Closed Generative QA with Follow-up questioning in Large Language Models

Chaitanya Chakka[1], Muskandeep Jindal[1]

Boston University, Massachusetts, USA[1]

Introduction

Users often struggle to fully convey their circumstances, resulting in incomplete and inadequate AI responses. To overcome this, we propose a conversational AI model that emulates human counselors by posing clarifying questions to identify and address gaps in the provided information. This approach enables the delivery of personalized and contextually accurate responses, which is particularly valuable in domains like immigration, healthcare, and fitness, where accurate advice relies on a thorough understanding of the user's circumstances.

The proposed methodology includes the development of a self-assessing framework that identifies incomplete inputs, formulates appropriate follow-up questions, and integrates the additional context into its responses. This model is tested in the Visa and Immigration querying to demonstrate the adaptability of the framework.

Motivation

The growing demand for expert guidance is often hindered by the high cost and time constraints associated with human counseling. This challenge disproportionately affects individuals with limited access to personalized services due to geographic or economic barriers.

Our study aims to enhance the accessibility and precision of AI-driven counseling services by providing reliable, context aware guidance in a scalable and cost-efficient manner. Focused on immigration-related data, we design a model that refines user interactions by using targeted questions to close information gaps, delivering precise and personalized guidance.

User Question: What is required of refugees and asylees after one year in the U.S.?

LLM Generated Response: Both refugees and asylees are required to apply for adjustment to Lawful Permanent Resident (LPR) status after one year of continuous presence in the US.

Generated Follow-up: What specific information do you need regarding application process of adjustment to Lawful Permanent Resident (LPR) as a refugee or asylee in the US?

User Response: I would like to know more about the required forms, documents, and fees to guide me through the adjustment process to LPR status as a refugee in US

Figure 1: Sample conversation with proposed QA model

Research Objectives

Our AI model is composed of three key modules: an LLM for generating responses, a Correctness Module to ensure answer completeness, and a Follow-Up Question Generator to prompt for additional context when necessary.

- RQ_1: What are the baseline performance metrics for the conversational AI system before fine-tuning?
- RQ_2: Implement all three modules and fine-tune them on our collected dataset
- RQ_3: What are the performance metrics for the conversational AI system after fine-tuning?

Closed Domain QA Dataset

One of the biggest sources of question answering is the stack exchange online forum where users post various queries and people try to answer them. All the data across the forums are publicly made available via a Big Data Explorer. We have devised a BigQuery to extract the desired question-answer pairs by checking if the question involves related terms like visa, immigration, citizenship and passport.

This query resulted in the extraction of around 26,000 question answer pairs with each question having at least one answer. We have also extracted the number of votes for the answer so that, it can be used to estimate a confidence value for the proposed architecture. The dataset has been preprocessed to remove HTML tagging and simple preprocessing like lowercasing and special character removal. A sample of the dataset is given below:

Sample Dataset Point

Question: Although I had requested an extension before my visa expired, I have not heard back from the government. I need to go to Italy in August, & wish to return to NYC in September. Will I be allowed to return if my visa has not been renewed by then?
Answer: If you want to return in M-1 status, you'll need a new M-1 visa. This is true even if your extension request is granted...
Tags: italian-citizens, student-visas
Vote Score: 6

The following are a few statistics concerning votes for each post. This helped us in choosing a heuristic for the confidence level threshold:

```
Min vote_score: -12
Max vote_score: 219
Mean vote_score: 6.011221320508967
Std vote_score: 9.93655226706058
```

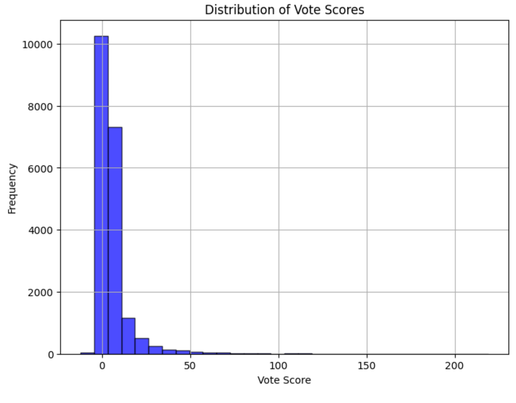

Figure 2: Distribution of Votes for all the data points

Model Framework

This architecture employs an iterative process for question answering with the following components:

- **Preprocessing Module**: User inputs are preprocessed and vectorized into an input vector V.
- **Large Language Model (LLM)**: Generates an answer based on V and any prior contextual information.
- **Correctness Evaluation Module**: Evaluates the answer's reliability by assigning a confidence score. If the score exceeds a threshold, the answer is finalized. Otherwise, follow-up generation is triggered.
- **Question Generation Module**: Formulates follow-up questions to gather additional user context.
- **Iterative Loop**: The new user response, along with previous inputs and answers, is fed back to the LLM. The cycle repeats until the confidence score meets the threshold or a maximum iteration limit is reached.

This iterative design ensures adaptive refinement and context-aware answers. The same has been shown in figure 3.

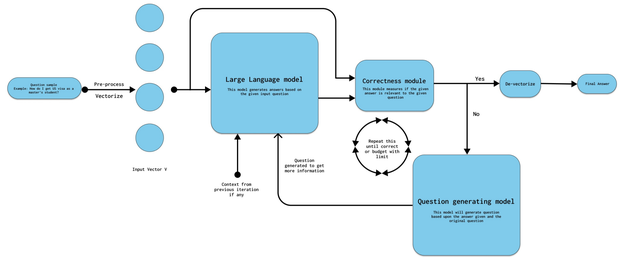

Figure 3: Conversation AI model architecture with follow-up questions generation functionality

Experimental Setup

For a baseline set-up, we have chosen a standard FLAN-T5 model available on Hugging Face. We have also fine-tuned the FLAN-T5 model and tested on the dataset for more robust comparison. For the proposed architecture, we have implemented the FLAN T5 model for generating the answer and, fine-tuned BERT as regression model for generating confidence score. To generate a follow-up question based on confidence level, we have used ChatGPT using the openAI library. Since each module has different tasks, we have fine-tuned each model separately and combined them into a single pipeline to perform the final inferencing. The correctness module is designed with BERT attached with a classification head consisting of a multi-layer perceptron with 2 hidden layers of 512 and 64 neurons each. The first layer has Tanh activation while the second one uses ReLU. The hyperparameters were optimized through multiple rounds of validation testing, using a 65-15-15 train-dev-test split.

Results and Findings

The training and validation losses during the epochs have been shown in 1. The model has been checkpointed at regular intervals and based on these plots, the final models at taken at epoch 20 and epoch 10 for the correctness model and QA model respectively.

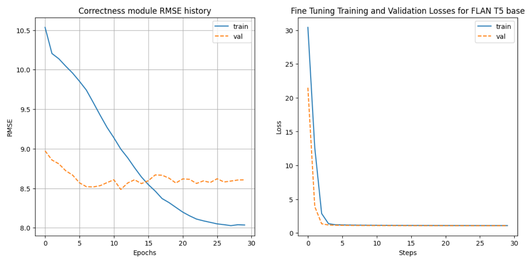

Figure 4: Left: RMSE loss history of the correctness module over 30 epochs. Right: Cross Entropy Loss history of FLAN T5 model on question answering

The BLEU and ROGUE scores are used to evaluate the predictions and findings are tabulated in table 1.

Model	BLEU	ROUGE-1	ROUGE-2	ROUGE-L
Standard FLAN T5	2.166e-06	0.1209	0.02421	0.0915
Fine-tuned FLAN T5	2.167e-06	0.1208	0.02425	0.0917
Our Model	-	0.1306	0.02513	0.0957

Table 1: Comparison of BLEU and ROGUE scores of different experiments

Future Scope

We aim to explore the following areas in the future, given more time and resources.

1. To gather more reliable data for immigration questions to improve model's performance, reduce bias, and enhance the accuracy of generated answers
2. With improved data quality, we can use Retrieval Augmented Generation(RAG) model to extract contextually relevant information before the generative answering
3. Training the model on QA pairs from diverse domains
4. Currently, the model uses the OpenAI API to generate follow-up questions, but in the future one can fine-tune a model like T5 to generate follow-up questions.
5. Create pipelines for continual learning, where the model can be periodically fine-tuned with new data.

Conclusion

The proposed architecture has shown a marginal improvement over the baseline models for closed generative question answering. This emphasizes the importance of context-awareness in question answering, where the architecture's ability to self-evaluate by generating follow-up questions has enhanced its performance on the ROUGE and BLEU evaluation metrics.