Title

Building Agentic AI Systems: A Practical Guide with LLMs and Retrieval-Augmented Generation

Abstract

Large Language Models (LLMs) and Retrieval-Augmented Generation (RAG) are rapidly transforming how businesses interact with and utilise information. While LLMs, such as GPT and Claude, have achieved unprecedented performance in natural language understanding and generation, they often exhibit limitations, including hallucination, outdated knowledge, and a lack of interpretability. RAG addresses these challenges by combining LLMs with external information retrieval mechanisms, enabling systems to generate responses grounded in relevant, real-time data. This hybrid architecture follows a retrieve-then-generate paradigm, retrieving contextually relevant documents from a knowledge base and then using them to guide the generation of more accurate and trustworthy responses.

This tutorial offers a hands-on introduction to integrating LLMs and RAG into practical business workflows. Participants will explore the foundational concepts of tokenisation, embedding-based retrieval, vector databases, and prompt engineering. Emphasis is placed on applying RAG to knowledge-intensive applications such as customer service chatbots, intelligent document summarisation, and dynamic content generation. By incorporating up-to-date and domain-specific information into the generation process, RAG allows LLMs to deliver more reliable outputs for enterprise use cases.

Live demonstrations will guide attendees through building an end-to-end intelligent chatbot using LLMs, LangChain, and Streamlit, showcasing how RAG can be deployed with minimal infrastructure using cloud-based environments such as Google Colab. Attendees will receive codebases and implementation templates to replicate and customise in their workflows.

The session concludes with a discussion of implementation challenges, including latency, retrieval accuracy, and the ethical risks associated with automated decision-making. We also explore future research directions such as adaptive retrieval agents, integration with multimodal inputs, and responsible fine-tuning techniques to ensure fairness, transparency, and accountability.

Collectively, this tutorial equips participants with the practical tools and theoretical insights to harness LLMs and RAG for intelligent automation, enhancing the quality and efficiency of digital decision-making in modern enterprises.

Biographies

Charles Liu

A researcher at the Australian Artificial Intelligence Institute (AAII), University of Technology Sydney. His research lies at the intersection of intelligent computing and systems engineering, with a strong emphasis on applied AI and end-to-end autonomous systems. His interdisciplinary work spans smart agriculture, carbon intelligence, medical AI, business and financial intelligence, renewable energy, and large-scale integrated data infrastructures. In the domain of carbon neutrality, he proposed the KACINO system, an intelligent computing framework that enables the digitisation and integration of carbon-neutral assets across both physical and financial ecosystems, thereby bridging environmental performance with tokenised carbon credit infrastructures. With a focus on advanced computing technologies, he is dedicated to developing intelligent, adaptive systems by integrating real-time data, composable workflows, and ubiquitous AI architectures. His work aims to build trusted, transparent, and scalable decision-support systems for complex, data-driven environments.

Imani Abayakoon

Imani Abayakoon is a PhD student at the University of Technology Sydney, specialising in Large Language Models (LLMs), Retrieval-Augmented Generation (RAG), and Explainable Artificial Intelligence (XAI). Her research is dedicated to advancing the capabilities of AI systems, with a particular emphasis on generating outputs that are accurate, reliable, and interpretable.