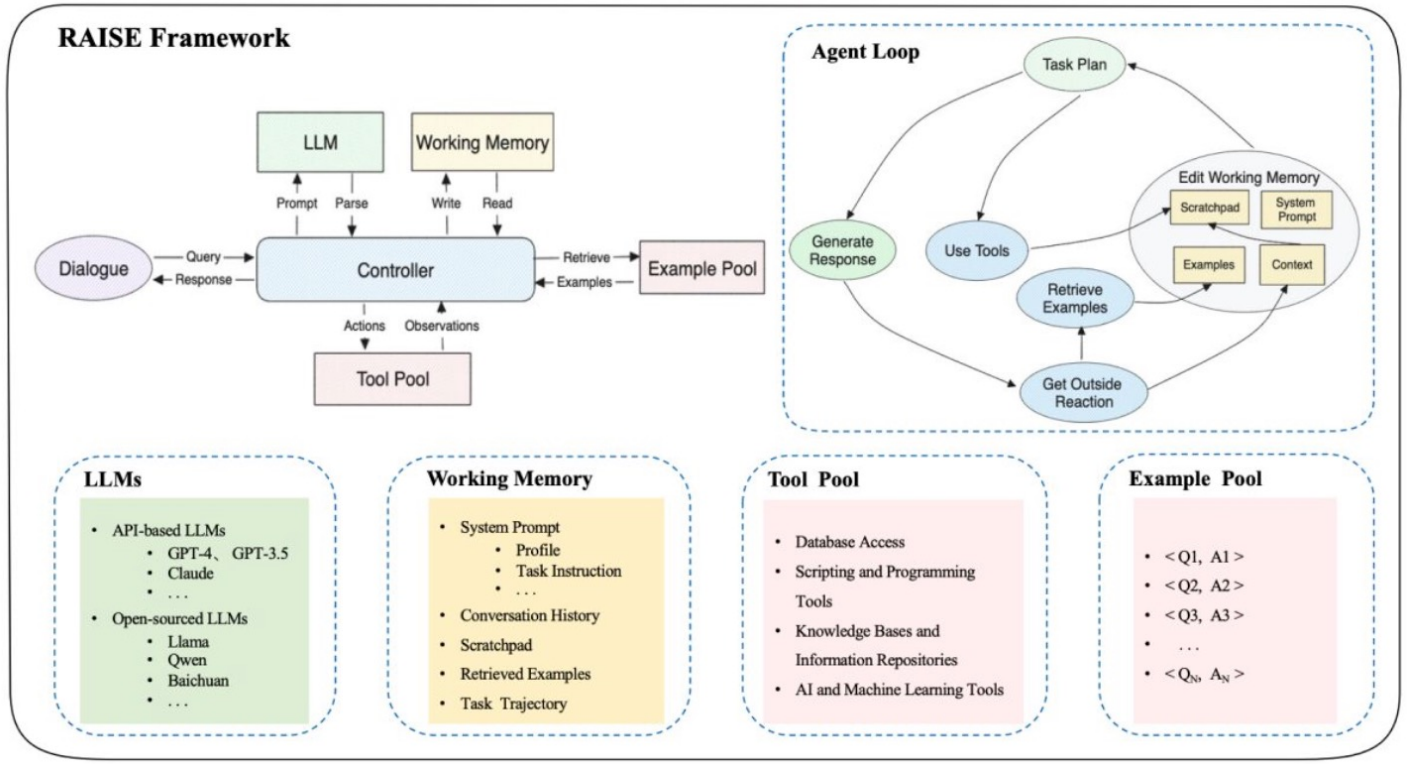
From LLM to Conversational Agent: A Memory Enhanced Architecture with Fine-Tuning of Large Language Models - Liu et al. 2024



From LLM to Conversational Agent

Proposes RAISE, an advanced architecture to enhance LLMs for conversational agents.

It's inspired by the ReAct framework and integrates a dual-component memory system.

It utilizes a scratchpad and retrieved examples to augment the agent's capabilities.

The scratchpad serves as a transient storage (akin to short-term memory) and the retrieval module operates as the agent's long-term memory. So you can think of this as a framework that combines ReAcT, a scratchpad, and a retrieval system.

This system mirrors human short-term and long-term memory and helps to maintain context and continuity which are key in conversational systems.

One benefit of such a system is the ability to customize and control the behavior of the conversational system.

This work also shows the potential to fine-tune the LLM within the RAISE framework which leads to enhanced controllability.

RAISE was tested on the real-estate domain and showed superior performance as compared to the conventional conversational agents.

For every use case, there might be a need to control for different aspects of an LLM-powered agent. This can range from customizing the working memory components to changing the effectiveness of retrieval of certain types of information.

I like the idea of modularity in a conversational agent but that means there are more variables to control. Lots of potential with this framework for building more tailored and personalized agents.