TrustinLLM

Trustworthy Digital Systems Assisted by Large Language Models (LLMs)

Richard Hohensinner[1], Roman Kern[2], Belgin Mutlu[1]

Pro2Future GmbH[1], TUG-IML (Institute of Machine Learning and Neural Computation)[2]

[1] Sandgasse 34, 8010 Graz, Austria

[2] Inffeldgasse 16b/I, 8010 Graz, Austria

Pro²Future

MOTIVATION & GOALS

As digital infrastructures increasingly incorporate Artificial Intelligence, particularly **Large Language Models (LLMs)**, ensuring their **trustworthiness** is essential. This project seeks to establish a foundational framework to strengthen trust in LLM-assisted systems by focusing on **Transparency**, **Reliability**, and **Safety**. We propose novel methodologies that enable the generation of **factually accurate responses** by explicitly integrating context from verified and reliable sources. These approaches will address both system-level attributes (e.g., architecture, robustness) and data-level qualities (e.g., integrity, bias), contributing to the development of AI systems that are **dependable**, **transparent**, and **ethically aligned**.

Project FactBox

Project Name TrustinLLM
Project ID FFG 915295
Duration 36 Months

Area 3
Area Analytics

Project Lead
DI Dr. Belgin Mutlu

APPROACH

Retrieval-Augmented Generation (**RAG**) is a promising approach to **mitigate hallucinations** in LLMs by **grounding responses** in source documents provided by users. This enables a fast and secure mechanism for validating the correctness of responses for users. This project enhances RAG through improved pre- & post-retrieval processing, **transparency** and **traceability** measures, and UI-based **interpretability features**.

CONTRIBUTION

Scientific contribution

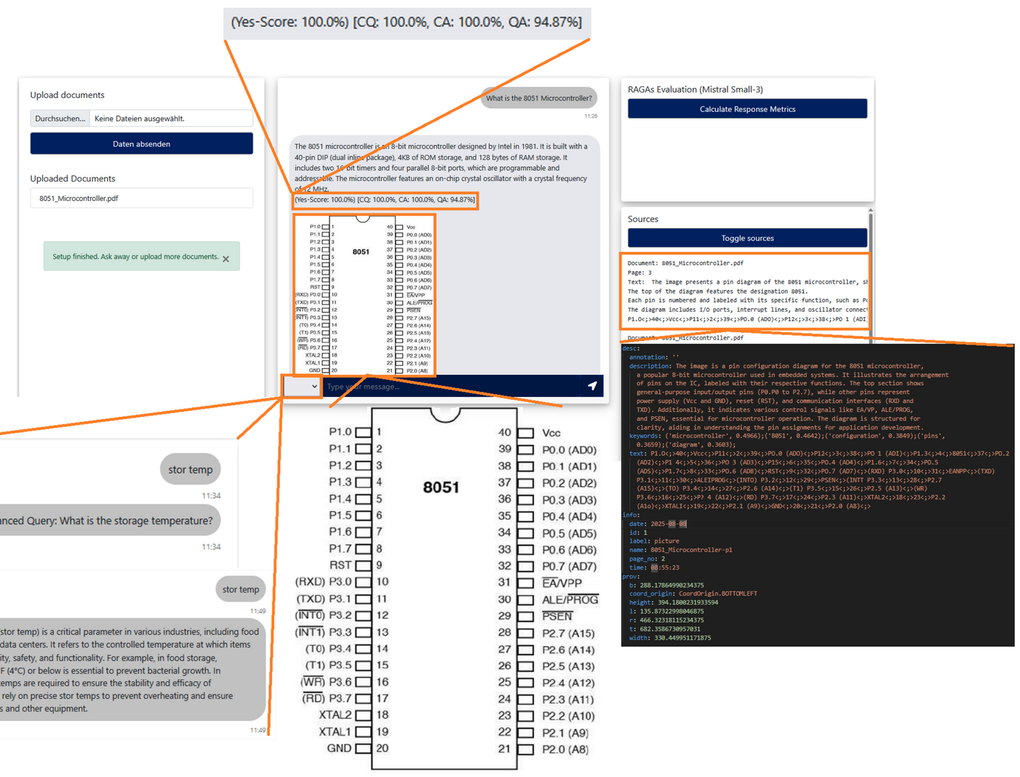
- Increasing transparency capabilities of LLM-based RAG systems
- Increased trustworthiness for AI-based systems
- Data provenance and traceability solutions to tackle rising data obfuscation
- Mitigation of hallucinations and deceiving responses from LLM inference

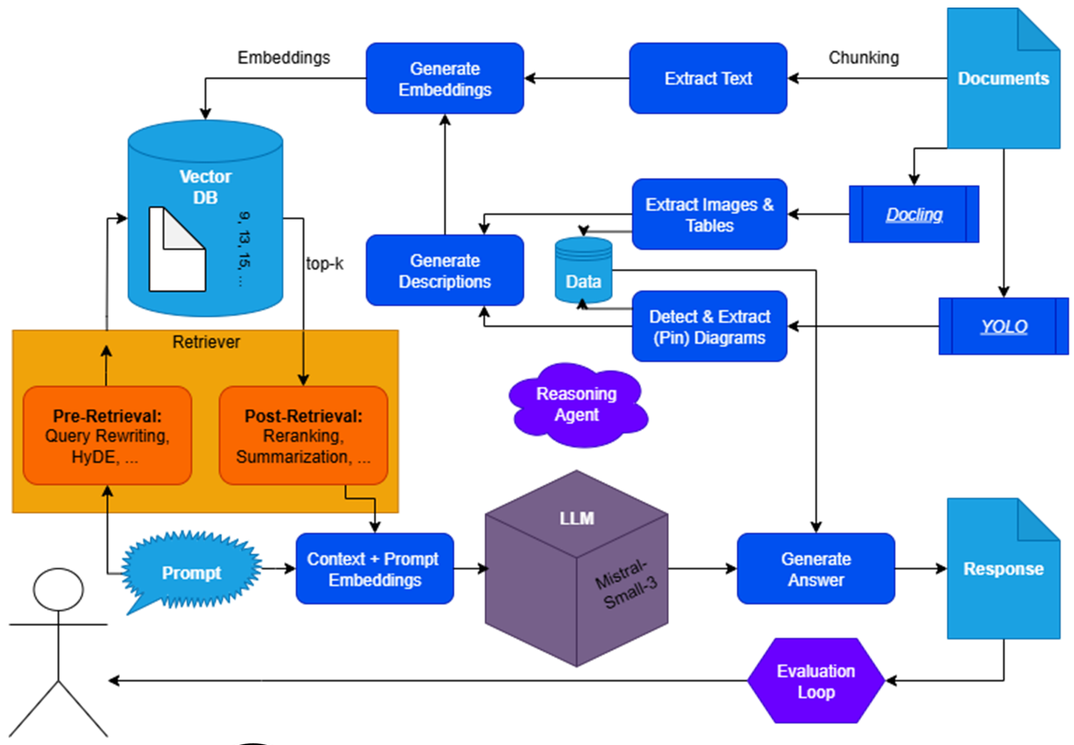
Economic contribution

- Lower risk of costly errors through reliable AI outputs
- Higher productivity via reduced manual oversight
- Faster adoption due to increased user trust
- Improved compliance with safety and transparency
- Competitive advantage through trustworthy AI offerings

SYSTEM ARCHITECTURE

Documents are uploaded onto a locally hosted RAG-Framework, pre-processed to extract **multimodal embedded items** and their textual content chunked into a vector store. Users can then ask questions, and the LLM-based system generates answers sourced from the uploaded documents providing a tracible way of information generation.





Contact: DI Richard Hohensinner, Pro2Future GmbH, richard.hohensinner@pro2future.at

Acknowledgement: This work was supported by Pro²Future II (FFG, 911655) and Robert Bosch AG.

BOSCH

JOHANNES KEPLER UNIVERSITÄT LINZ · TU Graz · AVL · Fronius · UAR Upper Austrian Research GmbH · Bundesministerium Innovation, Mobilität und Infrastruktur · Bundesministerium Wirtschaft, Energie und Tourismus · FFG Forschung wirkt. · ober österreich · Das Land Steiermark Wirtschaft, Arbeit, Finanzen, Wissenschaft und Forschung · SFG NEUES DENKEN. NEUES FÖRDERN.