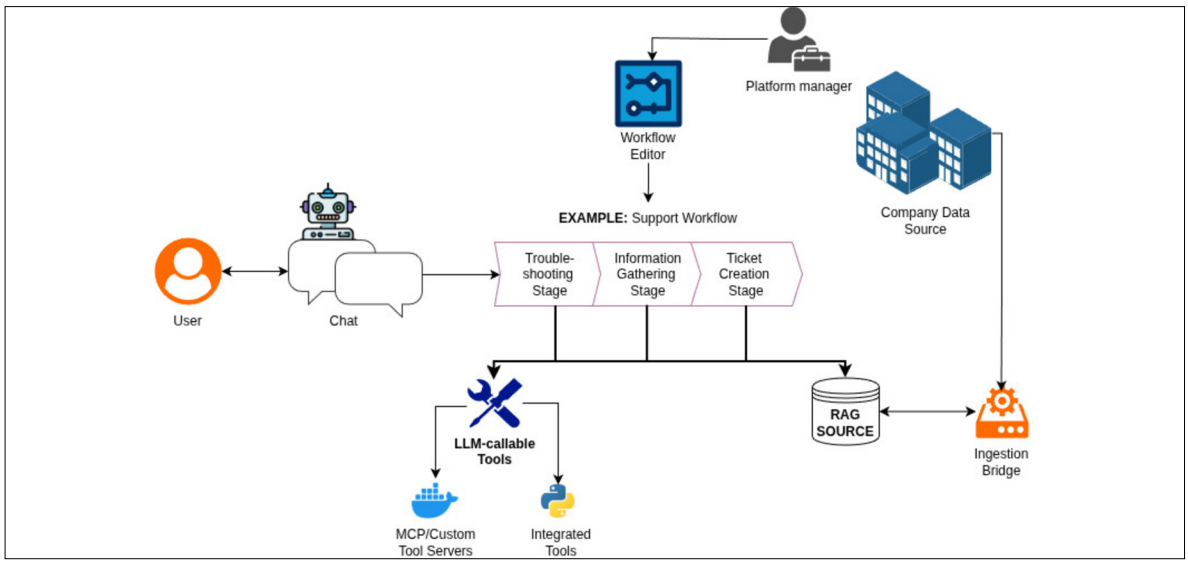
Operational AI Platform Overview
Own presentment



Retrieval-Augmented Generation Pipeline
Own presentment



Workflow of a first-level IT Support Assistant
Own presentment



Advisor
Prof. Dr. Marco Lehmann

Subject Area
Artificial Intelligence, Software Engineering, Network and Cloud Infrastructure