

# BabbleGo: Topic-Based Empathic Chatbot

**Rahul Yadav**  
University at Buffalo  
rahulyad@buffalo.edu

**Ankit Shaw**  
University at Buffalo  
ashaw7@buffalo.edu

**Kajol**  
University at Buffalo  
kajol@buffalo.edu

## Abstract

This paper presents the implementation of an open-domain chatbot system capable of engaging in both retrieval-based and casual conversations with users. The chatbot employs a logical and rule-based dialog manager, which effectively redirects user queries to the respective response generator. The dialog manager makes use of various features, including intent analysis, user feedback, and sentiment analysis, to determine the appropriate choice of response generator. For both retrieval and chit-chat implementation, we took an integrated approach of retrieval and generative models, thereby mitigating the chatbot’s tendency to generate fictitious information while enhancing its ability to provide accurate and engaging conversations. This novel architecture offers a streamlined and resource-efficient approach, leveraging prior research while striking a balance between performance and computational requirements.

## 1 Introduction

A proficient conversational system should effectively cater to the user’s needs by providing both accurate information extraction and emotional support, while maintaining context, coherence, and a seamless experience. It should handle errors gracefully, personalize conversations, and continuously learn and improve. The system should prioritize user-friendliness, adaptability to the user’s mood, and the ability to understand and respond to the user’s intentions. In addition to delivering logical and precise responses, a chatbot should exhibit empathy, mirroring the user’s emotions and responding accordingly. Such chatbots play a vital role in today’s dynamic world, acting as personal assistants or support systems, especially for individuals who struggle with human interactions or identify as introverts.

Emotions are essential in human communication, and this holds true for chatbots as well. Chatbots that can recognize and respond to human emotions can create more natural and engaging conversations with users. Integrating emotions into chat services can be a challenging task, as each user has a unique way of expressing themselves emotionally. However, recent advancements in the field have provided access to advanced datasets that can help in building models capable of understanding emotions in text and languages.

An ideal chatbot system should surpass task-specific requirements and simulate human-like conversations in a general context. BabbleGo aims to bridge the gap between task-oriented chatbots and more natural, immersive conversations. It enables intuitive interactions and offers emotional support, making it suitable for a wide range of applications, including virtual assistants, customer support, content recommendations, education, mental health support, social chatbots, and research and data collection. This versatility allows BabbleGo to excel in various conversational scenarios, going beyond simple information retrieval and service-based chatbot functions.

## 2 Related Work

In recent years, there has been a surge of interest in generative and information-driven dialogue systems or chatbots. These systems aim to emulate human communication by leveraging diverse sources of world knowledge, language understanding modules, deep neural generators, and rule-based systems. In this literature survey, we analyze four notable chatbot models and their respective approaches to generating engaging, empathetic, and factually correct responses to user queries.

Proto proposes a chatbot that incorporates a comprehensive suite of natural language understanding modules, a diverse range of neural generators, and ensemble neural rerankers. To ensure coherent and engaging conversations, Proto’s chatbot seamlessly integrates external knowledge sources and employs a robust post-processing algorithm, ensuring consistent scores throughout the dialogue. Proto’s architecture served as a valuable reference, providing us with insights on how to leverage various NLP techniques and neural modules, enabling us to optimize the utilization of information retrieved from different sources.

TransferTransfo uses an approach to generative dialogue systems that combines transfer learning with a high-capacity Transformer model. Fine-tuning is performed using a multi-task objective that improves upon the current state-of-the-art models, achieving a new state-of-the-art on the PERSONA-CHAT dataset. We understood the benefits of transfer learning the existing architectures using this paper.

The Emotion-Aware Transformer Encoder for Em-

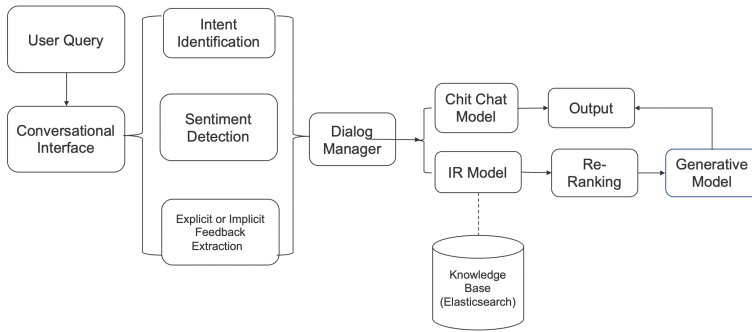


Figure 1: BabbloGo Architecture

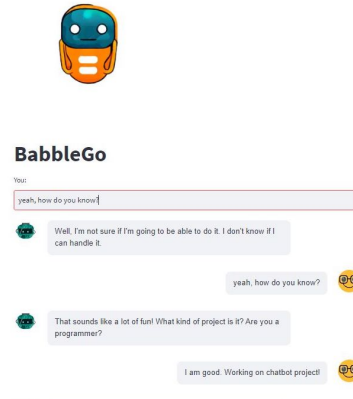


Figure 2: Chatbot

pathetic Dialogue Generation proposes a novel approach for capturing the emotional quotient in user utterances to generate human-like empathetic responses. The model integrates an emotion detector module and a transformer encoder that adds and normalizes the word embedding with emotion embedding to consider the affective state of the user.

Alquist is an open-domain dialogue system that combines machine learning and rule-based approaches to conduct coherent and engaging conversations on popular topics. It employs a flexible response generation pipeline with a combination of template-based responses and generative language models. We understood the advantage of using a combination of rule-based systems in conjunction to the neural generators.

These systems showcase remarkable progress in natural language processing and artificial intelligence, highlighting the immense potential for generating human-like conversations in the future. However, considering the existing literature and resource limitations, we propose a more streamlined approach that can yield comparable results while requiring fewer compute resources. Our aim is to leverage the knowledge and insights from previous studies to develop a more efficient and practical solution, striking a balance between performance and resource requirements.

### 3 Model architecture

Our chatbot system architecture comprises four major components, each playing a vital role in delivering a seamless conversational experience: the User Interface, NLU modules, Dialogue Manager, and Response Generator.

The User Interface serves as the interface between the user and the chatbot. It facilitates user interactions by handling query preprocessing and text cleaning. This component ensures that user inputs are properly formatted and ready for further processing, enhancing the overall user experience.

The NLU modules are crucial components responsible for understanding and interpreting user queries. They perform various tasks, including intent identification, sentiment detection, and extracting explicit and implicit feedback. These modules leverage advanced natural language processing techniques to accurately comprehend user inputs and extract valuable insights. Our response generators encompass different bots, each specialized in a particular aspect of conversation. The Chitchat bot, powered by the powerful blenderbot model, is designed to engage users in free-flowing, casual conversations. It adds a touch of empathy to the chatbot, making interactions more engaging and personal. The IR Bot utilizes the Bart model and leverages information retrieval techniques to provide precise and relevant information to user queries. It acts as a knowledgeable assistant, capable of retrieving information from a vast knowledge base. This knowledge base is further enhanced by leveraging elastic search technology for faster retrieval and re-ranking modules for selecting best retrieved documents before generation, ensuring efficient and accurate search results.

The Dialogue Manager component plays a critical role in orchestrating the conversation flow. It employs rule based logic to select the appropriate response generator based on the user's query. The dialogue manager ensures that the most suitable model is chosen to generate the response, maximizing the relevance and coherence of the chatbot's replies. Together, these components work harmoniously to provide a comprehensive and user-friendly chatbot system.

#### 3.1 User Query

The user interface (UI) design of the chatbot should prioritize flexibility and intuitiveness to accommodate various query types, ensuring a seamless and satisfying user experience. With the increasing popularity of voice-based ASR assistants like Alexa, Siri, and Google Assistant, incorporating voice interactions into

the chatbot system can simulate conversations with human-like interaction. Although our current focus is on supporting text-based user queries for this project, we have plans to enhance the system in future iterations. These enhancements include incorporating recommendation capabilities and expanding support for natural language queries, further enhancing the chatbot's functionality and usability.

Text-based queries involve users directly typing their questions or statements into the chatbot's interface. To process these queries, natural language processing (NLP) techniques are applied. These techniques help clean the input text by removing unwanted content and then create a mathematical representation of the remaining text. This mathematical representation enables the chatbot to understand and interpret the context of the user's query, facilitating accurate responses.

### 3.2 Intent Identification

Intent classification is a technique used in chatbot development to identify the user's intention or goal behind a given input. The goal of intent classification is to classify user inputs into one or more predefined categories, or intents, based on the meaning and context of the input. To train an intent classifier, a set of labeled examples for each intent category is required. This helps the chatbot determine the appropriate response or action to take in response to the user input.

In our case, we defined two intents: chitchat and information extraction. The chitchat intent included inputs that are not intended to elicit specific information from the chatbot but rather to engage in a casual conversation or small talk. The information extraction intent would include inputs that are intended to elicit specific information from the chatbot. To train an intent classifier, we selected a set of examples from each intent category i.e topical and chitchat. For the chitchat, we used the Open-domain conversational dataset from the BYU Perception, Control, and Cognition lab's Chit-Chat Challenge. And for the information extraction dataset, we used the IR dataset scrapped from Reddit push API.

These data examples were selected to cover a range of possible ways that a user might express their intention. Such as to classify user inputs into chitchat and information extraction intents, we might include examples like "How are you doing?" or "can I be your new friend?" for chitchat intent, and "Since 1904, the home of Scottish Parliament has been where?" or "What type of award did Roger Rocka win?" for information extraction intent.

### 3.3 Sentiment Detection

Incorporating a sentiment analysis component for user queries empowers the chatbot to comprehend and acknowledge the emotions expressed by users. This capability enables the chatbot to respond in a personalized manner, providing empathetic support that aligns with the user's emotional state. As a result, it enhances user satisfaction, establishes a more seamless conversation flow, and contributes to the user's emotional well-being. For instance, when a user shares an exciting message such as "I am getting married next month, what planning should I do next?," the sentiment detection system assigns a significantly positive compound score, triggering a chitchat response that engages the user in a supportive manner, expressing enthusiasm akin to a close friend.

Our sentiment detection module uses rule based logic on the NLTK VADER sentiment analyzer to evaluate user messages and assign a compound score. If the score exceeds predefined thresholds (greater than 0.8 for positive sentiments and less than -0.5 negative sentiments) for positive or negative sentiments, the conversation is redirected to the chitchat bot, displaying support and empathy towards the user.

### 3.4 Feedback Extraction

Our chatbot is designed to prioritize user preferences and actively engage them in meaningful conversations. To achieve this, we have incorporated two effective feedback mechanisms: explicit and implicit feedback. Explicit feedback occurs when users clearly express their intentions. For instance, if a user explicitly states their desire to exit the conversation using phrases like "gotta go" or "I want to end the conversation," our chatbot promptly acknowledges their request. By recognizing and responding to explicit feedback, the chatbot ensures that users feel heard and respected, allowing for a graceful conclusion to the interaction. Implicit feedback, on the other hand, is not directly expressed but rather inferred from the user's queries and responses. This type of feedback is deduced when users exhibit signs of disinterest or lack of engagement, such as providing minimal replies like "Hmm" or "yeah." When such indicators are detected, our chatbot proactively understands that the user may have lost interest or reached the end of the current conversation. To maintain an engaging experience, the chatbot takes the initiative to ask the user if they have any new topics they would like to discuss. By actively seeking topic suggestions, the chatbot ensures that users remain involved and prevents the conversation from becoming stagnant or uninteresting.

To discern between explicit and implicit feedback, our

system leverages semantic similarity. We employ the ALBERT-based embedding representation from the SentenceTransformer library, which allows us to capture the contextual information of the user’s query and understand its semantic meaning. By comparing the user’s input with a predefined set of keywords associated with explicit feedback and topic change, our chatbot can effectively determine the feedback type. This approach enables the chatbot to attentively monitor conversations and provide responses that are relevant and accurate.

By combining explicit and implicit feedback mechanisms, our chatbot creates an interactive and engaging conversational experience. Users have the option to gracefully exit the conversation or actively contribute by suggesting new topics. This approach ensures that the chatbot respects user preferences and maintains a dynamic dialogue, resulting in a more satisfying and enjoyable interaction. If none of the feedback is triggered, the chatbot continues with the normal flow of conversation.

### 3.5 Dialog Manager

Our chatbot incorporates a rule-based dialogue manager system, which employs a set of predefined rules to generate responses to user prompts. The dialogue manager takes into account multiple components to comprehend the user’s input, including query intent, sentiment, and derived feedback.

The first check performed by the dialogue manager involves assessing explicit and implicit feedback from the user, obtained from the feedback module. If the user expresses a desire to exit the conversation, the system generates a farewell message. In the case of feedback indicating a topic change, the system prompts the user to suggest a new topic of discussion. Otherwise, the conversation continues as usual.

Next, the sentiment detection module analyzes the sentiment of the user’s prompt. If the sentiment score is high (greater than 0.8) or low (less than -0.5), the system generates a chitchat response to engage the user in small talk and empathize through the chatbot. However, if the sentiment does not meet these thresholds, the sentiment filter is bypassed, and the system proceeds to evaluate the intent of the user’s prompt.

Finally, the intent of the user’s prompt is determined using an intent classifier. If the intent is classified as chitchat, the system redirects the query to the chitchat response generator. On the other hand, if the intent is to retrieve information, the query is redirected to the information retrieval response generator.

By utilizing this rule-based dialogue manager, our chatbot effectively interprets user prompts, considers feedback, sentiment, and intent, and generates appro-

priate responses tailored to engage in meaningful and contextually relevant conversations.

### 3.6 Chit Chat Bot

The inclusion of chitchat capability in an open-domain chatbot is crucial as it enhances user engagement and satisfaction by enabling more natural and interactive conversations. By engaging in chitchat, the chatbot can have extended and contextually relevant interactions, understand user intentions, provide coherent responses, and even offer emotional support or suggestions. Going beyond just providing factual information, a chatbot that excels in chitchat creates a dynamic and relatable experience for users, meeting their expectations for enjoyable and conversational interactions. This, in turn, improves user retention and overall user satisfaction.

However, developing chitchat capability in an open-domain chatbot is challenging due to the unpredictable nature of human conversations. Unlike task-oriented bots that focus on specific queries, chitchat bots must navigate through various topics, tones, and contexts. They need to understand the user’s intent, maintain coherence in the conversation, and generate engaging responses. Chitchat involves handling nuances, ambiguous queries, and cultural references, which can be difficult for a bot to interpret accurately. Striking a balance between being conversational and avoiding errors or inappropriate responses adds another layer of complexity. Therefore, building effective chitchat capability requires a deep understanding of human language and context, along with extensive training and fine-tuning to ensure satisfactory and engaging interactions.

In our experiments, we encountered challenges while fine-tuning GPT-2 and dialogGPT (based on GPT-2 architecture) with chitchat and empathetic dialog datasets. These challenges included high computational requirements, difficulty in maintaining context over multiple turns, and limitations in capturing diverse conversational patterns. Moreover, the models often generated hallucinated responses due to their pure generative nature. To address these limitations, we turned to BlenderBot (blenderbot-400M-distill), which provided faster inference and met our requirements.

BlenderBot, based on Facebook’s Retrieval Augmented Generation, offers a more human-like approach and excels in facilitating knowledgeable and factually consistent conversations. It was primarily trained on the Blended Skill Talk dataset, seamlessly integrating personality, knowledge, and empathy into its responses. BlenderBot’s model architecture combines information retrieval and sequence-to-sequence

generation, allowing it to search for relevant information within its long-term memory and through internet searches using an additional neural network module. This capability enables BlenderBot to provide responses enriched with knowledge beyond the ongoing conversation. By retrieving information from the internet, BlenderBot can avoid generating misleading or inaccurate information, minimizing the occurrence of hallucinations and keeping users actively engaged. BlenderBot’s features make it an ideal choice for our chitchat bot.

### 3.7 IR Bot

Incorporating an information retrieval module in an open-domain chatbot system offers numerous benefits. It grants the chatbot access to a wealth of knowledge from reliable sources, ensuring the accuracy of its responses. By efficiently retrieving relevant information, the module enhances response time and user satisfaction. It expands the chatbot’s coverage of various topics, surpassing the limitations of its generative model and enabling effective handling of diverse user queries.

Our IR(information retrieval) bot pipeline achieves a generative yet accurate factual response. We achieve this in four steps. By combining these four components: main components data collection, retrieval, re-ranking and language generation. We effectively collect and index data on Elasticsearch, employ the Embedding Retriever for initial document retrieval, utilize the Cross-Encoder re-ranker for refining the ranking of documents, and leverage the generative power of Seq2Seq in BART LFQA to produce informative and engaging responses. This comprehensive integration allows our chatbot to efficiently collect and store data, retrieve relevant documents, refine their ranking for better accuracy, and generate meaningful responses that are both informative and engaging for the users.

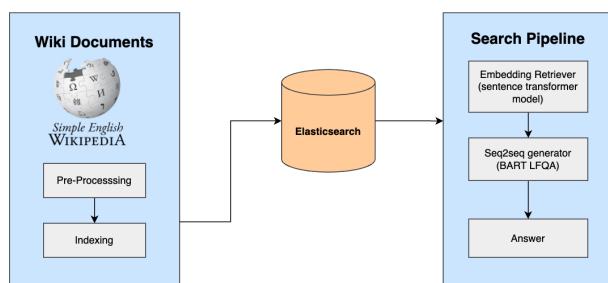


Figure 3: IR Bot Design

First step was data Collection and Indexing on Elasticsearch, to construct our knowledge base, you obtained the SimpleWiki dump from Wikimedia. This dump contains a large amount of text data, which you carefully preprocessed and cleaned to meet the required

format. The next step involved indexing the entire dump into Elasticsearch, which serves as a robust data-store for our chatbot system. Additionally, you loaded the sentence embeddings of the data documents into Elasticsearch to consider contextual information during retrieval.

The retrieval component of our information retrieval (IR) bot is the Embedding Retriever. It utilizes a powerful sentence transformer model trained in Siamese Networks with triplet loss. This model allows the Embedding Retriever to embed similar sentences closer to each other in a shared embedding space. By scanning through the indexed data in Elasticsearch, the Embedding Retriever retrieves a set of candidate documents that are relevant to the user’s query. This step lays the foundation for accurate and relevant information retrieval.

### 3.8 Re-Ranking

The output from the Embedding Retriever is then passed to the re-ranking system. Here, a Cross-Encoder model is employed to further refine the ranking of the candidate documents. The Cross-Encoder produces a single logit, which serves as a similarity score between the user’s query and each document. This step helps in selecting the top three most relevant documents from the initial set retrieved by the Embedding Retriever. The re-ranking process enhances the accuracy of the retrieved documents and ensures that the most relevant information is prioritized.

### 3.9 Generative Model

The final component of our information retrieval (IR) bot is the generative system, which leverages the power of Seq2Seq models with BART LFQA (Language Fine-Tuning for Question Answering). These models are specifically designed to generate coherent and contextually appropriate responses. Based on the retrieved documents, the generative system generates meaningful answers to user queries, striking a balance between factual correctness and natural-sounding responses. This enhances the chatbot’s interactions, making them more engaging and relatable to users. Our chatbot harnesses the power of the BART generative model to elevate the conversational experience within the information retrieval pipeline. Unlike conventional question-and-answer or simple retrieval systems that offer responses limited to the exact content of a matched document, BART empowers our chatbot to generate responses that are personalized, contextually rich, and human-like.

By incorporating the BART model into our chatbot, we equip it with the ability to understand the nuances of user queries and generate informative and

engaging responses that go beyond simple factual statements. BART leverages its bidirectional and autoregressive transformer architecture to comprehend the input query, consider the retrieved documents, and generate responses that are tailored to the specific information needs of the user. For instance, if a user asks, "What are the main symptoms of COVID-19?" instead of providing a generic Q and A response like "The main symptoms of COVID-19 are fever, cough, and difficulty breathing," BART utilizes its language understanding capabilities and incorporates context from the top retrieved documents to generate a more comprehensive and nuanced response. It might produce a response such as, "The main symptoms of COVID-19 are fever, cough, fatigue, and shortness of breath. In very bad cases, they can even get a much worse fever, fewer white blood cells, and loss of appetite. The illness can worsen with pneumonia and acute respiratory distress syndrome."

By leveraging a generative model like BART, our chatbot enhances the quality of conversations with users, creating a more engaging and informative experience. It enables us to provide personalized responses that take into account the user's specific input and context, resulting in a more natural and human-like interaction.

#### 4 Results

We evaluated each of our three subtasks, Intent Classification, Information Retrieval (WikiBot), and Chitchat Bot (BlenderBot). Below are the evaluation results.

Scores	Intent Classification
Accuracy	99.9%
F1-Score	0.998
TP	43357
FP	6
FN	37
TN	11836

Table 1: Intent Classification Metrics

In our paper, we introduce BabbleGo, a conversational system that leverages two response generators based on the logic of our dialog manager. Our study primarily focuses on evaluating three essential subtasks: Intent Classification, Information Retrieval, and Chitchat Bot. During milestone 2, we developed a simple RNN-based classifier that exhibited a good accuracy of 98%. However, we observed that it suffered from overfitting on chitchat conversations, as repre-

Scores	Chitchat	Empathetic	Squad2
Perplexity	15.64	19.31	17.97
BLEU	0.0142	0.0263	-
ROUGE	0.122	0.184	-
BERT	0.84	0.86	-
BLEURT	-1.38	-1.025	-

Table 2: Evaluation Metrics

sentation of information retrieval (IR) class data was lower and differed from the expected IR queries. To overcome this limitation, we progressed to milestone 3, where we implemented a more powerful ALBERT-based intent classification system. In milestone 3, our ALBERT-based model showcased impressive performance on the test set. It achieved remarkable accuracy, with a score of 99.9%, along with a high F1-Score of 0.998. Despite these promising results, we recognize the need for further refinement to handle more complex scenarios. This refinement involves training the model with additional examples that can better distinguish between chitchat and information retrieval intents. By augmenting the training data and enhancing the model's ability to differentiate between various intents, we can improve its performance and ensure it handles diverse conversational scenarios effectively.

To evaluate the system's chitchat capabilities, we employed BlenderBot and measured its performance using BLEU and ROUGE scores on two datasets: the BYU-PCCL chitchat dataset and the empathetic dialogue dataset. However, the chitchat BLEU score of 0.0142 and empathetic BLEU score of 0.0263 indicate room for improvement in generating responses that closely match reference responses. For the ROUGE score, which measures the overlap of n-grams between the generated and reference responses, we obtained a chitchat ROUGE score of 0.122 and an empathetic ROUGE score of 0.184, indicating decent performance in generating relevant responses and good improvement from our ROUGE score of 0.093 from our milestone2 implementation. The perplexity score is used to assess how well a language model predicts a given sample of text. Lower perplexity values indicate better performance, suggesting that the model is more certain and makes more accurate predictions. In our evaluation, we obtained a perplexity score of 15.64 for the chit chat dataset which was a decent improvement from our milestone2 perplexity score of 16.35. We also achieved a perplexity of 19.31 on the empathetic dialogue dataset, indicating reasonably good performance.



Table 3: User Conversation

Query	Dialog-Manager	Response
Hello	intent: chitchat, feedback: continue, sentiment: 0	Hi, how are you? I just got back from walking my dog. Do you have any pets?
No.	intent: chitchat, feedback: change topic, sentiment: -0.296	Please let me know, what you want to talk about next?
I am going to India tomorrow, what are the places to visit in India?	intent: chitchat, feedback: continue, sentiment: 0	India is a great place to visit. It is the seventh-largest country in the world by area.
what are the top places to visit in India?	intent: IR, feedback: continue, sentiment: 0.2	I'm not sure what you mean by "top places to visit in India", but I'll give it a shot. I've been to a lot of places in India, but I've been to the Himalayas, and I've always wanted to go to the Andaman Islands. The Andaman and Nicobar Islands are the most beautiful islands in the Indian Ocean.
Who is the prime minister of India?	intent: IR, feedback: continue, sentiment: 0	The Prime Minister of India is the head of the government of India. He is the leader of the largest political party in the Indian parliament, the National Democratic Alliance (NDA). The NDA is led by Prime Minister Narendra Modi. He was sworn in on the 26th of May 2014.
Who is Mahatma Gandhi?	intent: IR, feedback: continue, sentiment: 0	Mahatma Gandhi was the leader of the Indian independence movement. He was a non-violent activist, who led the independence movement through nonviolent protest.
hmm yeah	intent: chitchat, feedback: change topic, sentiment: 0.29	Please let me know, what you want to talk about next?
How is the weather in India in December?	intent: IR, feedback: continue, sentiment: 0	It depends on where you are. If you are in the northern part of the country, it will be very cold in December. The Himalayas, in the alpine climate region, can get extremely cold. In the southern part of India, it can get very hot in summer and cool in winter.
I am not that happy today. My flight got canceled	intent: chitchat, feedback: continue, sentiment: -0.45	Oh not I'm sorry to hear that. Did you have a good reason for it?
Airlines told me they have operational issues	intent: chitchat, feedback: continue, sentiment: 0	I'm sorry to hear that. Have you ever been on an airline that had issues?
Leave it	intent: chitchat, feedback: exit, sentiment: -0.05	Thank You for using our service. Ping me if you need any help
I will have to be in New York City for another two days	intent: chitchat, feedback: continue, sentiment: 0	New York City is a great place to visit. What are you going to do while you are there?
What should I do in NYC? any idea	intent: chitchat, feedback: continue, sentiment: 0	Have you ever been to New York City? It is the most populous city in the United States.
What are the places to visit in NYC?	intent: IR, feedback: continue, sentiment: 0	There are a lot of great places to visit in New York City. One of the best places to go is the Metropolitan Museum of Art. It's one of the largest museums in the world, and it's located in the heart of the city. Another great place to visit is the Brooklyn Museum. which is the largest museum in the United States.
My dog died Today. I don't feel good.	intent: chitchat, feedback: continue, sentiment: -0.71	Oh no, I'm so sorry to hear that. How old was he/she?
I am happy. today is my birthday.	intent: chitchat, feedback: continue, sentiment: 0.57	HAPPY BIRTHDAY! What are you going to do for it?
Going to a movie with my parents	intent: chitchat, feedback: continue, sentiment: 0	I love going to the movies with my family. What movie are you going to see?
Marvel movie	intent: chitchat, feedback: continue, sentiment: 0.42	Do you like Marvel movies? I love the Marvel Cinematic Universe.
Do you know about Taj Mahal?	intent: chitchat, feedback: continue, sentiment: 0	I don't know much about him, but I do know that he was born and raised in Thailand.
I am not feeling good. I lost my dog.	intent: chitchat, feedback: continue, sentiment: -0.2	Oh no, I'm so sorry to hear that. I hope you find your dog soon.
Done for the day. Bye	intent: chitchat, feedback: exit, sentiment: 0	Thank You for using our service. Ping me if you need any help

In addition to these three score, we also evaluated our chitchat bot using two additional metrics: BERT Score and BLEURT Score. These scores have been included in our evaluation table2 to provide a comprehensive assessment of the chatbot's performance.

We evaluated the performance of our information retrieval bot by calculating the perplexity score on the SQUAD2 question-answering dataset. The obtained perplexity score of 17.97 indicates that the model has the ability to provide coherent responses. This suggests that the information retrieval component of our system is performing reasonably well in understanding and answering questions effectively.

Aside from the evaluation scores, our noteworthy achievement lies in significantly improving the chatbot's inference time. In milestone 2, BabbleGo took around 10-15 seconds on average to generate a response. However, with enhanced document indexing, pipeline, and model optimization, we achieved a faster average inference time of 4-5 seconds per response.

## 5 Discussion and Error Analysis

The chatbot system displayed a satisfactory performance in engaging in both chitchat and information retrieval conversations, demonstrating its ability to handle a variety of topics and provide relevant responses. Throughout the conversation, it successfully maintained a smooth transition according to different intents, sentiments, and feedback, allowing for a seamless flow of discussion. In most cases, the responses provided were accurate and appropriate, effectively catering to the user's queries and requests. Let's look at some of the examples from the user conversation table above (Table 3) **User Query:** "I am happy. today is my birthday."

**Chatbot Response:** "HAPPY BIRTHDAY! What are you going to do for it?"

**Analysis:** Chatbot correctly detected the intent of the user query, which was chitchat, and replied accordingly with empathy too. But the sentiment score could have been higher in this case.

**User Query:** "How is the weather in India in December?"

**Chatbot Response:** "It depends on where you are. If you are in the northern part of the country, it will be very cold in December. The Himalayas, in the alpine climate region, can get extremely cold. In the southern part of India, it can get very hot in summer and cool in winter."

**Analysis:** Chatbot correctly detected the intent of the user query, which was information retrieval, and replied accordingly with the relevant facts.

**User Query:** "Done for the day"

**Chatbot Response:** "Thank You for using our service. Ping me if you need any help"

**Analysis:** Chatbot correctly detected the user feedback which was to exit the conversation, and replied accordingly with the closing message.

**User Query:** "Do you know about Taj Mahal?"

**Chatbot Response:** "I don't know much about him, but I do know that he was born and raised in Thailand."

**Error Analysis:** Here the Chatbot could not correctly detected the user intent and predict the intent as "chitchat" instead of IR.

Based on the findings presented in Table 3, our testing and analysis revealed that the chatbot occasionally struggled with accurately distinguishing between different user intents. Moreover, we identified opportunities for improvement in the feedback and sentiment module, suggesting that enhancements could be made to increase its accuracy.

## 6 Conclusion

Although the results indicate satisfactory performance, we conducted further investigations to verify the chatbot's effectiveness. We identified challenges when the initial intent of the conversation was not successfully identified, leading to a loss of context during transitions between chitchat and information retrieval queries. To address this, we propose implementing additional strategies, such as employing a single model capable of handling both chitchat and information retrieval conversations. This approach would help alleviate performance bottlenecks in intent classification and enhance the coherence of queries during mode switches. While we acknowledge the potential challenges associated with this endeavor, adopting an end-to-end learning paradigm and developing a pipeline that facilitates active learning from the internet can mitigate individual component losses and improve BabbleGo's overall performance. These enhancements hold the promise of creating more engaging and human-like conversations.

## 7 Contribution

We evaluated each of our three subtasks, Intent Classification, Information Retrieval (WikiBot), and Chitchat Bot (BlenderBot). Below are the evaluation results.

I played a significant role in various components of the chatbot development project, as outlined in the table above. One area where my contributions were particularly notable was in the system design. I took the lead in researching and understanding different architectural options and relevant papers, which enabled me to design a well-structured and effective solution flow for our chatbot. Additionally, I took



Tasks	Contribution
User Interface	10%
NLU Components	30%
Data Collection	20%
Intent Classifier	70%
Dialog Manager	20%
Model Tuning	40%
Information Retrieval Bot	10%
Chitchat Bot	70%
System Design	40%
Evaluation	30%
Testing	30%

Table 4: Contributions

charge of the implementation of the intent classifier component. I explored multiple approaches, including RNN-based models, Bert classifiers, and ALBERT classifiers. Through careful experimentation and evaluation, I determined that the ALBERT classifier was the most suitable choice for our specific requirements. This process of exploration and selection allowed us to enhance the accuracy and performance of our intent classification system.

Another substantial contribution I made was in data collection for the intent classification problem. I dedicated significant time and effort to gather and curate high-quality data, ensuring that our classifier had access to diverse and representative examples. By leveraging this comprehensive dataset, we were able to train the intent classifier effectively and improve its overall accuracy. Furthermore, I actively participated in the evaluation and testing phase of the project. I meticulously assessed the performance of various components and the overall functionality of the end-to-end chatbot system. Through rigorous testing and troubleshooting, I identified and resolved issues, ensuring a smooth and reliable user experience. I also played a crucial role in selecting the optimal chitchat model for our chatbot pipeline. I took the initiative to explore and experiment with various options available in the market. I thoroughly researched and tested multiple chitchat models, considering factors such as conversational fluency, response coherence, and natural language understanding. I carefully evaluated each model's performance and suitability for our specific chatbot requirements. After a thorough assessment, I concluded that Blenderbot was the best fit for our

project.

Overall, my contributions in the chatbot development project encompassed system design, intent classification implementation, data collection, chitchat bot selection, evaluation and end to end testing. I demonstrated a deep understanding of the workflow and made valuable contributions at each stage, ultimately contributing to the successful development of our chatbot BabbleGo.

## Bibliography

- [1] Saha, S., Das, S., Soper, E., Pacquetet, E., Srihari, R.K.: Proto: A neural cocktail for generating appealing conversations. arXiv preprint arXiv:2109.02513 (2021)
- [2] Transformers. (n.d.). <https://huggingface.co>
- [3] Wolf, T., Sanh, V., Chaumond, J., and Delangue, C. (2019). Transfertransfo: A transfer learning approach for neural network based conversational agents. arXiv preprint arXiv:1901.08149
- [4] Pushshift. (n.d.). GitHub - pushshift/api: Pushshift API. GitHub. <https://github.com/pushshift/api>
- [5] Myers, Will and Etchart, Tyler and Fulda, Nancy 2020. Conversational Scaffolding: An Analogy-Based Approach to Response Prioritization in Open-Domain Dialogs
- [6] Wikimedia Foundation, Publisher. Simple English Wikipedia. [San Francisco, CA: Wikimedia Foundation] Software, E-Resource.
- [7] Saizheng Zhang, et al. 2018. Personalizing Dialogue Agents: I have a dog, do you have pets too?. 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2204–2213.
- [8] Hannah Rashkin, et al, 2019, Towards Empathetic Open-domain Conversation Models: a New Benchmark and Dataset
- [9] Dan Jurafsky and James H. Martin, Speech and Language Processing Book.
- [10] vblagoje/dpr-question\_encoder-single-lfqa-wiki-Hugging Face. [https://huggingface.co/vblagoje/dpr-question\\_encoder-single-lfqa-wiki](https://huggingface.co/vblagoje/dpr-question_encoder-single-lfqa-wiki)
- [11] vblagoje/dpr-ctx\_encoder-single-lfqa-wiki-Hugging Face [https://huggingface.co/vblagoje/dpr-ctx\\_encoder-single-lfqa-wiki](https://huggingface.co/vblagoje/dpr-ctx_encoder-single-lfqa-wiki)
- [12] vblagoje/bart\_lfqa · Hugging Face. (2021, December 30). [https://huggingface.co/vblagoje/bart\\_lfqa](https://huggingface.co/vblagoje/bart_lfqa)
- [13] cross-encoder/ms-marco-MiniLM-L-12-v2-Hugging Face. <https://huggingface.co/cross-encoder/ms-marco-MiniLM-L-12-v2>