AskTelstra: An LLM-Powered QA System for an Enhanced Customer Support

Jack Latrobe, Andrea Chiechi, Danielle Nunns, Sirirat Wisessiri, Stuart Taylor, Vijayaraj Vinuvilla Rajan, Sing Teng, Sara Keretna, Sing Yatphaisan, Zhuohan Xie, Mudit Upadhyay, Anish Chelliah CR, Steve Morris, Rohit Lakhotia, Cat Smith
Telstra, Australia
Sara.keretna@team.telstra.com

Abstract

Large Language Models (LLMs) have opened the door to a wide range of Natural Language Processing (NLP) applications, including Question-Answering (QA) chatbots, by enabling more advanced and context-aware language understanding (Ray, P.P., 2023). In the corporate sector, enterprises started to leverage LLM capabilities to build their own QA chatbots on top of their internal knowledge. With minimal effort, these chatbots can understand questions, break them down, and give answers that mimic what a human would say, and the answers are usually of high quality.

Within this context, this proposal introduces "AskTelstra" an LLM-driven QA system designed to harness the wealth of information contained within Telstra's internal knowledgebase. The primary objective of AskTelstra is to empower customer support agents to efficiently address customer inquiries by providing them with a cutting-edge tool capable of accurately delivering responses. The proposal aims to showcase the innovative approach and promising potential of AskTelstra within the corporate sector.

1 Telstra

Telstra is Australia's leading telecommunications and technology company. It offers a wide list of services, ranging from mobile and broadband connectivity to entertainment services. Telstra serves a diverse customer base that includes individual consumers, wholesale partners, small and medium enterprises, large enterprises and governmental bodies. Telstra's operational reach extends beyond Australia's borders, with an international presence spanning over 20 countries, positioning it as an influential global player in the telecommunications sector (Telstra 2023).

Telstra has established itself as the leading player in the ever-evolving digital landscape. It remains committed to providing the best experience for its customers, leveraging recent tech advancements in the digital space. In this context, we introduce "AskTelstra," an innovative LLM-Powered QA system, which shows Telstra's commitment to enhancing the customer support experience. This proposal explores how LLMs can be used in the customer support context to make the process more efficient and relevant to individual customer inquiries.

2 AskTelstra

AskTelstra is an LLM-powered QA system that scans and indexes Telstra's public and non-sensitive knowledgebases and provide a chat interface for asking questions and getting answers based on specific references to articles or documents in the knowledgebase. It aims to provide customer support agents with a tool that can easily access all the knowledge base they need to provide them with a quick answer, supported by relevant citations regarding a specific inquiry.

The primary goals for AskTelstra are as follows:

- Enhanced Customer Experience: Provide users with a responsive and intelligent chatbot that can answer queries as accurately as possible and reference specific information from indexed documents.

- Scalability: Ensure the solution can handle a growing number of both customers asking questions, and new business units onboarding their own documents and knowledgebases.
- Security: Follow the existing security standards and controls platform to protect sensitive user and business data.

This solution leverages Azure Cognitive Search and GPT-3.5 capabilities to create a powerful, user-friendly chatbot that efficiently navigates knowledgebase articles and provides accurate answers to user queries.

The initial implementation focuses on indexing internal customer support knowledge base articles and deploying a chatbot accessible through a simple web user interface for internal employees.

2.1 Main Components and tools

The main components that form AskTelstra are illustrated in Figure 1. Ask Telstra is composed of five main components:

Storage: This is where the data that will be indexed reside. Data are curated and stored in an azure blob storage. This data represents the internal knowledge base that will be search and answers will be extracted from.

Search engine: token size limit is one of the main challenges for using LLMs. Exposing an entire knowledgebase to an LLM to answer an inquiry will not work because max token limit will be consumed very quickly, and no answer will be generated. That's why a powerful search engine needs to be used as an intermediate step to retrieve documents that are most relevant to the inquiry being analysed. The LLM will then only use the retrieved documents to generate an answer for the given inquiry. In AskTelstra, Azure cognitive search is used for that purpose. Bing enterprise search is also being utilized to enable the user to search Telstra.com.

Core Engine: This is the App service that forms the frontend and the backend of AskTelstra. The backbone of the frontend is ReactJS. The backend is highly dependent on the popular langchain library that helps modularizes the solution to make it scalable. Several concepts in langchain are utilized including agents, tools and chains for knowledge retrieval,

LLM: The LLM currently used to power the Core Engine is GPT3.5-turbo, but the solution can be adapted to work on other LLMs.

2.2 Challenges

The most common challenges from designing an LLM-based QA bot are:

- **Token size limits:** LLMs have token limits, this means that there is a maximum number of tokens that can be used per request for an LLM to generate an answer. If that number is exceeded, the LLM will throw an error. Prompts and context are counted as part of the total tokens used. If the text the contains the answer to the inquiry is long, the risk of exceeding the token limit increases. Token limits can lead to an incomplete or no answer at all (Wang, S. and Jin, P., 2023).

Hallucinations: LLMs tend to hallucinate information as they attempt to find a relevant answer (Telstra 2023.Telstra website, accessed 12 Sept 2023. <https://www.telstra.com.au/aboutus/our-company>

- Deng, J. and Lin, Y.) (Cao, Y., Li, S., Liu, Y., Yan, Z., Dai, Y., Yu, P.S. and Sun, L). At this stage, the generated answer is reviewed by a customer support agent before being used. In addition, prompt engineering is utilised in AskTelstra to put guard rails and reduce the chances of hallucinations, which proved to be very effective so far.

- **Data nature:** The data that AskTelstra is designed for has a challenging nature; it contains different articles about very similar topics that cover different contexts. Examples for that are prepaid vs postpaid articles where an inquiry can be very similar but depending on the context the answer can be very different.

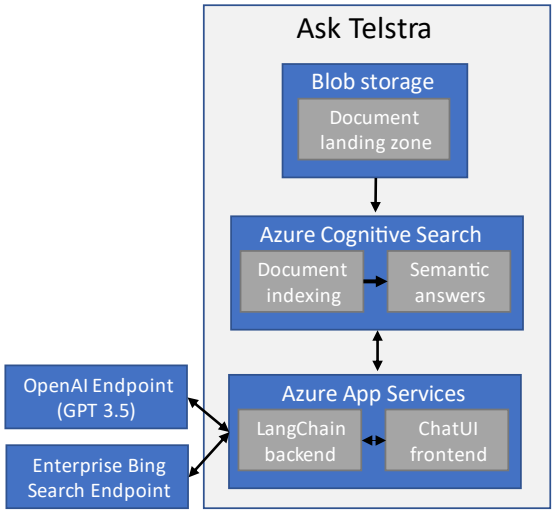

Figure 1: High level components for AskTelstra

2.3 Summary and future work

A beta version of AskTelstra is currently being released for customer support agents to test. Their feedback is being actively collected and acted upon. So far, the feedback has been very positive.

Future expansions of the solution will include integration with other systems such as SharePoint, Confluence, and public websites to export more relevant content to be able to answer wider range of queries.

References

Telstra 2023.Telstra website, accessed 12 Sept 2023. <https://www.telstra.com.au/aboutus/our-company>

Deng, J. and Lin, Y., 2022. The benefits and challenges of ChatGPT: An overview. Frontiers in Computing and Intelligent Systems, 2(2), pp.81-83.

Cao, Y., Li, S., Liu, Y., Yan, Z., Dai, Y., Yu, P.S. and Sun, L., 2023. A comprehensive survey of ai-generated content (aigc): A history of generative ai from gan to chatgpt. arXiv preprint arXiv:2303.04226.

Ray, P.P., 2023. ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope, Internet of Things and Cyber-Physical Systems 3 (2023) 121–154. doi: 10.1016/j. iotcps. 2023.04. 003. URL https://doi. org/10.1016/j. iotcps, 3.

Wang, S. and Jin, P., 2023. A Brief Summary of Prompting in Using GPT Models.