Chatbot Evolution at Swisscom Customer Support

Fatemeh Borran
fatemeh.borran@swisscom.com

Abstract

Large language models (LLMs), such as OpenAI's GPT-x, Meta's LLaMA, or Google's LaMDA are revolutionizing the world, especially since ChatGPT benefited from an immense press and social media coverage. These models claim to understand and generate human-like natural language conversations. In addition to producing convincing responses across different domains, the context awareness of these models makes them even more powerful. One can influence the response by providing contextual information and then ask the model to answer specific question in that context. Despite being one of the greatest advances in the field of Natural Language Understanding, using these models in production is NOT recommended. This is mainly due to unreliable content generated by these models and their non-deterministic nature. Thus, the question that we answer in this study is - how can industries benefit from those models? Conventional chatbots typically involve an intent recognition module and predefined flows specifically tailored for selected business scenarios. While these chatbots provide a considerable degree of control over flow execution, building a comprehensive customer care support across all business scenarios is often a laborious task. Conversely, generative bots relying on Large Language Models require little implementation effort at the expense of lack of control. When it comes to developing a customer support bot tailored for a particular case, one must either excel in prompt engineering or possess the necessary data and infrastructure for fine-tuning open-source GPT models, which is not affordable for most companies. The constraints on the token size and the prolonged response time of cloud based GPT models hinder the ability to encompass all business knowledge within a prompt. Retrieval Augmented Generation (RAG) combines both retrieval-based and generative methods to improve the performance of conversational AI systems, mainly, increasing quality by using most relevant information and reducing hallucination. RAG typically involves utilizing a Knowledge Base (KB) for retrieval and incorporating this information into the generative process. While RAG is a powerful approach in chatbot development, there are certain limitations, especially when it comes to using customer and dynamic data. In this case study, we show how Swisscom leverages LLMs to create values for productive systems. Following RAG methodology, we create a KB comprising detailed description of business scenarios along with their corresponding resolutions. We instruct LLM to call external APIs, use customer data, and execute specific actions as required in different situations. For a given customer request, we (1) retrieve data from the KB and (2) retrieve customer specific data (after asking customer to login), then (3) we use LLM model to (i) generate an answer and (ii) define next best action. By adopting this approach, we transition from the world of predefined scenarios in conventional chatbots to more scalable chatbot world with the capacity of handling unlimited scenarios with little implementation effort.