GERALD: A Conversational AI

George Mason University

Dr. Kevin Lybarger

Syed Abdul Hadi | Aidin Ziaee | Akash Hala Swamy | Vaishnavi Raju Echarlu Thimmaraju

---

Abstract: Open source and open domain question answering (OpenQA) is an active area of research in the Natural processing language (NLP); it uses a significant number of unstructured documents to attempt to provide a natural language answer to a user query. OpenQA typically calls for data-driven methods, which may be generally separated into two categories: retrieval-based and generation-based systems. Retrieval-based systems rely on existing data to retrieve responses. The quality of generated responses strictly depends on the program's ability to find a relevant response in the data. Although these models can provide contextually relevant responses if trained on large enough data, they suffer from a lack of relevance in the structure of the response, which is tied to the structure of the user query. On the other hand, generative models are trained on natural language and can generate structurally accurate sequences, however, these models suffer from a lack of contextual relevance and most responses can be short and meaningless. In this paper, we propose an open-source and open-domain conversation AI model that uses a sequential ensemble of a retrieval-based and generation-based system to intelligently respond to user queries while maintaining contextual and structural relevance. The retrieval-based system has been trained on a large dataset based on conversational AI question-answer pairs and natural language conversations from reddit, Counsel Chat and IRC Chat Rooms etc. Top k queries in the dataset that match closely with a user query are extracted and processed. These queries are then evaluated and passed on to the generative model which intelligently generates new responses which are contextually relevant and structurally aligned with the user query. The retrieval-based system uses a BERT transformer to contextually evaluate a user query against the dataset, and the generative model uses a T5 transformer model, which conditions the retrieved text and the user query to generate new sequences. This ensemble system is known to outperform retrieval-based and generation-based models working independently.

Index Terms: Conversational AI, open domain chatbot, NLP, BERT, T5 transformer.

---

1 INTRODUCTION

Chatbots are computer programs that are designed to simulate human-like conversations in a natural way. Advancements in the field of Artificial intelligence(AI) have led to better understanding of human behavior [1]. Most of the existing chatbots use the concept of specializing the dialog in the design phase. These chatbots are all closed domain, their functionality is limited for operating in one such domain such as MIMIC - chatbot for retrieving information about the movies[2]. For all open domain systems, the predefined

rule- based approach fails as it cannot handle diversity of the data. There are several reasons for it, but one major reason is that it could not handle large volumes of data. This limitation is overcome by Conversational AI that uses large volumes of data to imitate human-like conversations with the user. Based on the previous approaches for open domain conversational AI, using the data-oriented approach resulted in categorizing it into a retrieval dialog system and a generation dialog system [4].

The retrieval-based dialog systems work by searching for the most similar question present in the dataset and it outputs the corresponding answer to the matched question. The main limitation that we observe here is that the user will not be able to get any new utterances; the second major limitation observed is regarding the ranking of the answers stored in the database; ranking is usually performed by calculating the tf-IDF score and sometimes using word overlaps which do not address the semantics of NLP.

The generative-based dialog system addresses the second limitation of the retrieval-based dialog system using real value vectors called embeddings; there exists another layer that decodes the embedding to generate a reply. There are a few limitations of using this kind of dialog systems, there are many instances where the generated replies are short, very generic or universal, and sometimes meaningless.

Limitations observed by both types of dialog systems are overcome by assembling both the retrieval-based and generative-based approaches into one. This results in generated replies that are meaningful and relevant.

In this paper, we propose an open-source and open-domain conversation AI model named GERALD (named after Gerald Salton) that uses a sequential ensemble of a retrieval-based and generation-based system to intelligently respond to user queries while maintaining contextual and structural relevance. The retrieval-based system has been trained on a large dataset based on conversational AI question-answer pairs and natural language conversations from multiple sources. The top k queries in the dataset that match closely with a user query are extracted and processed. These queries are then evaluated and passed on to the generative model, which intelligently generates new responses which are contextually relevant and structurally aligned with the user query. The main objective of the program is to generate queries intelligently.

On performing the above-mentioned process, the experimental results show that combining both the retrieval and generative model into a single component overcomes the limitations of each of them and outperforms retrieval-based and generation-based models working independently.

2 DATA SOURCES

Data for developing 'GERALD: conversational AI' was gathered from multiple sources such as reddit, publicly available AI Q-A Datasets, CounselChat and IRC Chat Rooms constituting 20GB of data accounting for 1.5 million question answer pairs. Formats of data fetched were json, csv and txt files. Few of the data sources had direct question-answer pair data, whereas few were in the conversational format, which we had to manually segregate into question-answer pairs format.

The below figure shows the sample data used for our model.


```
{
"question": "What was Beyonc\u00e9's role in Destiny's Child?",
 "id": "56d43ce42ccc5a1400d830b4",
"answers": [{"text": "lead singer", "answer_start": 290}],
 "is_impossible": false
},
{"question": "What was the name of Beyonc\u00e9's first solo album?",
"id": "56d43ce42ccc5a1400d830b5",
"answers": [{"text": "Dangerously in Love", "answer_start": 505}],
"is_impossible": false}],
 "context": "Beyonc\u00e9 Giselle Knowles-Carter (/bi\u02d0\u02c8j\u0252nse\u026a/ bee-YON-say) (born September 4, 1981) is an American
singer, songwriter, record producer and actress. Born and raised in Houston, Texas, she performed in various singing and dancing
competitions as a child, and rose to fame in the late 1990s as lead singer of R&B girl-group Destiny's Child. Managed by her father,
Mathew Knowles, the group became one of the world's best-selling girl groups of all time. Their hiatus saw the release of Beyonc
\u00e9's debut album, Dangerously in Love (2003), which established her as a solo artist worldwide, earned five Grammy Awards and
featured the Billboard Hot 100 number-one singles \"Crazy in Love\" and \"Baby Boy\".\"},
 {"qas": [{"question": "After her second solo album, what other entertainment venture did Beyonce explore?",
"id": "56be86cf3aeaaa14008c9076",
"answers": [{"text": "acting", "answer_start": 207}],
"is_impossible": false},
 {
"question": "Which artist did Beyonce marry?",
"id": "56be86cf3aeaaa14008c9078",
"answers": [{"text": "Jay Z", "answer_start": 369}],
"is_impossible": false},
```


Fig 1: Data before processing

The data that we integrated from various sources was not readily usable as it had duplicates, null, fields, and indirect answers such as a unique ID to the answer maintained by some internet forum. Some of the questions had 'No Answer Given,' and similar data within the answer section, identifying and removing all these inconsistencies was a non-trivial task.

Preprocessing of data involved getting rid of duplicates, noise, and handling null/blank values. All this was achieved by using python libraries such as numpy and pandas. After processing the data, we were left with 0.6 million question-answer pairs.

3 LITERATURE REVIEW

Prior to now, developers have mostly focused on domain-specific and closed-domain conversational systems. Usually, pre-made themes like sports, psychotherapists, etc. These ontologies establish a limited number of data and values that must be used to address user queries. This diagram shows [Fig2] the traditional architecture conversational system

design[5], which primarily consists of three steps which are Question Analysis, Document Retrieval, and Answer Extraction.

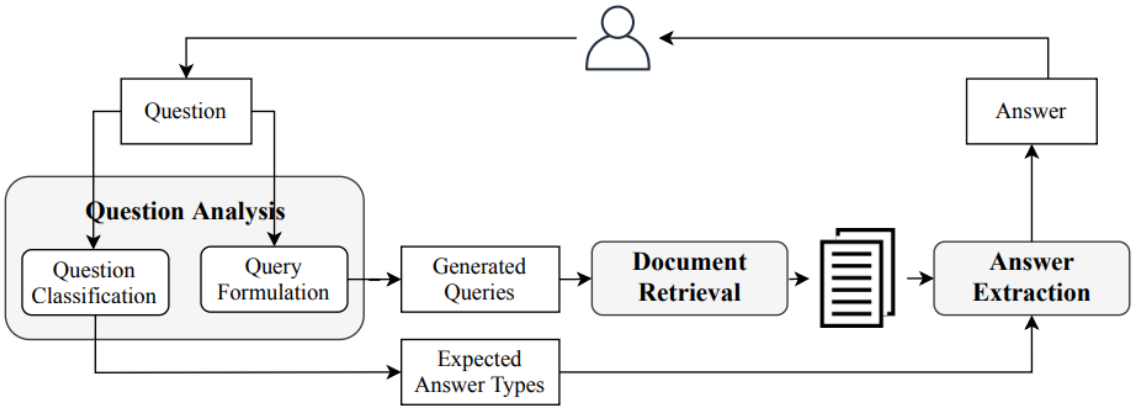

Fig 2: An illustration of the traditional architecture of OpenQA Conversational system [5]

When users ask the Question/query to the system, the Question Analysis module tries to understand the question and whether it is classified as a question based on a set of question types (e.g., where, when, who, what, how ...) manually defined by linguistic experts. Then it gets the expected answer from the corpus and fed to the Answer Extraction module, which returns the answer. If the user query is ambiguous, then, in Query Formulation to extract keywords for retrieval, linguistic procedures including POS tagging, stemming, parsing, and stop word removal are often used. Processed queries are then fed to Document retrieval to get a small number of relevant documents from a collection using an IR engine that most likely contains the response to a certain query which is fed to the Answer extraction module, this module is responsible for providing the best appropriate answer to a user's query. The complexity of the question determines this stage's effectiveness [5].

The system is predicated on domain knowledge, thus acquiring this data and optimizing the system will take much time. The user query may not always be system comprehensible, and the terminology and words used in the question are not always the same as those used in the documents containing the answer or response. This issue, often known as "term mismatch,"[5] will cause the model to return meaningless sentences, which is a vital issue in IR. This retrieval approach's efficacy and efficiency will be lower. This model and similar ones only utilize one of the two approaches—either retrieval or generative and not both. To address these issues, we have developed a model called ‘GERALD: Conversational AI’.

4 PROPOSED SYSTEM

GERALD is a system that intelligently responds to user queries while maintaining contextual and structural relevance. It does this by sequentially combining retrieval-based and generation-based systems. Below is the architecture diagram for the Gerald models.

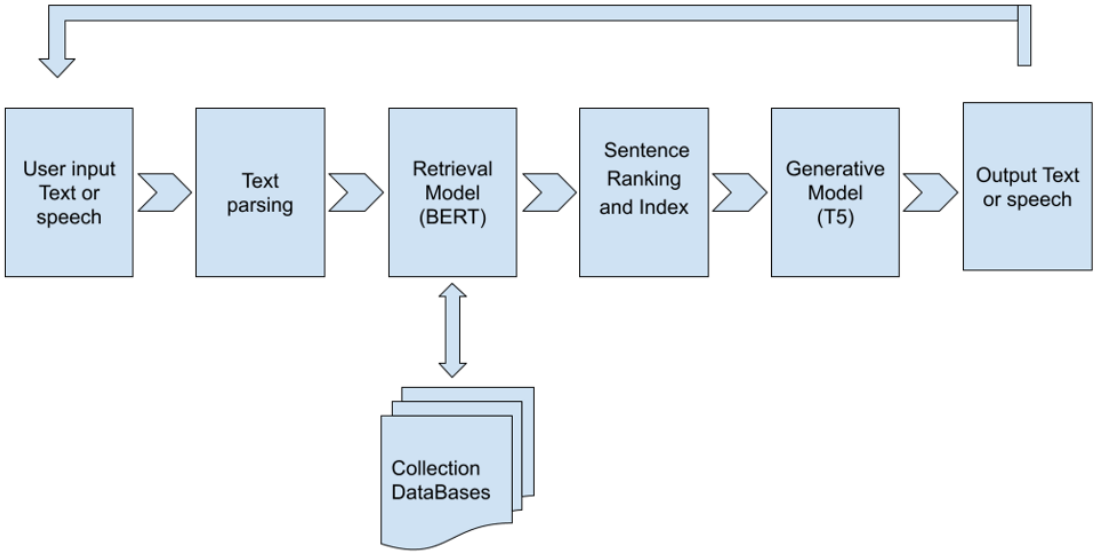

Fig 3 : An illustration of Gerald: A Conversational AI.

When a user enters a question or query in text or speech format, it is sent on to our text parsing module, which performs text parsing and preprocesses the query to best fit the user's query. It is then passed on to the Retrieval Approach Model utilizing the BERT (Bidirectional Encoder Representations from Transformers). A list of N relevant sentences to the user's query is obtained when we search a knowledge database for it. Then we send them to the sentence ranking and index module, which performs the task of removing any meaningless, brief responses, sentences that don't match, or responses that aren't as relevant as those returned by the retrieval module.

In terms of Generative Approaches Using T5 (the Text-To-Text Transfer Transformer), an encoder-decoder architecture is used to encode a query as vectors and to decode the vectors to generate a response. Instead of relying on retrieval-based methods, a generative system will produce fresh phrases. When a retrieved module’s context is considered for the generated response, it may be improved to be more meaningful in addition to being fluid and logical in relation to the question. After passing the generated sentences to the output

module, this module will finally provide the user with the most correct response to their question/query, either in text or speech format.

In terms of the objectives of the proposed model, this open-domain conversational AI aims to use the ensemble model and combine the retrieval-based and generation-based models to provide the most intelligent and human-like response to the users' questions and queries. In terms of the minimum hardware requirements to build this model, we needed at least 12GB of RAM memory for the machine to handle the large dataset. From a software requirement standpoint, we needed Python and various Python libraries to build our model. Below Python packages were used to build our model: TensorFlow, keras, keytotext, pipeline, time, sklearn.metrics.pairwise, cosine_similarity, sentence_transformers, SentenceTransformer, numpy, pandas, nltk.tokenize, word_tokenize, pyaudio, pyttsx3, speech_recognition, msvcrt, transformers, T5Tokenizer, T5ForConditionalGeneration, sklearn.decomposition, PCA, and pickle.

5 METHODOLOGY

The entire training data (after cleaning) is transformed into a vector space using a BERT transformer. Each of the 'questions' is given a vector representation of length 768. These weights are stored in a text file of size 9.8GB. For scalability, we reduce the size of these weights using Principal Component Analysis (PCA), preserving 95% of the variance within the weights. The length of these vector representations is reduced to a length of 278.

When a user presents a query to the model, a vector representation of length 768 is formed using the BERT transformer for the query. The dimensionality of this vector is then reduced using the same PCA model used to reduce the dimensions of the training data. When saved as a .pkl file, the PCA model does not need to be retrained every time the model is run. The reduced length vector is compared against all queries within the data using cosine similarity.

$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}},$$

A vector having the length of the number of data observations is created and the indices of this vector are sorted based on the match similarity. Indices corresponding to the top 3 matching queries are taken and the answers corresponding to these queries (which share the same indices) are extracted.

$$\boldsymbol{index_i = argsort_i\,(match\_vector)\,,\ for\ i\ in\ [0,1,2]}$$

	Question	Answer
2600	I've been married for 3 years and I have two k...	p Hi there Thank you reaching out for help I...
2601	I have been with my husband for almost 7 years...	p It s hard to let go of the dreams you had re...
2602	What can I do to stop grieving my mother's dea...	p I am sorry that you lost your mother That i...
2603	I'm having issues with my relative. The police...	p I think it would be wise for you to call a h...

Fig 4: Q-A Pairs in the Dataset

From the perspective of optimizing the algorithm, the next part of the program is categorized into a Layer 1 and Layer 2, and the top 3 retrieved responses are evaluated before being pushed to the generative model

Layer 1 is engaged when a similarity of 90% or greater exists between the user query and the first retrieved response. The query within the dataset and the answer corresponding to this query are fed to the pre-trained T5 Conditional Generator. The T5 model contextualizes the data query and response into a single output which is then presented to the user. The length of the output sequence is capped at 1.5 times the length of the answer within the dataset to ensure relevance and prevent the T5 model from including meaningless sequences to the response. When left uncapped, it has been observed that the T5 model elongates decoder output to include meaningless sequences of text.

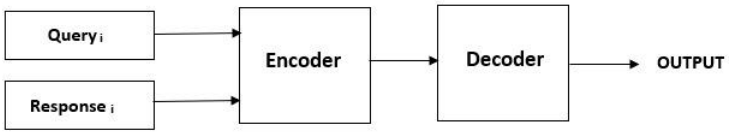

Fig 5: Layer 1 of Generative Model

Layer 2 is engaged when the retrieval system cannot find a response having a similarity higher than 90% within the dataset. In this case the top 3 data queries and their corresponding responses are extracted and fed individually to the T5 generator which creates generates a sequence of max length 25 for each of the 3 query-response pairs. In the next step, each of these 3 generated responses along with the user query are fed to the T5 generator again and this time a sequence of max length 60 is generated. This is presented as an output to the user.

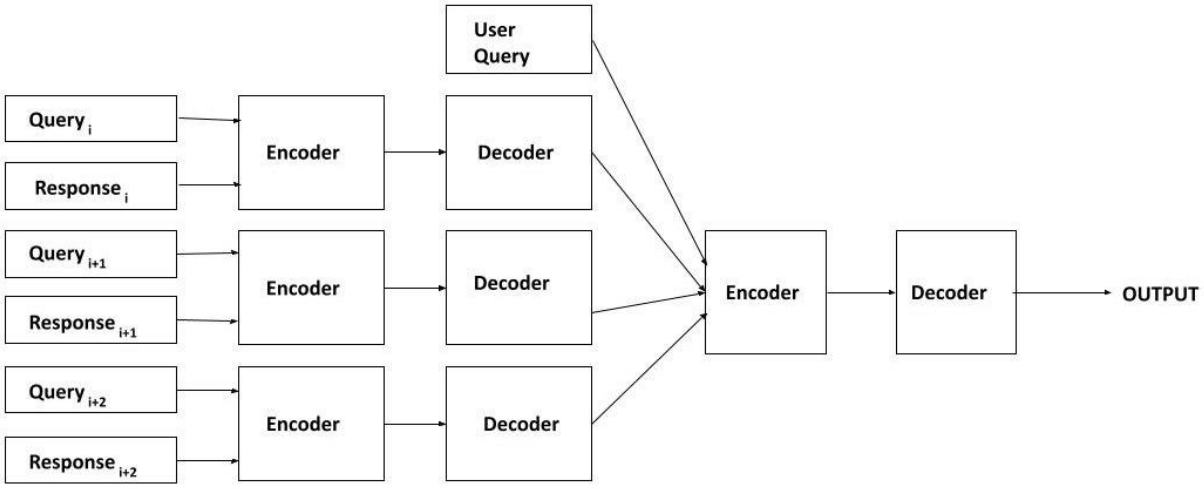

Fig 6: - Layer 2 of Generative Model

5 APPLICATIONS

Open domain conversational AI is purpose-built to achieve specific goals, these include psychotherapy, education and recommendation etc. For instance, Amazon Alexa and Apple's Siri are built to map a user query to an achievable task. Gerald has been trained on a large corpus of therapy data extracted from CounselChat where certified therapists respond to user questions. However, these therapist responses are not conversational in nature. Moreover, conversational therapy data is difficult to obtain because of challenges associated with this information. Client-therapist privilege prevents real-world to be made available and for this reason, Gerald cannot be reasonably scaled to provide psychotherapy oriented conversational systems as of now, although it is reasonably equipped to respond to non-conversational therapy-related questions, for instance, 'I feel, I experience, I don't understand' etc. On the other, the availability of massive Q-A datasets (especially from Wikipedia) makes Gerald particularly suitable for responding to factual questions presented by a user. The training data available to Gerald also consists of conversations on Reddit, which is useful in generating responses to non-factual queries a user may input.

Gerald is a development towards conversational AI, a program that can reasonably maintain a conversation with a user by responding relevantly, and the framework can be applied to solve specific goals by expanding the training data. A simple application of Gerald without any expansion is education and recreation.

Gerald can also be scaled for closed-domain applications. The program locates a query that is relevant to a user query and then processes the corresponding response. These responses can be mapped to specific tasks and specific outputs, which can enable Gerald to be used

in customer service applications i.e., connecting user to a relevant representative, make appointments, collect feedback about a specific product etc.

6 RESULTS AND DISCUSSIONS

The outputs of out of sample queries are compared against human generated responses using BLEU scores and vector-transformed representation cosine scores. While the BLEU score measures word overlap, the cosine scores gauge the semantic match between two sentences. A gold standard response and a model generated response are matched for semantic match, which is represented as the cosine similarity score.

Method	**Score**
BLEU SCORE	9.82 e-232
Cosine Similarity SCORE	0.815

The BLEU scores appear to be extremely low, but this can be attributed to the fact that BLEU scores gauge word overlap and this may necessarily result in a low match. Although we did not have the resources to develop a 'human score' metric, the responses were evaluated by human subjects, and many of the factual queries produced responses that were deemed relevant and appropriate.

Some of the sample responses generated by the program are:

```
Who is JK Rowling
jk rowling is a british novelist best known for her work on the Harr
y potter series. she is also the author of the harry potter series o
f books.

I often have headaches
headaches are caused by stress, tension or prolonged tension. the mo
st common cause of headaches is stress.

How do I relieve stress
meditation, yoga and regular exercise can relieve stress. if you tal
k to someone about your feelings exercise helps relieve stress.
```

```
Who was George Mason
mason was an american patriot statesman and a delegate from virginia
to the u s constitutional convention. mason was a member of the u s.
senate who was a member of the senate who was a member of the senate
who was a member of the senate.

How many awards did Beyonce win
beyonce won 20 grammys.

Are therapists any good
therapist can help you cope with feelings and symptoms and change be
havior patterns. a therapist can help you change your thinking and b
ehavior.

What is your age
what is your age, i am still young by your standards. i am still a s
tudent.
```

The program has its limitations when tested over a range of domains. Since the program begins by evaluating a user query against its own data through the retrieval system, the success of the generated output is strictly tied to the program's ability to find a related query within the data. If the program cannot find a related query (and any query that is remotely related), the output response is shown to be irrelevant. In the example response below, the program cannot find a specific comparison for 'coke' and 'pepsi', therefore it instead looks from a broader perspective and produces results based on a retrieved output that simply compares two liquids, 'gasoline' and 'diesel'.

```
is coke better than pepsi
diesel gives better mileage than gas yes diesel give better mileage
than gas.
```

Furthermore, the length cap of the T5 model fails in some cases and the program does not determine a good stopping point based on the retrieved outputs, as shown below. This is often the case when there is no consensus within the retrieved data.

```
Is coke better than pepsi?
diet coke is a sugar free soft drink. diet coke is a sugar free soft
drink. morphine is stronger. morphine is stronger. morphine is stron
ger. morphine is stronger. morphine is stronger. morphine is stronge
r.
```

7 Conclusion and Future Work

In this project, we created Gerald, an open-source/open-domain conversation AI model which applies a novel ensemble of retrieval-based and generation-based methods to receive the most accurate response based on the user's query. We concluded that the ensemble model performs much better than individually applying the retrieval-based or generation-based model.

We learned that using the retrieval method plays an essential role in our model. This method is doing a semantic match which is essential to match the meaning of the sentences to find relevant responses. The model often fails when a relevant response is not generated by the retrieval model. Such problems can be fixed by introducing a failure detection method that informs the user that the system is unable to find a relevant response. However, a more potent method for resolving this problem is to expand the training data for the model to be able to find responses to more queries. Moreover, the program must also be equipped with a better-engineered method for stopping criteria. Even when the model is able to find relevant responses within the data, it generates responses that exceed the length past the relevant part of the response.

Finally, the program can be equipped with a state of the art information retrieval system such as Lucene which can essentially help the model retrieve data faster and in a more efficient manner.

In terms of the limitations of our work, we understood that modeling the natural language could become difficult due to a large amount of data on Human-Human conversations and the variety of the natural language. In addition, we needed a powerful computer to handle the large dataset we used. If the dataset is expanded further, a more powerful machine and more optimized methodologies will need to be implemented for the program to be scalable.

Our work proposes new paths for future work. We recommend increasing the dataset size to cover more natural language and human-human conversations. To handle the larger dataset, we recommend using a more powerful computer to optimize the model and receive a faster response. Lastly, using our applied methods, future work can contain new approaches to combing the retrieval and generative dialogue systems to build the most optimum human-computer conversation system.

References

[1] Shingte, K. *et al.* (2021) *Chatbot development for Educational Institute, SSRN*. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3861241 (Accessed: December 3, 2022).

[2] Jennifer Chu-Carroll. MIMIC: An adaptive mixed initiative spoken dialogue system for information queries. In Proc. Conf. Applied Natural Language Processing, pages 97–104, 2000.

[3]] Tsung-Hsien Wen, Milica Gasic, Nikola Mrksic, Lina M Rojas-Barahona, Pei-Hao Su, Stefan Ultes, David Vandyke, and Steve Young. A network-based end-to-end trainable task-oriented dialogue system. arXiv preprint arXiv:1604.04562, 2016

[4] Charles Lee Isbell, Michael Kearns, Dave Kormann, Satinder Singh, and Peter Stone. Cobot in LambdaMOO: A social statistics agent. In AAAI, pages 36–41, 2000

[5] Zhu, F. et al. (2021) Retrieving and reading: A comprehensive survey on open-domain question answering, arXiv.org. Available at: https://arxiv.org/abs/2101.00774

[6] *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. (2018, October 11). arXiv. from https://arxiv.org/abs/1810.04805

[7] Roberts, A., & Raffel, C. (2020, February 24). *Exploring Transfer Learning with T5: the Text-To-Text Transfer Transformer*. Google AI Blog. from https://ai.googleblog.com/2020/02/exploring-transfer-learning-with-t5.html