A Grounded Well-being Conversational Agent with Multiple Interaction Modes: Preliminary Results

Xinxin Yan, and **Ndapa Nakashole**
Computer Science and Engineering
University of California, San Diego
La Jolla, CA 92093
x3yan@ucsd.edu, nnakashole@eng.ucsd.edu

Abstract

Technologies for enhancing well-being, healthcare vigilance and monitoring are on the rise. However, despite patient interest, such technologies suffer from low adoption. One hypothesis for this limited adoption is loss of human interaction that is central to doctor-patient encounters. In this paper we seek to address this limitation via a conversational agent that adopts one aspect of in-person doctor-patient interactions: A human avatar to facilitate medical grounded question answering. This is akin to the in-person scenario where the doctor may point to the human body or the patient may point to their own body to express their conditions. Additionally, our agent has multiple interaction modes, that may give more options for the patient to use the agent, not just for medical question answering, but also to engage in conversations about general topics and current events. Both the avatar, and the multiple interaction modes could help improve adherence.

We present a high level overview of the design of our agent, Marie Bot Wellbeing. We also report implementation details of our early prototype , and present preliminary results.

1 Introduction

NLP is in a position to bring-forth scalable, cost-effective solutions for promoting well-being. Such solutions can serve many segments of the population such as people living in medically underserved communities with limited access to clinicians, and people with limited mobility. These solutions can also serve those interested in self-monitoring (Torous et al., 2014) their own health. There is evidence that these technologies can be effective (Mayo-Wilson, 2007; Fitzpatrick et al., 2017). However, despite interest, such technologies suffer from low adoption(Donkin et al., 2013). One hypothesis for this limited adoption is the loss of human interaction which is central to doctor-patient encounters(Fitzpatrick et al., 2017). In this paper we seek to address this limitation via a conversational agent that emulates one aspect of in-person doctor-patient interactions: a human avatar to facilitate grounded question answering. This is akin to the in-person scenario where the doctor may point to the human body or the patient may point to their own body to express their conditions. Additionally, our agent has multiple interaction modes, that may give more options for the patient to use the agent, not just for medical question answering, but also to engage in conversations about general topics and current events. Both the avatar, and the multiple interaction modes could help improve adherence.

The human body is complex and information about how it functions fill entire books. Yet it is important for individuals to know about conditions that can affect the human body, in order to practice continued monitoring and prevention to keep severe medical situations at bay. To this end, our well-being agent includes a medical question answering interaction mode (**MedicalQABot**). For mental health, social isolation and loneliness can have adverse health consequences such as anxiety, depression, and suicide. Our well-being agent includes a social interaction mode (**SocialBot**), wherein the agent can be an approximation of human a companion. The MedicalQABot is less conversational but accomplishes the task of answering questions. The SocialBot seeks to be conversational while providing some information. And, there is a third interaction mode, the **Chatbot**, which in our work is used as a last-resort mode, it is conversational but does not provide much information of substance.

To test the ideas of our proposed agent, we are developing a grounded well-being conversational agent, called “Marie Bot Wellbeing". This paper presents a sketch of the high level design of our Marie system, and some preliminary results.

An important consideration when developing technology for healthcare is that there is *low toler-*

Proceedings of the 1st Workshop on NLP for Positive Impact, pages 143–151
Bangkok, Thailand (online), August 5, 2021.

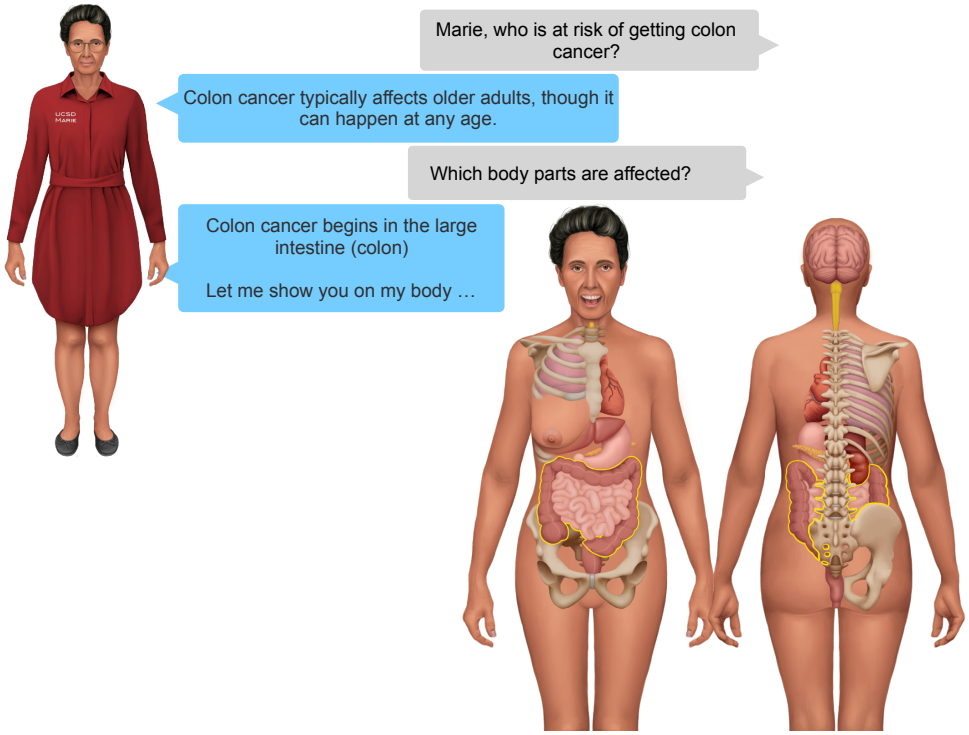

Figure 1: An illustration of the MedicalQA interaction mode. Here the agent's answer is grounded on our human avatar. The affected body part, the large intestine, is highlighted on the avatar.

ance for errors. Erroneous information can have severe negative consequences. We design the medicalQABot, and the SocialBot with this consideration in mind. Our design philosophy consists of the following tenets:

1. **Reputable answers:** Only provide answers to questions for which we have answers from reputable sources, instead of considering information from every corner of the Web.

2. **Calibrated confidence scores:** Even though the answers come from reputable sources, there are various decisions that are involved that the model must make including which specific answer to retrieve for a given question. For these predictions by our models, we must know what we do not know, and provide only information about which the model is fairly certain.

3. **Visualize**: Whenever an answer can be visualized to some degree, we should provide a visualization to accompany the text answer to help clarify, and reduce misunderstanding.

4. **Graceful failure**: when one of the interaction modes fails, another interaction mode can take over.

Organization In what follows, we discuss how the above tenets are manifested in our agent.

The rest of the paper is organized as follows: We begin with a high-level overview of the design of the different parts of the agent (Sections 2 to 4); We next discuss the current prototype implementation and preliminary results (Section 5); We next present related work (Section 6); and close with a discussion (Section 7) and concluding remarks (Section 8).

2 Interaction Modes and Dialog Management

In navigating between the different interaction modes, we design our system as follows. Based on the user utterance, we automatically predict using a binary classifier to switch between different interaction modes (MedicalQABot vs SocialBot). Suppose that the classifier predicts that the utterance is a question asking for medical information on a topic, and suppose our medicalQA determines that we have no information on that topic, our goal is to then let the SocialBot take over if it has information on that topic and can meaningfully hold a conversation about it. For the SocialBot, when missing the necessary information, our goal is to have it fall back to Chatbot mode.

Figure 2: An illustration of the SocialBot interaction mode

3 MedicalQABot Mode

3.1 Knowledge vault of QA pairs

Some aspects of the human body are well-understood, many diseases and medical conditions have been studied for many years. Thus a lot of medical questions have already been asked, and their answers are known. Thus one approach to medicalQA is a retrieval-based one which consists of two steps: First, we collect and create a knowledge vault of frequently asked questions and their curated answers from reputable sources.

Second, given a user question, we must match it to one of the questions in the QA knowledge vault. However, when people pose their questions, they are not aware of the exact words used in the questions of the knowledge vault. We must therefore match user questions to the correct question in the knowledge vault. A simple approach is keyword search. However, this misses a lot of compositional effects. One other way is to treat this as a problem of entailment. Where given a user question, we can find, in the knowledge vault, the questions that entail the user question.

3.2 Grounding to Human Anatomy Avatar

We develop a human avatar to help users better understand medical information. And also to help them to more precisely specify their questions. The avatar is meant to be used in two ways. The human avatar was illustrated by a medical illustrator we hired from Upwork.com.

Bot → Patient: When an answer contains body parts, relevant body parts are highlighted on the avatar. "this medical condition affects the following body parts ". An illustration of this direction is shown in Figure 1.

Patient → Bot: When the user describes their condition, they can point by clicking. "I am not feeling well here".

4 SocialBot Mode

For the SocialBot, we propose to create a knowledge vault of topics that will enable the bot to have engaging conversations with humans on topics of interest including current events. For example, the bot can say “Sure, we can talk about German beer" or. "I see you want to talk about Afghan hounds"". The topics will be mined from Wikipedia, news sources, and social media including Reddit. For the SocialBot, we wish to model the principles of a good conversation: having something interesting to say, and showing interest in what the conversation partner says (Ostendorf, 2018)

5 Prototype Implementation & Preliminary Experiments

Having discussed the high-level design goals, in the following sections we present specifics of our initial prototype. Our prototype’s language understanding capabilities are limited. They can be thought of as placeholders that allowed us to quickly develop a prototype. These simple capabilities will be replaced as we develop more advanced language processing methods for our system.

5.1 Data

We describe the data used in our current prototype.

Medline Data We collected Medline data [1], containing 1031 high level medical topics. We extracted the summaries and split the text into questions and answers. We generated several data files from this dataset: question-topic pair data, answer-topic pair data and question-answer pair data. The data size and split information is presented in Table 3. We will describe their usage in detail in the following sections

Medical Dialogue Data We use the MedDialog dataset(Zeng et al., 2020) which has 0.26 million dialogues between patients and doctors. The raw dialogues were obtained from healthcaremagic.com and icliniq.com.

We also use the MedQuAD (Medical Question Answering Dataset) dataset (Ben Abacha and Demner-Fushman, 2019) which contains 47457 medical question-answer pairs created from 12 NIH[2] websites.

News Category Dataset We also use the News category dataset from Kaggle[3]. It contains 41 topics. We use the data in 39 topics, without "Healthy Living" and "Wellness", which might be related to the medical domain. We extract the short description from the dataset.

Reddit Data We collected questions and comments from 30 subreddits. We treat each subreddit as one topic. The number of questions for each topic is shown in Table 7. This Reddit data is to be used for our SocialBot.

5.2 System Overview

As shown in Figure 3, our system makes a number of decisions upon receiving a user utterance. First, the system predicts if the utterance should be handled by the MedicalQABot or by the SocialBot.

If the MedicalQABot is predicted to handle the utterance, then an additional decision is made. This decision predicts which Medical topic the utterance is about. If we are not certain, the system puts the user in the loop, by asking them to confirm the topic. If the user says the top predicted topic is not the correct one, we present them with the next topic in the order, and ask them again, up to 4 times.

[1]https://medlineplus.gov/xml.html
[2]https://www.nih.gov/
[3]https://www.kaggle.com/rmisra/news-category-dataset

Train	286370
Valid	35796
Test	35797

Table 1: Interaction Mode Prediction Data

Valid accuracy	0.9970
Test accuracy	0.9972

Table 2: Interaction Mode Prediction Evaluation Results

If the SocialBot is predicted to handle the utterance, the goal is to have the system decide between various general topics and current events for which the system has collected information. If the topic is outside of the scope of what the SocialBot knows, the system resorts to a ChatBot, that may just give generic responses, and engage in chitchat dialogue.

5.3 Mode Prediction Classifier

We train this classifier to determine whether the user's input is related to the medical domain. We use the output from BERT encoder as the input to a linear classification layer trained with a cross-entropy loss function.

We choose the positive examples from MedQuAD Dataset, and negative examples from News Category Dataset. The training data information is shown in Table 1. And the evaluation results are shown in Table 2. This performance is potentially better than in real-life settings, because the medical (medline) vs non-medical (Kaggle news) data is cleanly separated. In reality, a user utterance might be "I am not happy, I have a headache" they may not want to get medical advise, but simply to just chat a bit to distract them from the headache.

5.4 MedicalQA Implementation

Medical Topic Classifier If the user utterance is routed to the MedicalQABot, the MedicalQABot first predicts the medical category of the user's input. We use Medline Data, which contains 1031 topics, to train this classifier. The dataset information is shown in Table 3. The evaluation results of our medical topic classifier is shown in Table 4.

Topic Posterior Calibration As shown in Figure 3, we ask a topic confirmation question after the topic classifier, which is used to let the user confirm the correctness of the output from Topic

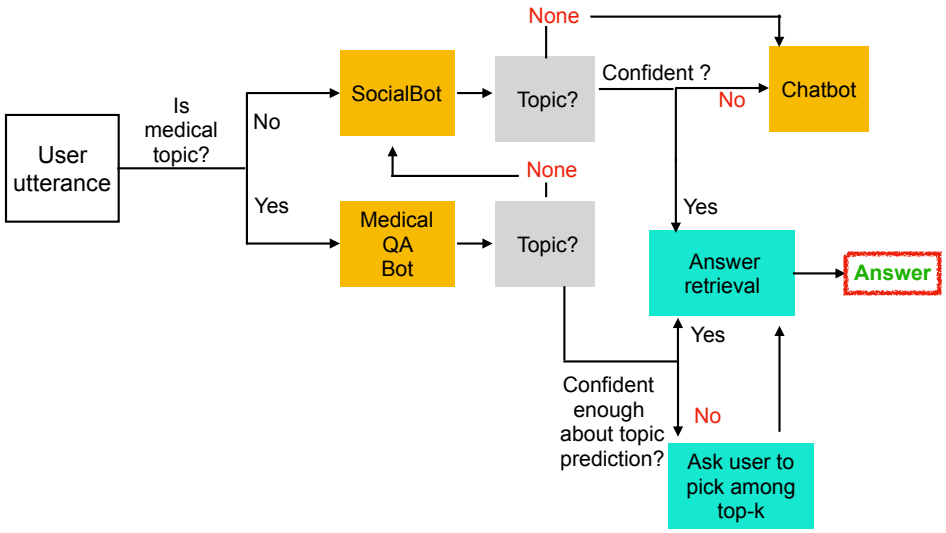

Figure 3: Our proposed pipeline. Section 5 has more details on the implementation of our current prototype.

Train	12082
Valid	3021
Test	615

Table 3: Medical Topic Classifier Training Data Information

Train accuracy	0.8812
Test accuracy	0.8358

Table 4: Medical Topic Classifier Evaluation Results

classifier. But we do not always need the confirmation. We set a threshold for the confidence score of the classifier. If the confidence score is higher than the threshold, meaning that our classifier is confident enough in the output, we will skip the confirmation question and retrieve the answer directly.

To make the classifier confidence scores more reliable, we use posterior calibration to encourage the confidence level to correspond to the probability that the classifier is correct (Chuan Guo, 2017; Schwartz et al., 2020). The method learns a parameter, called temperature or T. Temperature is introduced to the output logits of the model as follows:

$$pred = arg_i max \frac{exp(z_i/T)}{\Sigma_j exp(z_j/T)} \quad (1)$$

$\{z_i\}$ is the logits of the model and T is the temperature that needs to be optimized. T is optimized on a validation set to maximize the log-likelihood.

Precision	0.7585
Recall	0.7621
F-1 score	0.7603
Accuracy	0.7597

Table 5: MedicalQA Retriever Evaluation Results

MedicalQA Retriever After we determine the topic of the user's input, we can retrieve the answer from the Medline Dataset. We split the paragraphs in Medline data into single sentences and label them with the topics they belong to. We train the retriever using the augmented Medline data. We split the dataset into train, validation and test set using the ratio 8:1:1. The current retriever is based on BERT NextSentencePrediction model. We use the score from the model to determine the rank of each answer, and concatenate top 3 as the response of the agent. The evaluation result is shown in Table 5.

5.5 MedicalQA Grounding with Human Avatar

Our initial version for the human avatar contains 49 key body parts for front and 33 key body parts for the back. The front and back body part keywords are shown in Table 8 and 9. As future work, our goal is a more complete avatar with a comprehensive list of body parts.

Example grounded answers in our prototype system are shown in Figures 4 and 5 .

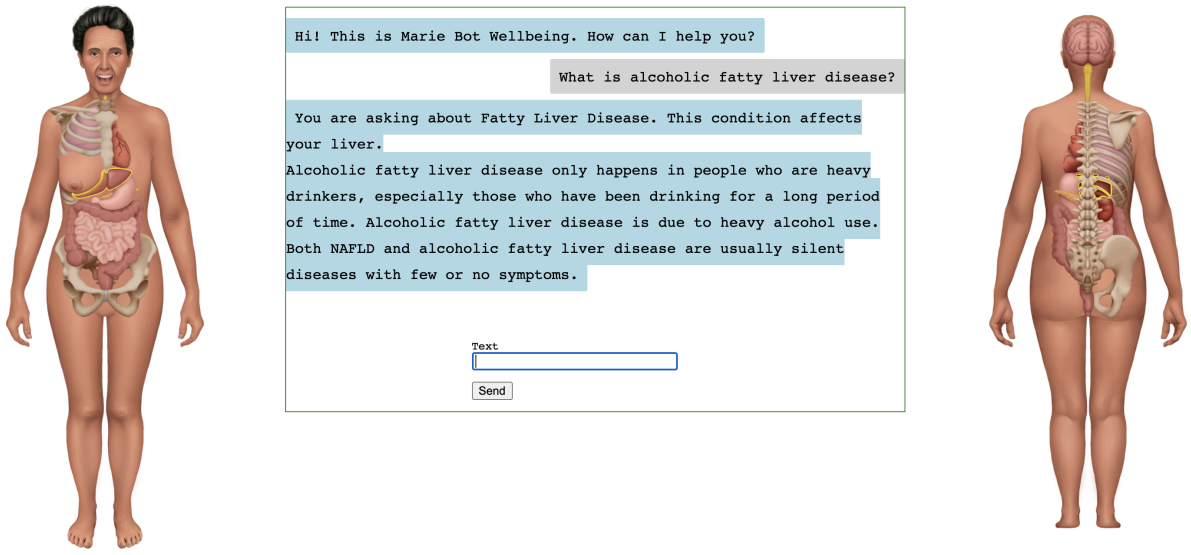

Figure 4: Human avatar visual answer example from our prototype: The affected body part, the liver, is highlighted on the avatar.

5.6 SocialBot Implementation

For our SocialBot, we currently have collected data from Reddit where each subreddits corresponds to a topic as shown in Table7. The topic classifier, posterior calibrator, and answer retriever are the same as in the MedicalQABot.

5.7 ChatBot Implementation

What is implemented is the last resort ChatBot, for which we have two versions: one is derived from a base language model, and another derived from a fine-tuned language model.

Language Models We use a large scale pre-trained language model, OpenAI GPT, as our base language model. We use the idea of transfer learning, which starts from a language model pre-trained on a large corpus, and then fine-tuned on end task. This idea was inspired by the huggingface convai project (Wolf, 2019).

Fine-tuning on Medical Dialogue Dataset: We use the Medical Dialogue Data(Zeng et al., 2020) to fine-tune the pre-trained language model. We use the questions as chat history and answers as current reply. The training set contains the portion from healthcaremagic and the test set the portion from icliniq

The evaluation results of our language model ChatBot are shown in Table 6.

	NLL	PPL
pre-trained model	5.4277	227.6291
fine-tuned model	3.2750	26.4423

Table 6: Language Model Evaluation. Negative log likelihood (NLL) and Perplexity (PPL)

6 Related Work

Medical Conversational Agents Academic and industry NLP research continues to push the frontiers of conversational agents, for example Meena from Google trained on a large collection of raw text (Daniel Adiwardana, 2020). In that work, it was found that end-to-end neural network with sufficiently low perplexity can surpass the sensibleness and specificity of existing chatbots that rely on complex, handcrafted frameworks. Medical dialogue has also been pursued from various angles for automatic diagnosis (Wei et al., 2018; Xu et al., 2019).

Grounding to Human Avatar IBM Research developed a human avatar for patient-doctor interactions (Elisseeff, 2007) with a focus on visualizing electronic medical records. By clicking on a particular body part on the avatar, the doctor can trigger the search of medical records and retrieve relevant information. Their focus on electronic medical records is different from our grounded medical question answering focus.

Another work (Charette, 2013) analyzed whether

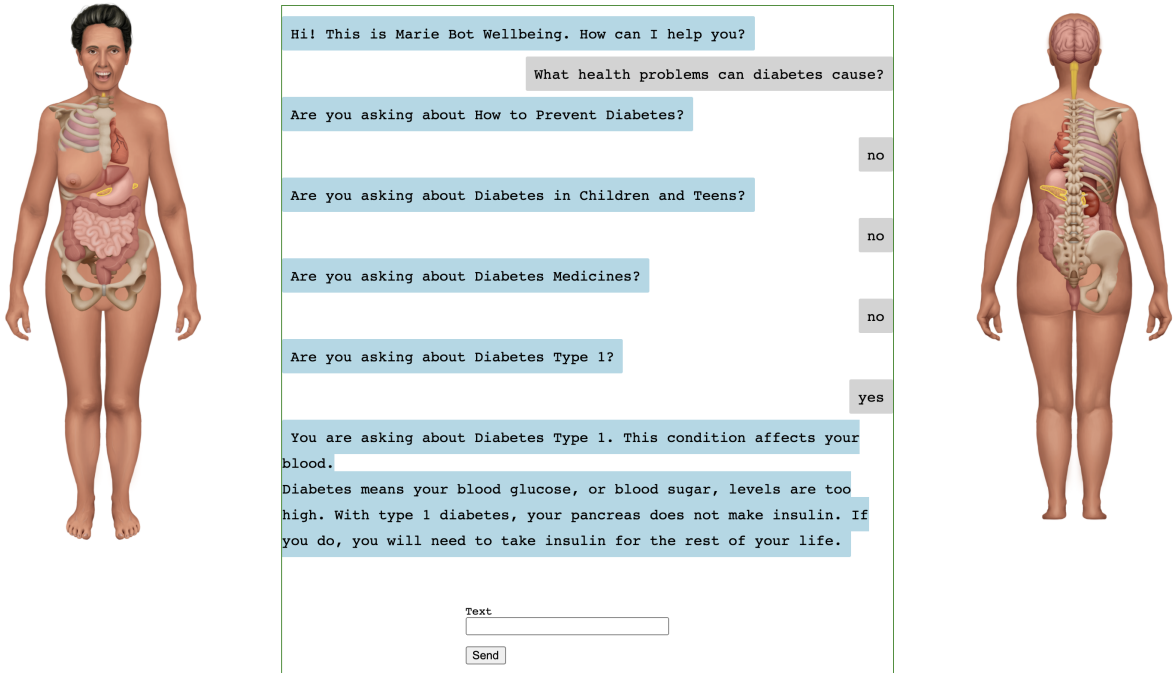

Figure 5: Human Avatar Visual Answer Example From our Prototype: Diabetes/Blood Sugar

and how the avatars help close the doctor-patient communication gap. This study showed that poor communication between doctors and patients often leads patients to not follow their prescribed treatments regimens. Their thesis is that avatar system can help patients better understanding the doctor's diagnosis. They put medical data, FDA data and user-generated content into a single site that let people search this integrated content by clicking on a virtual body.

7 Discussion

7.1 Technical Challenges

Quality and Quantity of Data In order for users to find the agent useful, and for the agent to really have a positive impact, we must provide answers to more questions. We need to extract more questions from a diverse set of reputable sources, while improving coverage.

Comprehensive Visualizations For the visualization, and human avatar grounding to be useful, a more comprehensive avatar is required, with all the parts that make up the human body. Medical ontologies such as the SNOMED CT part of Unified Medical Language System (UMLS)[4] contain a comprehensive list of the human body structures, which we can exploit and provide to a medical illustrator.

7.2 Ethical Considerations

Privacy When we deploy our system, we will respect user privacy, by not asking for identifiers. Additionally, we will store our data anonymously. Any real-world data will only accessible to researchers directly involved with our study.

False Information False or erroneous information in our data sources could lead our agent to present answers with potentially dire consequences. Our approach of only answering medical questions for which we have high quality, human curated answers seeks to address this concern.

System Capabilities Transparency Following prior work on automated health systems, our goal is to be clear and transparent about system capabilities (Kretzschmar et al., 2019).

8 Conclusion

We have presented a high level overview of the design philosophy of Marie Bot Wellbeing, a grounded, multi-interaction mode well-being conversational agent. The agent is designed to mitigate the limited adoption that plagues agents for healthcare despite patient interest. We reported details of our prototype implementation, and preliminary results.

[4] https://www.nlm.nih.gov/research/umls/index.html

There is much more to be done to fully realize Marie, which is part of our ongoing work.

References

Asma Ben Abacha and Dina Demner-Fushman. 2019. A question-entailment approach to question answering. *BMC Bioinform.*, 20(1):511:1–511:23.

Robert N. Charette. 2013. Can Avatars Help Close the Doctor-Patient Communication Gap?

Yu Sun Kilian Q. Weinberger Chuan Guo, Geoff Pleiss. 2017. On Calibration of Modern Neural Networks. arXiv:1706.04599v2.

David R. So Daniel Adiwardana, Minh-Thang Luong. 2020. Towards a Human-like Open-Domain Chatbot. arXiv:2001.09977v3.

Liesje Donkin, Ian B Hickie, Helen Christensen, Sharon L Naismith, Bruce Neal, Nicole L Cockayne, and Nick Glozier. 2013. Rethinking the dose-response relationship between usage and outcome in an online intervention for depression: randomized controlled trial. *Journal of medical Internet research*, 15(10):e231.

Andre Elisseeff. 2007. IBM Research Unveils 3D Avatar to Help Doctors Visualize Patient Records and Improve Care.

Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. 2017. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): a randomized controlled trial. *JMIR mental health*, 4(2):e19.

Kira Kretzschmar, Holly Tyroll, Gabriela Pavarini, Arianna Manzini, Ilina Singh, and NeurOx Young People's Advisory Group. 2019. Can your phone be your therapist? young people's ethical perspectives on the use of fully automated conversational agents (chatbots) in mental health support. *Biomedical informatics insights*, 11:1178222619829083.

Evan Mayo-Wilson. 2007. Internet-based cognitive behaviour therapy for symptoms of depression and anxiety: a meta-analysis. *Psychological medicine*, 37(8):1211–author.

Mari Ostendorf. 2018. Building a socialbot: Lessons learned from 10m conversations. *NAACL Keynotes*.

Roy Schwartz, Gabriel Stanovsky, Swabha Swayamdipta, Jesse Dodge, and Noah A. Smith. 2020. The right tool for the job: Matching model and instance complexities. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 6640–6651. Association for Computational Linguistics.

John Torous, Rohn Friedman, and Matcheri Keshavan. 2014. Smartphone ownership and interest in mobile applications to monitor symptoms of mental health conditions. *JMIR mHealth and uHealth*, 2(1):e2.

Zhongyu Wei, Qianlong Liu, Baolin Peng, Huaixiao Tou, Ting Chen, Xuan-Jing Huang, Kam-Fai Wong, and Xiang Dai. 2018. Task-oriented dialogue system for automatic diagnosis. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 201–207.

Thomas Wolf. 2019. How to build a State-of-the-Art Conversational AI with Transfer Learning.

Lin Xu, Qixian Zhou, Ke Gong, Xiaodan Liang, Jianheng Tang, and Liang Lin. 2019. End-to-end knowledge-routed relational dialogue system for automatic diagnosis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7346–7353.

Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang, et al. 2020. Meddialog: A large-scale medical dialogue dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9241–9250.

A Appendix

SubReddit	Question Num
AskPhotography	996
NoStupidQuestions	912
AskHistorians	985
askscience	998
AskWomen	525
AskReddit	925
AskUK	781
AskMen	200
AskCulinary	998
AskEconomics	560
AskAnAmerican	850
AskALiberal	830
askaconservative	775
AskElectronics	842
Ask_Politics	999
AskEngineers	912
askmath	999
AskScienceFiction	652
AskNYC	994
AskTrumpSupporters	357
AskDocs	684
AskAcademia	987
askcarsales	995
askphilosophy	981
AskSocialScience	487
AskEurope	844
AskLosAngeles	400
AskNetsec	995
AskFeminists	978
AskWomenOver30	838

Table 7: Number of questions we extracted from each SubReddit

ankle	arm	breast
cheeks	chin	collar bone
ear lobe	ear	elbow
eyebrows	eyelashes	eyelids
eyes	finger	foot
forehead	groin	hair
hand	heart	hip
intestines	jaw	knee
lips	liver	lungs
mouth	neck	nipple
nose	nostril	pancreas
pelvis	rectum	ribs
shin	shoulder blade	shoulder
spinal cord	spine	stomach
teeth	thigh	throat
thumb	toes	tongue
waist	wrist	

Table 8: Human Avatar Front Body Parts

ankle	anus	arm
back	brain	buttocks
calf	ear lobe	ear
elbow	finger	foot
heart	intestines	kidney
knee	liver	lungs
neck	palm	pancreas
pelvis	rectum	ribs
scalp	shoulder blade	shoulder
spinal cord	spine	stomach
thigh	thumb	wrist

Table 9: Human Avatar Back Body Keywords. Some body parts can be visualized from both the front and back.