A Smart Q&A System for Modern Tourism Using GNNs and LLMs

VIVA VOCE

Presentation by,
Aditi Kannan (203002004)
Dhasharadharami Reddy Bommana (203002021)

Guided by,
Dr. Actlin Jeeva
Assistant Professor, ECE Dept

OUTLINE

- Objective
- Motivation
- Literature Survey
- Analysis of available data
- Proposed Method
- Components required
- Dataset Creation

OUTLINE

- Pre-Processing steps for RoBERTa
- Fine-Tuning RoBERTa
- Automatic Speech Recognition
- Testing
- User Interface
- Demo
- Results and Conclusion
- References

OBJECTIVES

- Develop a smart question and answer system for modern tourism, utilizing LLMs.
- Provide accurate responses to traveler queries, offering personalized information on tourist attractions, and other cultural insights.
- Enhance tourism experiences, and increase revenue and GDP through informed and customized travel interactions.

MOTIVATION

- As of 2021, tourism generated ₹13.2 lakh crore (US$170 billion) or 5.8% of India's GDP and supported 32.1 million jobs. It is the largest growing sector.
- As the demand for personalized travel experiences and modern amenities increases, the need for an advanced Q&A system to satisfy tourists' needs becomes crucial.
- Moreover, catering to diverse traveler needs promote international engagement.

LITERATURE SURVEY

Sl. No	Title of the paper	Authors & year of publication	Techniques used	Inference
1.	***Graph Neural Networks for Natural Language Processing: A Survey***	L. Wu, Y. Chen, K. Shen, X. Guo, H. Gao, S. Li, J. Pei, and B. Long (2021)	The survey explores the growing use of Graph Neural Networks (GNNs) in Natural Language Processing (NLP) for graph-related tasks	A variety of problems in NLP can be best expressed with graph structure. The paper organizes GNN research, covering graph construction, representation learning, and encoder-decoder models, providing a comprehensive overview of this emerging field.
2.	***QA-GNN: Reasoning with Language Models and Knowledge Graphs for Question Answering***	Yasunaga, M., Ren, H., Bosselut, A., Liang, P., & Leskovec, J. (2021)	Relevance scoring to estimate Knowledge graph node importance and joint reasoning through graph neural networks	Using language models (LMs) for question answering faces challenges in extracting relevant knowledge from large knowledge graphs (KGs) and performing effective joint reasoning with the QA context and KG. The paper addresses these challenges by employing relevance scoring for KG nodes based on the QA context, resulting in improved performance, interpretability, and structured reasoning compared to existing LM and LM+KG models.

LITERATURE SURVEY

Sl. No	Title of the paper	Authors & year of publication	Journal /Conference Name	Techniques used	Inference
3.	***An Interpretable Question Answering Method Based on Heterogeneous Graph Neural Networks***	Y. Wu, Q. Zhou, H. Yin and D. Liu (2022)	2022 International Conference on Computer Engineering and Artificial Intelligence (ICCEAI)	IQAGNN, Heterogenous Graph Neural Networks, Multi-Type entity embeddings, Entity Correlation	The paper introduces IQAGNN, an interpretable question answering method using Heterogeneous Graph Neural Networks for predicting relevant answers and offering explanations. IQAGNN utilizes a heterogeneous network and graph neural networks for entity embeddings, yielding improved interpretive question answering over baseline methods in community-based scenarios.
4.	***Mutually improved dense retriever and GNN-based reader for arbitrary-hop open-domain question answering***	Li, Ronghan ,Wang, Lifang , Jiang, Zejun & Hu, Zhongtian , Zhao, Meng & Lu, Xinyu. (2022)	Neural Computing and Applications, ResearchGate	OpenQA, Dense Passage Retriever,Asynchronous Multi Grained Graph Network	The paper extends the GNN-based reader AMGN for multi-hop OpenQA. It introduces a training strategy improving dense retriever and AMGN mutually. Results demonstrate the effectiveness of asynchronous reasoning, retrieval-similarity attention, and reader-score guided MIPS for interpretable multi-hop reading.

LITERATURE SURVEY

Sl. No	Title of the paper	Authors & year of publication	Journal /Conference Name	Techniques used	Inference
5.	***Relevance-guided Supervision for OpenQA with ColBERT***	Omar Khattab, Christopher Potts, Matei Zaharia (2021)	Transactions of the Association for Computational Linguistics, Volume 9	OpenQA, ColBERT, Iterative Weak Supervision Strategy	The paper introduces ColBERT-QA, adapting the neural retrieval model ColBERT for Open-Domain Question Answering (OpenQA), significantly improving retrieval performance through fine-grained interactions between questions and passages. Efficient training strategy iteratively employs ColBERT for creating training data, achieving state-of-the-art extractive OpenQA performance.
6.	***Exploratory Inference Chain: Exploratorily Chaining Multi-hop Inferences with Large Language Models for Question-Answering***	S. Haji, K. Suekane, H. Sano and T. Takagi, (2023)	IEEE 17th International Conference on Semantic Computing (ICSC) , 2023	Multi Hop Inference, Exploratory Inference Chain (EIC) framework	The Exploratory Inference Chain (EIC) framework enhances question answering by combining implicit language model processing with explicit inference chains. Experimental results validate EIC's superiority over existing approaches in multi-hop QA, achieving more accurate and knowledge-rich answers.

LITERATURE SURVEY

Sl. No	Title of the paper	Authors & year of publication	Journal /Conference Name	Techniques used	Inference
7.	***IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages***	Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar (2020)	Findings of the Association for Computational Linguistics: EMNLP 2020	NLP resources for 11 major Indian languages, ALBERT Based Model, IndicGLUE Benchmark	The resources encompass 8.8 billion tokens from news sources, FastText-based word embeddings suitable for Indian languages' morphological complexity, compact ALBERT-based language models, and various NLU evaluation tasks. These resources aim to advance Indic NLP research and facilitate evaluation across a diverse set of languages, benefiting over a billion people

ANALYSIS OF AVAILABLE DATA

- The data used is from the Samanantar dataset from AI4Bharat website. It is the largest publicly available parallel corpora collection for Indic languages: Assamese, Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Oriya, Punjabi, Tamil, Telugu. The corpus has 49.6M sentence pairs between English to Indian Languages.
- The dataset consisted of sentences without a label indicating a domain, so after necessary text preprocessing steps and using an unsupervised approach using BERT Word Embedding (GLoVe) to represent each word as an array and then summarize each sentence by taking the mean of the word embeddings.
- Then using cosine similarity on predefined labels, a label was assigned to the sentence with the label with most similarity.

	sentences	cleaned_sentences
0	Have you heard about Foie gras?	heard foie gras
1	I never thought of acting in films.	never thought acting film
2	Installed Software	installed software
3	A case has been registered under Sections 302 ...	case registered section 302 376 ipc
4	Of this, 10 people succumbed to the injuries.	10 people succumbed injury
...	...	...
995	It has a post and telegraph office.	post telegraph office
996	Who's the lovely?	who lovely
997	Director: Amit Roy	director amit roy
998	Youth arrested on theft charges	youth arrested theft charge
999	350 crore .	350 crore

1000 rows × 2 columns

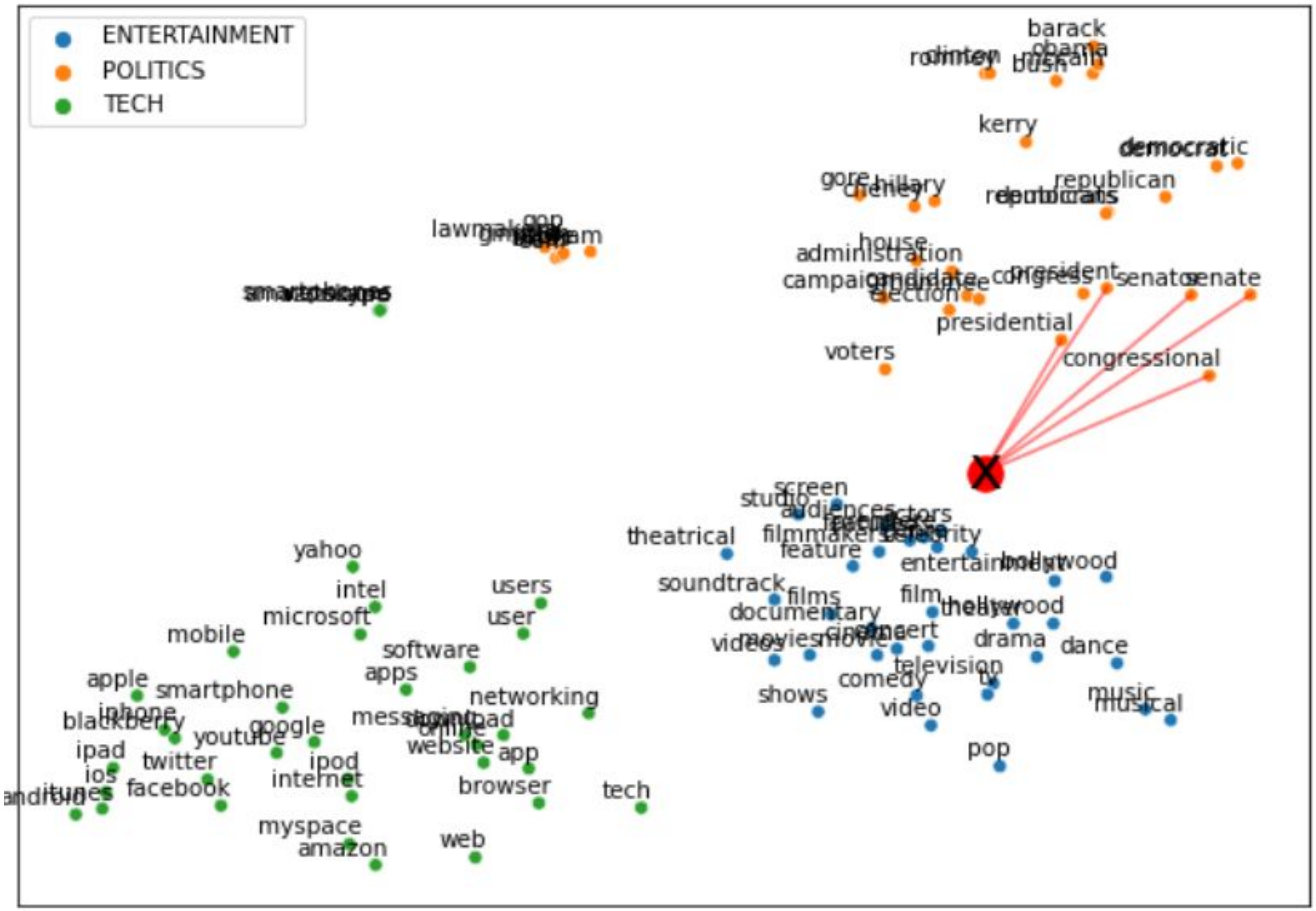
sentences	cleaned_sentences	0_x	1	2	3	4	5	6	7	...	759	760	761	762	763	764	765
The market valuation of ICICI Bank advanced by...	market valuation icici bank advanced 54178 cro...	0.202042	0.025990	0.257640	-0.311033	0.590954	0.166579	0.131796	0.544231	...	0.287364	0.431821	-0.577444	0.479203	-0.195919	-0.092383	-0.219508

- But this is not a very good way of processing embeddings as the sentences lose meaning in the process. So a special method was required to create a sentence's embedding as it is along with semantic considerations.

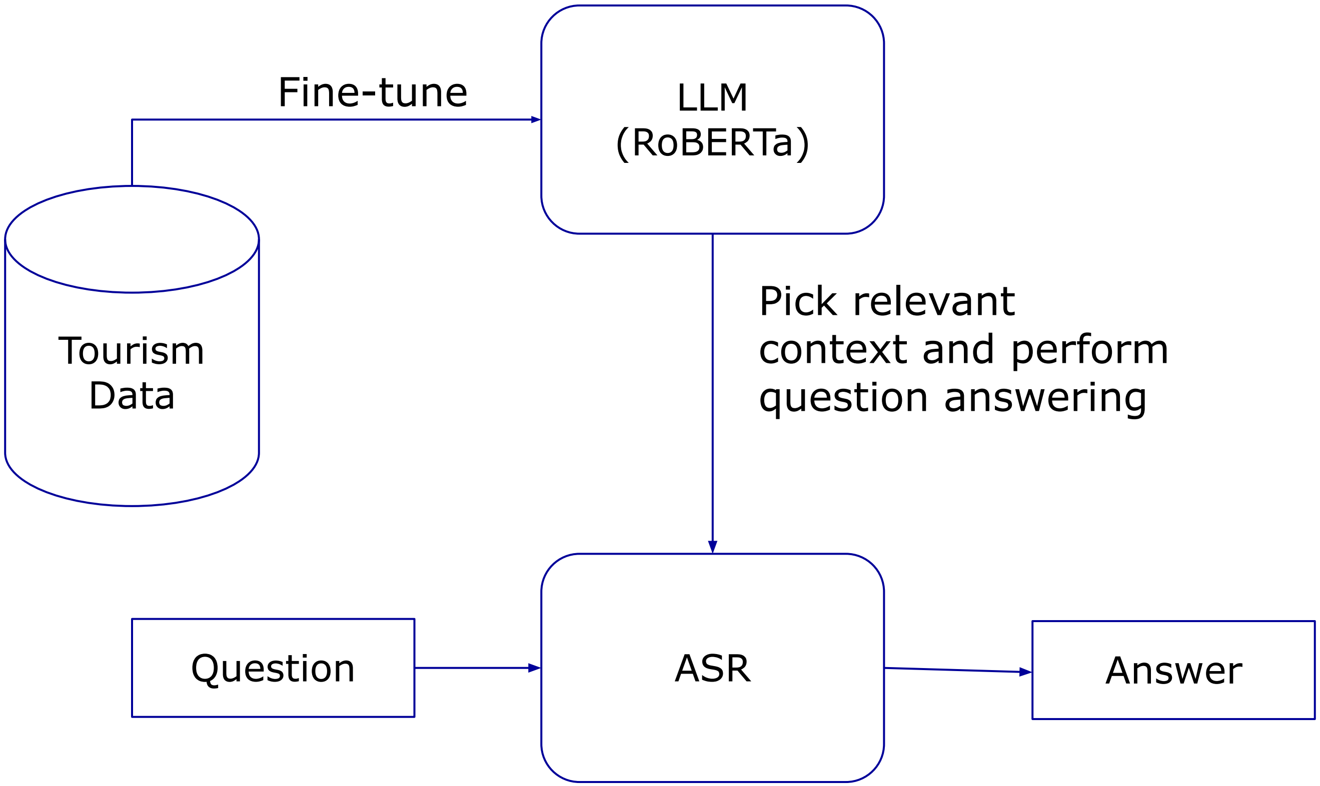
- So what was used is sentence transformers using **SentenceTransformer**. There are two types of models used,
  - all-MiniLM-L6-v2 => model size = 80MB, dimensions = (384,)
  - all-mpnet-base-v2 => model size = 420MB, dimensions = (768,)
- It is pretty evident that the larger model performs better with better and much more detailed embedding for each sentence.
- all-MiniLM-L6-v2 embeddings clusters:

```
A case has been registered under Sections 302 and 376, IPC.
Of this, 10 people succumbed to the injuries.
The incident was recorded in the CCTV footage.
Woman killed in stampede
He was immediately rushed to a nearby hospital where doctors declared him dead, the official said.


Australian batsman David Warner.
On the second day of Navratri, Maa Brahmacharini is worshipped.
Unlike in Gujarat State, alcohol is legal in Diu.
Rs 6.02 crore.
Chief Minister YS Jagan Mohan Reddy clearly stated that the building was constructed in blatant violation of all laws and regulations, hence it should be de


I never thought of acting in films.
Her acting has been praised by critics.
The movie also stars Kajal Agarwal in a prominent role.
Salman Khan will be essaying the role of circus artist in the film.
Starring Amitabh Bachchan and Ayushmann Khurana, Gulabo Sitabo is directed by Shoojit Sircar.


Installed Software
The Bibles viewpoint on this is clearly indicated at Colossians 3: 9: Do not be lying to one another.
5 lakh would be provided.
Education institutions are closed across the country in the wake of lockdown due to coronavirus outbreak.
The smartphone was recently launched in Indonesia.


Have you heard about Foie gras?
Respect privacy
Super Bowl.
It cannot work.
A few days ag...
```

DOMAIN CLASSIFICATION FOR SAMANANTAR DATASET



PROPOSED METHOD



COMPONENTS REQUIRED

- Python and the required libraries such as:
  - **numpy**
  - **pandas**
  - **huggingface_hub**
  - **sklearn**
  - **transformers**
  - **speech_recognition**
  - **gtts**
  - **pyaudio**
  - **sentence_transformers**
  - **tkinter**

DATASET CREATION

Feature	Description
Source	Wikipedia, Goibibo, Makemytrip, TN Govt. Websites
Tourist Attractions	31 tourist attractions in Chennai only
Format	CSV
Training Dataset Rows	270 rows of data
Test Dataset Rows	50 rows of data

DATASET CREATION

- Attributes: "review" (context/paragraph) and "qas_id" , "question," , "human_ans_spans." and “human_ans_indices”
- "qas_id": Unique identifier for the question.
- "question": The question being asked.
- "human_ans_spans": correct answer to the question.
- “human_ans_indices”: where the answer starts and ends in the context paragraph

LIST OF TOPICS AND NO. OF QUESTIONS FROM EACH TOPIC

S.No	**Topic**	**No. of questions from each topic**
1	Chennai Culture	14
2	Guindy National Park	12
3	Vandalur Zoo	11
4	Madras Crocodile Bank	9
5	Vedanthangal Bird Sanctuary	13
6	Adyar Eco Park	11
7	Marina Beach	14
8	Kovalam Beach	7
9	Elliot's Beach	6
10	Blue Flag Beach Muttukadu	8

LIST OF TOPICS AND NO. OF QUESTIONS FROM EACH TOPIC

S.No	Topic	No. of questions from each topic
11	Muttukadu Boathouse	7
12	Fort St. George	13
13	Government Museum Chennai	10
14	DakshinaChitra Heritage Museum	12
15	Chennai Rail Museum	10
16	Birla Planetarium	8
17	Chennai Lighthouse	7
18	Valluvar Kottam	13
19	Chepauk Stadium	9
20	Napier Bridge	8

LIST OF TOPICS AND NO. OF QUESTIONS FROM EACH TOPIC

S.No	Topic	No. of questions from each topic
21	MGR Film City	12
22	Shopping Malls	14
23	T Nagar	9
24	Parry's Corner	5
25	Moore Market	7
26	Santhome Church	11
27	Annai Vailankanni Shrine	8
28	Thousand Lights Mosque	6
29	Kapaleeshwarar temple	12
30	Parthasarathy Temple	14
31	Mahabalipuram	13
32	Shore Temple	7

	A	B	C	D	E
1	qas_id	review	question	human_ans_spans	human_ans_indices
2	q1	Chennai, form	What is Chennai formerly k	Madras	(27, 33)
3	q2	Chennai, form	What is the capital city of T	Chennai	(0, 7)
4	q3	Chennai, form	What is Chennai known fo	Its rich history, vibr	(189, 247)
5	q4	Chennai, form	In which region of India is	South India	(150, 161)
6	q5	Chennai, form	What are some of the key	Cultural, economic,	(104, 146)
7	q6	Chennai, form	What is the state to which	Tamil Nadu	(78, 88)
8	q7	Chennai, form	What are some notable lar	The Marina Beach,	(304, 685)
9	q8	Chennai, form	What is the Marina Beach?	The Marina Beach,	(304, 406)
10	q9	Chennai, form	Which temple is famous in	Kapaleeshwarar Te	(461, 485)

Fig 1: Example picture of curated dataset

PREPROCESSING STEPS FOR ROBERTA

Tokenization and Converting to Numerical Format:

- Tokenization is the process of breaking down the input text into individual tokens (words, subwords, or characters).
- RoBERTa's tokenizer uses a Byte-Pair Encoding (BPE) algorithm to split words into subword tokens.

Padding:

- Padding is added to ensure all sequences have the same length.
- The padding="max_length" argument ensures that all tokenized sequences are padded to the maximum length specified by max_length.

Overflow Handling and Truncation:

- return_overflowing_tokens=True ensures that overflow tokens are returned. This is useful for handling long sequences by dividing them into multiple segments with a certain overlap defined by stride.

PREPROCESSING STEPS FOR ROBERTA

Offset Mapping:

- The return_offsets_mapping=True argument returns the mapping between tokenized tokens and their corresponding character positions in the original text.
- This mapping is essential for identifying the start and end positions of the answer spans in the tokenized sequences.

Answer Position Calculation:

- The code iterates over the tokenized sequences and calculates the start and end positions of the answer spans within the context.
- It uses the character offsets provided by offset_mapping to map the answer span from the original text to the tokenized sequences.
- If the answer span is not fully contained within the context, it's labeled as (0, 0).

FINE-TUNING ROBERTA

Epoch	Training Loss
1	1.2009
2	0.6775
3	0.4106
4	0.275
5	0.2142
6	0.1426
7	0.0973
8	0.0575
9	0.0611
10	0.0405

Epoch	Training Loss
11	0.0434
12	0.0289
13	0.0312
14	0.0279
15	0.02
16	0.0264
17	0.02
18	0.0152
19	0.0183
20	0.0178

FINE-TUNING ROBERTA

- RoBERTa was fine-tuned for the curated dataset
  - Model Fine-tuned: deepset/roberta-base-squad2
  - 124M parameters
  - Training Loss of 0.0178 after 20 epochs
  - By default validation loss is not returned for RoBERTa
  - commit URL: aditi2212/roberta-finetuned-ChennaiQA-final · Hugging Face
  - Fine-Tuning was performed on Microsoft Azure

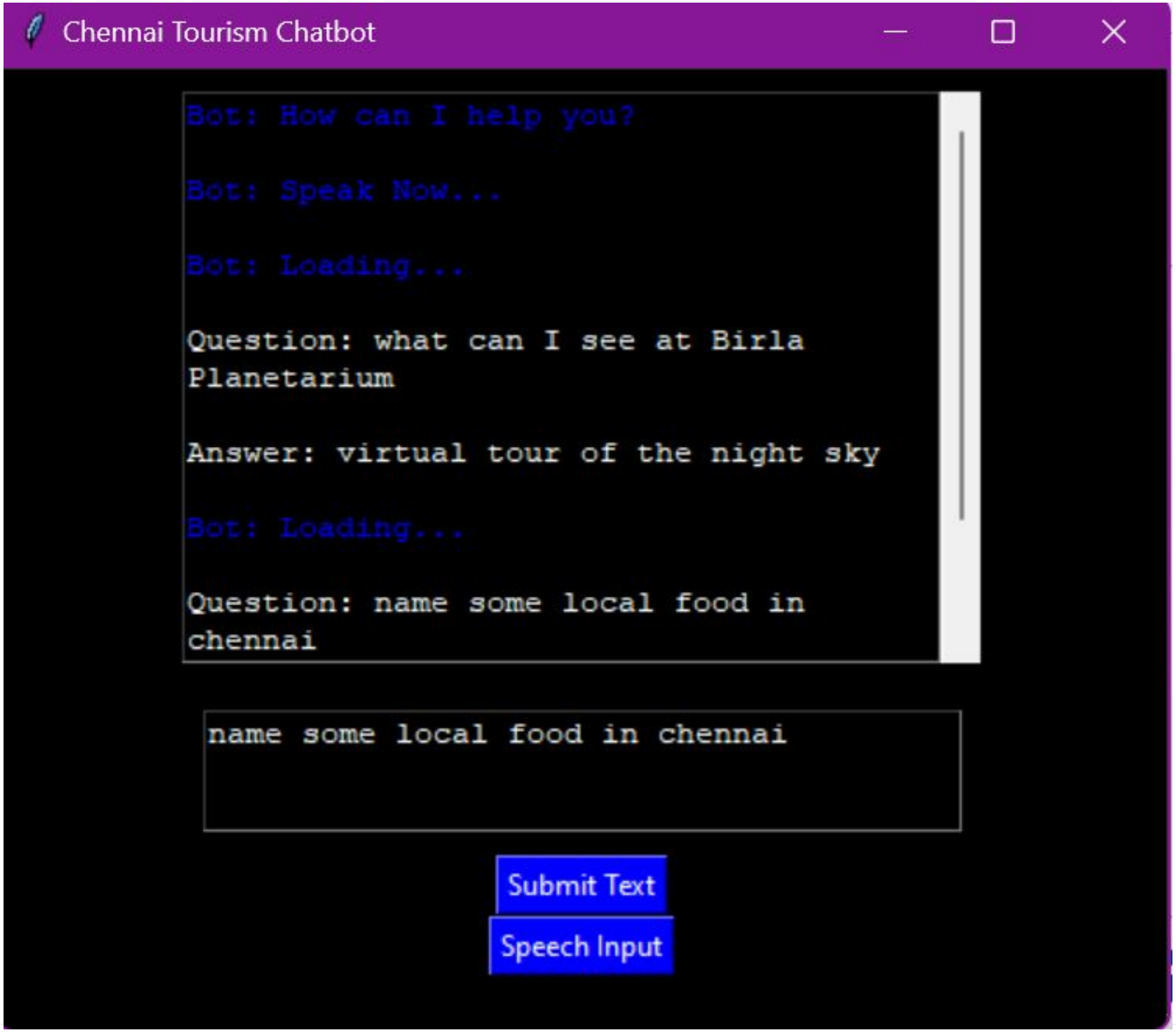
AUTOMATIC SPEECH RECOGNITION

- This code relies on Google Speech-to-Text, which requires an internet connection.
- The accuracy may vary depending on background noise, microphone quality, and speech clarity.
- It's a basic implementation and doesn't handle functionalities like continuous listening or advanced noise cancellation.
- There are libraries like DeepSpeech for a more offline-capable solution, but the setup complexity is very high.
- Python modules such as **pyttsx3, pyaudio** were used for speech output whereas **speech_recognition** module was used for speech input.

TESTING

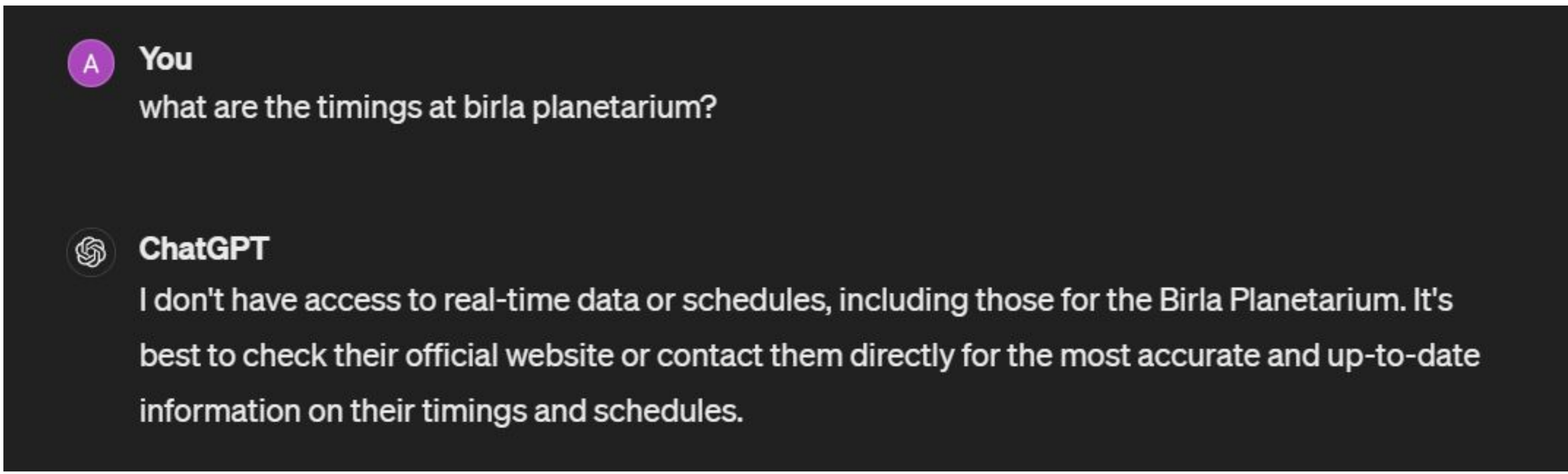
- The model was **tested for all 320 questions** that were formulated
- **Out of 320 questions**, the model was able to answer **282 questions** accurately
- This implies that the model shows 88.125% accuracy while testing.
- Similarity measures such as Jaccard Similarity and Levenshtein distance were used as evaluation metrics
- **Jaccard Similarity formula:**
- **$J(A, B) = |A \cap B| / |A \cup B|$**
- High Jaccard scores indicate a strong overlap in content between the expected and retrieved answers, while low Levenshtein distances suggest minimal edits are necessary for an exact match

USER INTERFACE



WHY USE OUR CHATBOT?

- Domain specificity
- Multimodal input output capabilities



DEMO

Disclaimer: The chatbot can accurately provide information and details about a tourist attraction (like where it is situated, what are the timings). I does not have the ability to measure distances between two places as map facilities are not integrated.

RESULTS AND CONCLUSIONS

- A Q&A system was developed to address tourism inquiries for Chennai, India, integrating ASR and TTS for a user-friendly experience.
- Roberta was fine-tuned on a dataset of 320 Q&A pairs about Chennai, with a training set of 270 and a testing set of 50 examples.
- After 20 epochs, the model achieved a significant loss reduction to 0.0178, indicating effective learning and error minimization.
- The model showed 88.125% accuracy while testing with similarity metrics such as Jaccard similarity and Levenshtein distance

FURTHER WORK

- The training dataset can be expanded to include more places for tourism other than Chennai, expanding our region of focus
- Multilingual Support can also be a future area of research
- Building a feedback system to improve the model based on user queries is another area of future research
- The current ASR system has some difficulties recognizing Indian names (like Kapaleeshwarar temple), although it fares well with an Indian accent. Improving the ASR for Indian names is another area of research.

REFERENCES

[1] L. Wu, Y. Chen, K. Shen, X. Guo, H. Gao, S. Li, J. Pei, and B. Long, (2021) "Graph Neural Networks for Natural Language Processing: A Survey," https://doi.org/10.48550/arXiv.2106.06090.

[2] Yasunaga, M., Ren, H., Bosselut, A., Liang, P., & Leskovec, J. (2021). “QA-GNN: Reasoning with Language Models and Knowledge Graphs for Question Answering”. https://doi.org/10.48550/arXiv.2104.06378.

[3] Y. Wu, Q. Zhou, H. Yin and D. Liu, "An Interpretable Question Answering Method Based on Heterogeneous Graph Neural Networks," 2022 International Conference on Computer Engineering and Artificial Intelligence (ICCEAI), Shijiazhuang, China, 2022, pp. 576-580, doi: 10.1109/ICCEAI55464.2022.00123.

[4] Li, Ronghan & Wang, Lifang & Jiang, Zejun & Hu, Zhongtian & Zhao, Meng & Lu, Xinyu. (2022). “Mutually improved dense retriever and GNN-based reader for arbitrary-hop open-domain question answering. Neural Computing and Applications”. 34. 10.1007/s00521-022-07072-0.

REFERENCES

[5] Omar Khattab, Christopher Potts, Matei Zaharia; "Relevance-guided Supervision for OpenQA with ColBERT." Transactions of the Association for Computational Linguistics 2021; 9 pp. 929–944. doi:https://doi.org/10.1162/tacl_a_00405

[6] S. Haji, K. Suekane, H. Sano and T. Takagi, "Exploratory Inference Chain: Exploratorily Chaining Multi-hop Inferences with Large Language Models for Question-Answering," 2023 IEEE 17th International Conference on Semantic Computing (ICSC), Laguna Hills, CA, USA, 2023, pp. 175-182, doi: 10.1109/ICSC56153.2023.00036.

[7] Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. "IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages". In Findings of the Association for Computational Linguistics: EMNLP 2020, pp. 4948–4961, Online. Association for Computational Linguistics.

THANK YOU