Context relevant Questioning for Better Answer Generation in Large Language Models

Chaitanya Chakka
chvskch@bu.edu

Muskandeep Jindal
mujindal@bu.edu

Abstract

In numerous scenarios, individuals encounter challenges in fully articulating their circumstances to AI models, resulting in incomplete or less-than-ideal answers due to the omission of critical details. This limitation constrains the contextual accuracy of the model's outputs. We propose a model that emulates the behavior of a human counselor by posing contextually relevant questions to capture missing information and enhance its understanding of the user's situation before delivering a final response. By employing this interactive methodology, the model not only augments the depth of its responses but also ensures that its advice is precisely tailored to the user's specific needs across various domains, such as visa applications, nutrition, and fitness.

1 Introduction

Human counseling services in domains such as immigration, nutrition advice, fitness coaching, and other personalized areas are highly sought after but often come with considerable costs in terms of time and financial resources. Expert human counselors provide invaluable insights by thoroughly comprehending the nuances of an individual's circumstances before offering tailored recommendations. This comprehensive interaction is crucial for delivering accurate and relevant advice, as it enables counselors to pose clarifying questions, gather specific details, and consider various facets of a person's situation that might otherwise be overlooked.

Despite advancements in AI systems designed to replicate such expertise, existing models frequently fail to capture the full complexity of a user's needs. These systems typically rely on the user's initial input, which may omit essential information due to the user's inability to articulate their situation completely. This lack of detailed understanding results in suboptimal responses, thereby diminishing the effectiveness of the AI system.

This paper explores the development of a model designed to bridge this gap by engaging in dialogue with users to obtain a more comprehensive understanding of their circumstances. With this enhanced information, the model can refine its responses to better align with the individual's unique situation. The contributions of this research include:

1. Design a conversational model that can self-assess to ask questions to the user if needed and give more context-relevant answers to the original question.

2. Tabularize results by experimenting with different large language models and compare them with different existing architectures on various evaluation grounds.

The rest of the paper is organized as follows: Section 2 covers the impact of such a model in real-life situations. With the few domains in consideration, we brief on the available datasets for those domains in Section 3. Section 4 covers the proposed model architecture and the various components involved. It also briefs on the basic training paradigm we plan to follow to learn the weights of the model.

2 Motivation and Impact

Implementing a model capable of engaging in follow-up dialogue to gather contextual information before delivering responses has the potential to significantly impact the general population. It can enhance the accessibility of expert advice in critical areas such as immigration, healthcare, legal assistance, education, and financial planning. Individuals who lack access to personalized professional guidance due to geographic, economic, or social barriers can benefit immensely from this technology. Empowering users to receive tailored, context-aware advice without the need for prolonged in-

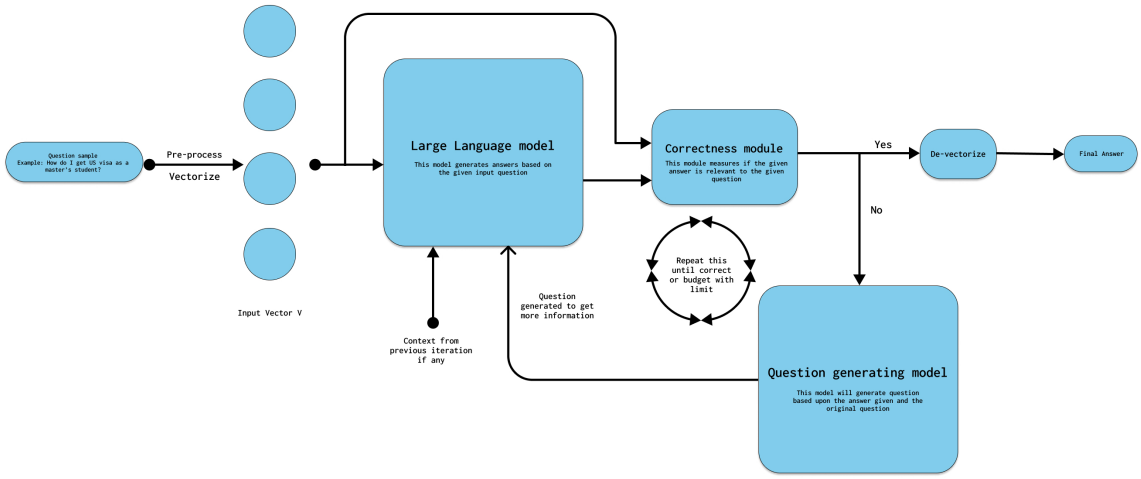

Figure 1: Question Answering model architecture to ask follow-up questions

teractions with human professionals, streamlining processes that often require multiple appointments and significant waiting times.

Additionally, the model's ability to ask clarifying questions would mitigate the risk of incomplete or inaccurate responses, ensuring that the user receives more reliable and actionable information. This could help in preventing missteps in visa applications, ensuring personalized nutrition and fitness plans, or offering timely medical advice—all of which contribute to improved quality of life and better decision-making. On a broader scale, the automation of these advisory tasks could free up human professionals to focus on more complex cases, creating a more efficient ecosystem where AI and human expertise complement one another.

3 Dataset Description

The architecture proposed in this paper is a generalized question-and-answer system. Assuming a good quality dataset is available, this model can be used in many applications like Visa inquiry, travel and trip advising, food and nutrition, etc. The subsequent sections determine datasets for each domain and the final domain will be decided in the upcoming milestone.

3.1 Visa and Immigration queries

A study on the effects of Immigration policies on the dynamic of the inflow and outflow of people into the country has shown that complicated policies significantly decrease as the complexity of the restrictions increases(Czaika and De Haas, 2017). This shows the rising importance of conversational agents for answering overseas travel questions.

We have extracted the travel questions from the famous Stack Exchange forum and filtered the data related to visas and abroad traveling. This resulted in around ten thousand posts that contained the question and various answers given by the community. The unfiltered dataset can be found here.

3.2 Healthcare

With the rising awareness of diseases and their cures and an elevated medical literacy in common people, there is a steady increase in healthcare content like WebMD. However, there is always a chance for misinformation, and therefore trustworthy and up-to-date sources are important. A conversational agent for answering any queries regarding healthcare would help patients make informed decisions.

(Ben Abacha and Demner-Fushman, 2019) have devised a question-answer dataset called MedQuAD which contains about forty-seven thousand questions related to different medical conditions and answers backed up by genuine sources like the National Institutes of Health(NIH). The dataset can be found here.

4 Research framework and Methodology

This section describes the model architecture and the training framework that we employ in this paper. The model consists of two large language models each tuned for a different purpose. We first clean the dataset with the standard pre-processing techniques - tokenization, stopword removal, lemmatization, and removal of non-alphanumeric characters. We then pass the question of each data point into a large language model which will be tuned to generate answers in the relevant field. Next, we will pass the output vector into the confidence module which will take both the question and answer and predict whether the answer is context relevant to the question. Based on the result of this module, we will pass the output to another large language model which is tuned to generate questions based on the answer given by the first model. This will be then passed back into the first model to generate a different answer. This cycle is carried on until either the module outputs a relevant answer or a certain number of questions have been already asked. This architecture has been shown in figure 1.

References

Asma Ben Abacha and Dina Demner-Fushman. 2019. A question-entailment approach to question answering. *BMC bioinformatics*, 20:1–23.

Mathias Czaika and Hein De Haas. 2017. The effect of visas on migration processes. *International Migration Review*, 51(4):893–926.