MEDCOD: A Medically-Accurate, Emotive, Diverse, and Controllable Dialog System

Rhys Compton∗
Ilya Valmianski
Li Deng
Costa Huang†
Namit Katariya
Xavier Amatriain
Anitha Kannan
Curai

Abstract

We present MEDCOD, a Medically-Accurate, Emotive, Diverse, and Controllable Dialog system with a unique approach to the natural language generator module. MEDCOD has been developed and evaluated specifically for the history taking task. It integrates the advantage of a traditional modular approach to incorporate (medical) domain knowledge with modern deep learning techniques to generate flexible, human-like natural language expressions. Two key aspects of MEDCOD’s natural language output are described in detail. First, the generated sentences are emotive and empathetic, similar to how a doctor would communicate to the patient. Second, the generated sentence structures and phrasings are varied and diverse while maintaining medical consistency with the desired medical concept (provided by the dialogue manager module of MEDCOD). Experimental results demonstrate the effectiveness of our approach in creating a human-like medical dialogue system. Relevant code is available at https://github.com/curai/curai-research/tree/main/MEDCOD

1. Introduction

The development of natural language (NL) understanding and dialogue systems, both spoken and text-based, has over 30 years of history and can be divided into three generations according to the disparate styles of system design: (1) use of expert systems based on symbolic-rules and templates (Allen et al., 1996; Rudnicky and Xu, 1999), (2) use of (shallow) statistical learning (Wang et al., 2005; Tur and Deng, 2011; Wang et al., 2011), and (3) use of deep learning (Tur et al., 2018; Dhingra et al., 2016). The earlier two generations of dialogue systems were usually designed with a number of separate modules: textual (or spoken) natural language understanding (NLU), dialogue manager, natural language generation (NLG), and (optionally) spoken language generation. The main advantage of modular designs is their ability to easily incorporate domain knowledge. The main disadvantage is their weakness in generating flexible, human-like responses. The third-generation language understanding and dialogue systems, driven by deep learning technology (Hinton et al., 2012; Huang et al., 2013; Deng and Yu, 2014; Deng, 2016; Deng and Liu, 2018), adopt the end-to-end neural network architecture approach. This provides the opportunity to learn all parts of the dialogue system jointly and the ability to produce more human-like responses (Vinyals and Le, 2015; Chen et al., 2016; Wu et al., 2020; Hosseini-Asl et al., 2020). The main weakness, however, is that end-to-end learned systems require large amounts of training data to implicitly acquire domain knowledge and suffer from poor control over the system’s output.

In this paper, we present a hybrid modular and deep learning approach to designing a medical dialogue system targeted for the history taking task called MEDCOD, which integrates domain knowledge and controllability from a modular design with human-like NLG of a deep learning system. Medical dialogues between patients and doctors are one key source of information for diagnosis and decision making (Chen et al., 2020; Soltau et al., 2021). No-
tably, the task of history taking in such medical dialogues is an important, time consuming but, in many circumstances, low-complexity part of a medical encounter as it involves asking a series of closed-ended questions to ascertain the patient’s current condition; this makes it a prime target for automation, decreasing the clerical load on physicians and allowing them to practice at the top of their scope. Other clinical use cases include AI-driven online symptom checkers and automatic patient triaging.

In our history-taking dialogue system, the dialogue manager uses both an expert system and a machine learned emotion classifier to control a deep-learning-based NLG module. The expert system uses a medical knowledge base (KB) (c.f. Miller and Masarie Jr. (1989)) that contains rich medical domain knowledge to identify which patient-reportable medical finding should be asked next. The emotion classifier then predicts the emotion with which the NLG module should ask the question. The NLG module is implemented using a deep learning approach to generate variable medical questions while maintaining medical consistency with the expert-system-derived finding, while containing emotion-classifier specified emotion.

The technical contributions of this paper are as follows. First, we developed a novel method of using “control codes” (Keskar et al., 2019) to within the medical dialogue data for training DialoGPT (dialogue generative pre-trained transformer) (Zhang et al., 2020), which serves as the NLG model in our dialogue system. This use of control codes aims to maintain medical consistency in the generated questions while creating diversity that exhibits human-like attributes. Second, we train an emotion classifier for use in the inference stage of NLG. This gives our system the ability to generate emotive sentences simulating human doctors’ behavior. Finally, to overcome the problem of sparsity in the dialogue training data, we made effective use of GPT-3 (Brown et al., 2020) to augment finding-NL paired data jointly for both diversity and emotion while maintaining medical consistency in the NL output. With these technical innovations, we have built MEDCOD, the first medically consistent and controllable history taking dialogue system with human-like NL expression as the system output in each dialogue turn.

2. Related Work

The task-oriented dialog system is one of four major types of dialogue systems in common use, the other three types being for non-goal-oriented applications in information consumption, decision support, and social interactions; see a review in (Celikyilmaz et al., 2018). This paper is devoted to task-oriented systems only, for the task of medical history taking.

The classic task-oriented dialog systems incorporate several components including Speech Recognition, Language Understanding, Dialog Manager (consisting of State Tracker and Dialog Policy), NLG, and Speech Synthesis. In the preliminary development, MEDCOD has not yet incorporated the speech components, although this would have a positive impact on further usability (He and Deng, 2013; Huang and Deng, 2010; He et al., 2011; Deng and O’Shaughness, 2003; Yu and Deng, 2015).

Until recently, a majority of task-oriented dialogue systems were based primarily on hand-crafted rules (Aust et al., 1995; Simpson and Eraser, 1993) or (shallow) machine learning techniques for all major components of the systems (Gorin et al., 1997). This work formulated the dialogue as a sequential decision making problem based on Markov decision processes and reinforcement learning (Young et al., 2010). With the introduction of deep learning in speech recognition, spoken language understanding, and dialog modeling, incredible successes were demonstrated in the robustness and coherency of dialog systems, especially in the NLU component of the system. (Tur et al., 2012; Mesnil et al., 2015; Hakkani-Tür et al., 2016; Li et al., 2016; Lipton et al., 2016).

The approach we take in developing our current medical dialogue system has been motivated by successes in the related work discussed above. Our work is also inspired by Paranjape et al. (2020), who demonstrated an open-domain, non-target-oriented dialogue system capable of empathetic conversations with emotional tone. Further, we have benefited from the work of (Keskar et al., 2019), which proposed the use of “control codes” to construct a generic language model for controllable NLG. We have used a similar mechanism to train the NLG module of our domain-specific medical dialogue system to maintain medical accuracy while generating diverse NL sentences. Finally, the approach to few-shot NLG using structured knowledge by Chen et al. (2019) relates to our work in that we have also made use of a specific medical knowledge base (KB) (as part of our dialogue manager) to provide control over the NLG model; one key difference is that our KB is used to create medical concepts for their diverse NL expressions rather than to provide better generalization across domains.
Medical dialogue systems have been reported in the literature only in recent years, with focus on NLU for clinical documents. Prior to the work presented in this paper, however, there does not appear to be any previous work that focused on NLG for medical dialogue systems. Enarvi et al. (2020) developed a system to generate NL medical reports based on patient-doctor dialogue transcripts but they did so from speech recognition outputs instead of from the semantic representation of a dialogue manager as in our work presented here. Below we briefly review the related work on NLU, which can be considered as complementary work to our current focus.

In the area of mapping extracted medical concepts in conversations to a knowledge base, (Du et al., 2019a,b) introduced a hierarchical two-stage approach to infer clinical entities (e.g., symptoms), their properties (e.g., duration), and relationships between them. Selvaraj and Konam (2019) focused on the problem of treatment regimen extraction while Khosla et al. (2020) studied the problem of extracting relevant task-relevant utterances from medical dialogues. The problem of intent detection in doctor-patient interactions is a new area and was recently explored in (Rojowiec et al., 2020). Further, Soltau et al. (2021) recently reported a spoken medical dialogue system aiming to extract clinically relevant information from medical conversations between doctors and patients. But the work did not proceed further to use the extracted information as input to a dialogue manager and then to produce the NL response, which our MEDCOD approach does. Bo et al. (2019) introduces a medical knowledge-based dialogue system that acts as a health assistant. Lin et al. (2020) explored a natural paradigm for low-resource medical dialogue generation using very small amount of data for adaptation. Finally, Liu et al. (2019) presented a domain-aware automatic chest X-ray radiology report generation system. All the above work benefited the design of the NLG module of our MEDCOD system.

3. Approach
We present details of MEDCOD, our medical dialogue system for history taking which combines expert-system-driven structured history taking (i.e. generate “Next Finding” using a medical KB) with deep-learning-driven emotion classification and controllable NLG. This integration allows us to use the expert system to determine “what” to ask (by the system to the user) in an explainable and auditable way, and to use the deep-learning components to determine “how” to ask with human-like natural language. To enable this separation of “what” and “how” we developed a NLG module in the dialogue system that uses control codes provided by the expert system and the emotion understanding component to guide the formation of the NLG module’s output. Figure 1 provides overview of MEDCOD. Its Dialogue manager consists of three components:

1. **Conversation tracker**: this component tracks patient demographic information, reason for encounter, what findings have been reported by the patient, the text of the previous questions, and the text of patient responses, which our MEDCOD approach does. Bo et al. (2019) introduces a medical knowledge-based dialogue system that acts as a health assistant. Lin et al. (2020) explored a natural paradigm for low-resource medical dialogue generation using very small amount of data for adaptation. Finally, Liu et al. (2019) presented a domain-aware automatic chest X-ray radiology report generation system. All the above work benefited the design of the NLG module of our MEDCOD system.

2. **Next Finding**: this expert-system component takes patient demographics and patient findings and generates the target finding (next finding control code) to be asked next by the NLG. NLG component uses previous findings as well as control codes for the target finding and emote to generate a human-like NL question about the target finding (§ 3.2). This component was trained using the Emote dataset (§ 4.3.1).

3. **Emotion Classifier**: this model takes the conversation context and predicts the appropriate emote (emote control code) to be used by the NLG model (§ 3.1). This component was trained using the Emote dataset (§ 4.3.1).
3.1. Emotion Classifier

During medical history taking, the patient may provide sensitive or emotionally charged information (e.g. severe pain); it is imperative that an automated dialogue system reacts and emotes appropriately to this information, similarly to how a human doctor would (e.g. “Oh that’s unfortunate...”). When analyzing patient-provider medical conversations, we identified four broad classes of emote control codes that reflect emotional phrasing medical professionals use when talking with their patients. The control codes are Affirmative, Empathy, Apology, and None (see § 4.3.1 for more details). The goal of the emotion classifier is to predict the emote control code based on the conversational context. The conversational context contains three pieces of information: (1) previous question (2) patient response, and (3) target finding (which is the output of Next Finding module).

The model consists of embedding the contexts independently (using a pretrained model) to capture the semantics of the entirety of text, and then learning a linear layer of predictors, on reduced dimensionality, over the emote control codes. We first independently embed the three pieces of context using Sentence-BERT (SBERT) (Reimers and Gurevych, 2019) that takes as input a variable-size string (up to 128 tokens) and outputs a fixed-size vector containing semantically-meaningful feature values. We then apply principal component analysis (PCA) (Pearson, 1901; Hotelling, 1933) to the embeddings of each input type independently and then concatenate the embeddings. Finally, we train a logistic regression classifier over the four emote control code classes. The model is trained on the Emote dataset (§ 4.3.1).

3.2. Natural Language Generator

We developed the domain-specific NLG module of MEDCOD with three key constraints and goals:

1. **Medical Consistency**: Generated questions by the system must ask only about the target finding (e.g. if the target finding is “abdominal pain” then an acceptable question would be “Is your belly hurting?” while an unacceptable question would be “Do you have severe abdominal pain?”).

2. **Phrasing Diversity**: Generated questions must present phrasing variability, a major improvement over using templated questions (e.g. if the target finding is “abdominal pain”, instead of asking every time “Is your belly hurting?”, the model can generate alternative paraphrases such as “Do you have pain in your belly?”)

3. **Emotional Awareness**: Generated questions must be empathetic when appropriate. When gathering pertinent findings from the patient, we would like the NLG output to emote appropriately based on the context: did the patient say anything particularly difficult that we should empathize with (e.g. a patient complaining about severe pain)? Are we about to ask a highly relevant (for a presenting symptom) but sensitive question (e.g. checking with the patients if they have multiple sexual partners)?

We achieve these three goals simultaneously by fine-tuning a pretrained DialoGPT model (Zhang et al., 2019). In the fine tuning process, we use control codes for dialogue prompts to help guide the NLG output (Keskar et al., 2019) at inference-time. Apart from the control codes, we also prompt with the previous findings, patient’s age and gender, as well as patient’s reason for visit. The full control codes consist of the next finding control code and the emote control code. At training time, we use the Medical Conversations dataset (§ 4.3.2). At inference time, the control codes are generated by the dialogue manager: next finding control code comes from the Next Finding module while the emote control code comes from the Emotion Classifier.

4. Datasets

The development of our medical dialogue system relies on a number of datasets. The process for constructing these datasets is presented in Figure 2(a). We start with two preexisting datasets: KB and Doctor Edits dataset. The KB is used to generate two additional intermediate datasets, the Structured Clinical Cases dataset (§ 4.2.1) and the Paraphrased Questions dataset (§ 4.2.2). The Doctor Edits dataset is used to construct the Emote dataset (§ 4.3.1). Finally the Structured Clinical Cases, Paraphrased Questions, and Doctor Edits datasets are combined to produce the Medical Conversations dataset (§ 4.3.2). The final datasets consist of the Emote dataset and Medical Conversations dataset.
MEDCOD: A Medically-Accurate, Emotive, Diverse, and Controllable Dialog System

(a) Schematic of the dataset construction.

(b) Examples illustrating training set construction. We append control codes during training to condition the models on them, which can then be used for guiding NLG during inference.

Figure 2: Medical conversation and emote datasets construction.

4.1. Pre-existing datasets

4.1.1. Knowledge Base
The knowledge base (KB) is an AI expert system similar to Quick Medical Reference (QMR) (c.f. Miller and Masarie Jr., 1989) that is kept up-to-date by a team of medical experts. Currently, it contains 830 diseases, 2052 findings (F) (covering symptoms, signs, and demographic variables), and their relationships. It also contains human generated patient-answerable questions for ascertaining the presence for every finding. Finding-disease pairs are encoded as evoking strength (ES) and term frequency (TF), with the former indicating the strength of association between the constituent finding-disease pair and the latter representing the frequency of the finding in patients with the given disease. The inference algorithm of KB computes differential diagnosis and also facilitates the next finding to ask using Next Finding module.

4.1.2. Doctor Edits dataset
This is a small in-house dataset containing 3600 instances of doctor-edited questions as well as doctor and patient dialogue turns preceding the doctor-edited questions. For training the emotion classifier, we perform a random 80/20 train/test split.

4.2. Intermediate datasets

4.2.1. Structured Clinical Cases dataset
Structured Clinical Cases dataset contains simulated cases that consist of patient demographic data (age, sex), the chief complaint also known as reason for encounter (RFE) and a set of findings with “present” or “absent” assertion. The data is produced using an in-house simulator, Structured Clinical Case Simulator (SCCS). SCCS uses the KB as a clinical decision support system. Unlike previous works that simulate clinical cases for a desired diagnosis (Parker and Miller, 1989; Kao et al., 2018; Ravuri et al., 2018; Kannan et al., 2020), SCCS starts with a finding and goes through the process of history taking to lead up to a diagnosis: It first samples demographic variables and a finding \( f \in F \) that serves as the chief complaint/RFE. Then, it computes the differential diagnosis (distribution over the diseases given the finding) using the underlying expert system and then samples findings taking into account the differential diagno-
sis. The newly sampled finding is asserted randomly to be present \( f_{pos} \) or absent \( f_{neg} \) with a slight bias to absent. If asserted as present, then the findings that are impossible to co-occur are removed from consideration (e.g., a person cannot have both productive and dry cough). The next iteration continues as before: computing differential diagnosis and then identifying next best finding to ask. The simulation for a case ends when a random number (5-20) of findings are sampled or the difference in score between the first and second ranked diagnosis is at least 20 (a desired minimum margin under expert system). Simulated samples with margin higher than that are added to the Structured Clinical Cases dataset. See § A for an example simulated clinical case.

### 4.2.2. Paraphrased Questions Dataset

The Paraphrased Questions dataset contains findings and an associated set of questions, these questions being different ways (paraphrases) to ask about the finding (for examples see Table B.1). The goal of this dataset is to imbue variability into the NLG model with examples of different question phrasings.

We use carefully primed GPT-3 (Brown et al., 2020) to generate a large number of candidate questions for each finding. We curate a small but diverse set of thirty findings and manually paraphrase the expert-written question already available from the KB. We randomly sample 10 findings for priming (Liu et al., 2021; Chintagunta et al., 2021) GPT-3 to paraphrase new unseen findings. See Appendix I for example invocations. To restrain the generations but still acquire a diverse set of paraphrases, we limited the output to a single paraphrase at a time. We repeatedly invoke GPT-3 (\( \text{temp}=0.65 \)) until we have the desired number of distinct paraphrases, each time priming GPT-3 with random sample of ten findings.

The key strength of this approach is the minimal human effort required; only a single manually-written paraphrase is required for each finding in our small set, which is then used as guidance for GPT-3 to mimic the task on new findings. We manually validate the candidate questions using in-house medical professionals by asking them to label if the candidate question is medically consistent with the target finding, and keep only those that are. We achieved 78% correctness of finding-question pairs. Analysing the failure cases, we found that the error was either due to minor grammar issues or bad timing (i.e., "Are you sleepy?", which implies right now as opposed to intermittently throughout the day). We collected question paraphrases for the 500 most frequent findings in the Structured Clinical Cases dataset.

### 4.3. Final datasets for training

#### 4.3.1. Emote Dataset

The Emote dataset contains a set of emote phrases, their corresponding emote control codes, and patient and medical professional dialogue turns that preceded the use of the emote phrase. This dataset is directly used to train the Emotion Classifier (§ 3.1). The emote phrases are directly extracted from medical professional messages, while the emote control codes are manually assigned to each emote phrase.

We mined the Doctors Edits dataset for medical professional messages that express emotion and identified three broad classes of emote control codes:

- **Affirmative:** A neutral confirmation of the user’s response, flexible to be used in many conversation situations (e.g. thanks for the input).
- **Empathy:** A more emotionally charged response, implying something negative/painful about the conversation.
- **Apology:** This is an apology for asking a personal or sensitive question (e.g., "Sorry for asking a personal question, do you have multiple sexual partners?").

We also included None as an emotion code to reflect no emotion is added. We associated with each emote code a set of emote phrases that are frequently used by medical professionals to express these codes. We provide additional details on data mining for emotion from conversations and examples of emote phrases corresponding to codes in Appendix C.

#### 4.3.2. Medical Conversations Dataset

The Medical Conversations dataset consists of dialogue context, next finding and emote control codes, and medical finding questions with emotional responses; the NLG model is trained on this dataset.

Figure 2(b) provides a walk through for constructing Medical Conversation samples. To generate dialogue context we sample from the Clinical Cases dataset. From this we extract the RFE, patient demographic information, target finding, and the structured preceding finding and patient response. The preceding finding and the target finding are converted into a preceding medical professional question and
5. Experimental Results

In this section we present both subjective and objective evaluation results which robustly demonstrate the improved output from MEDCOD when compared to counterparts that use the fixed-template approach to asking questions or can not emote.

5.1. Experimental Setup

5.1.1. Training details

NLG: We use a pretrained DialoGPT-Medium from HuggingFace (Zhang et al., (accessed July, 2021) as our underlying NLG model. We train on 143,600 conversational instances, where each instance has only one previous conversation turn as context. We use a batch size of 16 with 16 gradient accumulation steps for an effective batch size of 64, for 3 epochs with a learning rate of 1e-4 and ADAM optimiser.

Emotion Classifier: We apply pretrained paraphrase-mpnet-base-v2 SBERT for embedding the conversational contexts. The Logistic Regression model is trained with C=10 and class re-weighting (to compensate for the data skew of the training data § 4.3.1). PCA is applied down to 70 components.

5.1.2. Ablations of MEDCOD

We ablate MEDCOD by varying data/control codes supplied to each underlying NLG model during training, with all other parameters kept consistent. This allows us to understand the importance of variability, medical consistency and ability to emote. We use Expert to denote the variant of MEDCOD in which the NLG module is trained only on expert questions (single question per finding). MEDCOD-no-Emote’s NLG module is trained on the Medical Conversations dataset (§ 4.3.2) with paraphrases but no emote codes while MEDCOD is our feature-complete dialog system trained on the entire Medical Conversations dataset including emote.

5.2. End2End Evaluation: Main Results

This is our main evaluation that is targeted at understanding if the patient experience on the end-to-end system can be improved by providing them with a more natural conversational dialog. For this, we instantiate two identical medical dialog interfaces with different driving systems: Expert and MEDCOD. A set of 30 commonly occurring chief complaints along with demographic information such as age and gender were collected from a telehealth platform. We recruited five medical professionals for the labeling task such that each medical professional will go through the conversational flow for 18 chief complaints, giving three labels for each case.

Labeling instruction: While the focus is on patient experience, we engaged medical professionals because of the dual patient/doctor role-play for this evaluation. When they start on a chief complaint, they were to choose a relevant final diagnosis and answer questions to substantiate that final diagnosis; this ensures that the sequence of questions asked during conversation are clinically grounded. While doing so, they were also acting as a patient, answering and responding (e.g., by volunteering extra information) as someone presenting with the symptoms would. The medical professionals converse simultaneously with the Expert and MEDCOD systems (the UI presents an anonymized A/B label) by providing identical answers for each conversation step between the two interfaces (but different answers between steps).

Once they perform 10 question responses or the conversation terminates (due to reaching a diagnosis), the encounter is over and they label each instance as follows: Enter a 1 for either Model A or Model B, based on how you think a new patient using the service for an ailment would feel. If you prefer the encounter with Model A, enter a 1 in the Model A column and 0 in the Model B column, and vice-

|                | A   | B   | Equal |
|----------------|-----|-----|-------|
| Total Pts      | 63  | 30  | -     |
| Mut. Excl. Pts | 49  | 16  | 25 (54.4%) |
| Mut. Excl. Pts | 20  | 2   | 8 (66.7%) |

Table 1: End2End Evaluation comparing between MEDCOD (A) and Expert (B)
versa if you prefer Model B. To avoid spurious ratings when the two models are very similar, we also allowed the same grade to be given to both models if they were equally good/bad, but required a comment explaining the decision.

Results: Table 1 shows the evaluation results (refer to Appendix H.1 for a full conversation example). When simply summing up the scores, MEDCOD achieves a score of 63 (max 90) — over twice as high as Expert. When separating scores into instances where one model is picked exclusively over the other or both are rated equally (“Mut. Excl. Pts”), we see a similarly strong result for our model; in over half the conversations enacted, MEDCOD is preferred holistically over Expert, while only 17.8% of the time is Expert preferred. When we inspect the difference, it’s often the case that MEDCOD emoted with Affirmative when not emoting (None) would have been more appropriate.

We also considered majority voting for each of the patient complaints, which shows an even more exaggerated improvement by our model. Two-thirds of the time (66.7%), MEDCOD is exclusively preferred over Expert, while only 6.6% of the time latter is preferred. In roughly one-fourth of the chief complaints, both models are rated equally.

In the majority of cases, MEDCOD is preferred, indicating that within an automated dialog situation, the contributions discussed in this paper provide a marked improvement to patient experience.

5.3. NLG Module Evaluation

The goal is to evaluate MEDCOD and its ablations individually along three important aspects for medical dialog (c.f. (Rashkin et al., 2018)):

1. Medical Consistency: How well does the question capture the clinical implication of the target finding?

2. Fluency: How fluent/grammatically correct is the candidate question?

3. Empathy: How empathetic/emotionally appropriate is the candidate question, given the conversational context and the finding to ask next?

To collect the data for this evaluation, we begin with an in-house dataset of conversations from a tele-health platform. We decompose each conversation into three-turn instances (same form as Emote Dataset §4.3.1), then attach an emotive control code to each instance by performing prediction with the Emotion Classifier. To exaggerate the difference between instances, we only keep instances where the predicted class’ probability > 0.8. We then randomly sample 25 instances from each of the four predicted classes to create our final set of 100 evaluation instances. Finally, we generate a candidate question for the instances by passing the conversation context to each of the model variants for generation. A team of five medical professionals label each example along each of three axes on a scale of 1 to 5.

Results: Table 2 provides the comparative results. MEDCOD scored significantly higher in Empathy, showing that the Emote additions improve human-evaluated empathy in a significant way. This result also indicates that the emotive code is appropriately predicted by the Emotion Classifier.

There are many correct ways to query a finding, however the Expert model is trained on data with precisely one way, which is expert-annotated, so is likely to have optimal medical consistency (and also be the most fluent for the same reason). Because of this, we view Expert as close to best performance achievable along Medical Accuracy and Fluency. MEDCOD and MEDCOD-no-Emote are still comparable to Expert indicating that the variations in how questions are framed do not significantly affect medical accuracy or fluency. As expected, given that its impossible to encode empathy preemptively (in the expert-annotated or paraphrased questions), Expert and MEDCOD-no-Emote score low on empathy. Note that it is not always necessary to emulate, hence they receive non-zero score.

Appendix G provides qualitative examples of comparing generations from model variants where MEDCOD is expressive and incorporates empathy.

5.4. Emotion Classifier Evaluation

Evaluating emotion is difficult as it is subjective and can be multi-label: in a situation, there may be multi-
MEDCOD: A Medically-Accurate, Emotive, Diverse, and Controllable Dialog System

Figure 3: Emotion Classifier Evaluation

ple “correct” ways to emote so comparing predictions to a single ground-truth label (i.e., physician’s emote) is unlikely to give an accurate notion of performance. We instead measure the emotional appropriateness using a small team of medical professionals.

For each instance in the Emote dataset test split, we pass the predicted emote control codes to a team of three medical professionals. They are tasked with labelling whether the emote code is appropriate, given the previous context in a conversation (input to our model). When the emote is not appropriate, an alternate emote is suggested by the labeller. We use majority voting on this data to obtain the final label, creating an alternate human-augmented test set. We evaluate the model’s predictions against this human-augmented test set; Figure 3 shows the confusion matrix on the full Emote dataset test split and Table D in Appendix shows the complete results.

On the human-augmented test set, our Emotion Classifier reached 0.9 accuracy with macro-F1: 0.8 and PR-AUC: 0.69. Looking at precision/recall statistics, each non-None emote class (and the model predictions overall) achieved precision >= recall, which is desired due to the high cost of inappropriate emoting; we want high confidence when we actually emote something, otherwise we should safely emote None. It should be noted that we did not tune the classification boundary but simply took the max-probability class as prediction; a high prediction threshold (e.g., 0.8) would further increase precision.

The confusion matrix (Figure 3) illustrates a similarly strong picture of our Emotion Classifier’s performance; the large majority of non-None predictions are correct. There are 54 ‘incorrect’ None predictions by our Emotion Classifier, however, these are low- or zero-cost mispredictions, as it is always appropriate not to emote (the same cannot be said for Empathy, for example, where it can be very costly to empathise when not appropriate).

We also analyzed how conversational context affects predicted emotion. See Appendix E for complete attribution details. We find that empathy is strongly influenced by the previous patient response, apology by the next question and affirmative by all three parts of the input.

6. Conclusion & Future Work

We introduced MEDCOD\(^1\), a novel approach to developing medical dialog systems, which combines the traditional modular design with a deep learning based NLG model. In particular, our approach allows us to incorporate, in a medically consistent fashion, the knowledge of medical findings and an appropriate emotional tone when generating human-like NL expressions through the use of their respective control codes, both provided by the dialog manager. The highly positive experimental results presented demonstrate the effectiveness of our approach.

Our current approach still leaves a number of open problems. A broad problem for making the dialogue even more human-like is the effective use of implied references; as an example, when a patient mentions they have diarrhea, the system would ask them “Is it bloody?” as a follow-up question, as opposed to “Do you have bloody diarrhea?”. Alternatively, historical context references can be used: asking “You mentioned you have headache. Is this recurrent?” instead of “Do you have a recurrent headache?”.

Another area of improvement in our current system is increasing the granularity of emotion classes. This may include both the addition of new classes and incorporating a notion of emotional intensity, e.g., “OK” has a different strength than “Thank you!”

In the longer term, our dialogue system should accommodate speech recognition and speech generation. This presents a new class of challenges such as emotion detection from speech and synthesis of properly empathetic speech. Multimodal inputs beyond text and speech may further advance the functionality of dialogue systems, especially in the user-interface aspect (Yu and Deng, 2008). For example, the video-feed of a patient may present additional opportunities for refining our dialogue manager module by collecting non-verbal cues from the patient.

\(^1\) https://github.com/curai/curai-research/tree/main/MEDCOD
References

James F. Allen, Bradford W. Miller, Eric K. Ringger, and Teresa Sikorski. A robust system for natural spoken dialogue. In Proceedings of the Annual Meeting of the Association for Computational Linguistics, pages 62–70, 1996.

H. Aust, M. Oerder, F. Seide, and V. Steinbiss. The philips automatic train timetable information system. Speech Communication, 17:249–262, 1995.

Lin Bo, Wenjuan Luo, Zang Li, and Xiaqing Yang. A knowledge graph based health assistant. AI for Social Good workshop at NeurIPS, Vancouver, Canada., 2019.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. 2020.

Asli Celikyilmaz, Li Deng, and Dilek Hakkani-Tür. Deep learning in spoken and text-based dialogue systems. Chapter 3 of Book: Deep Learning in Natural Language Processing, 2018.

Shu Chen, Zeqian Ju, Xiangyu Dong, Hongchao Fang, Sicheng Wang, Yue Yang, Jiaqi Zeng, Ruisi Zhang, Ruoyu Zhang, Meng Zhou, Penghui Zhu, and Pengtao Xie. Meddialog: A large-scale medical dialogue dataset. CoRR, 2020.

Yun-Nung Chen, Dilek Hakkani-Tür, Gokhan Tur, Jianfeng Gao, and Li Deng. End-to-end memory networks with knowledge carryover for multi-turn spoken language understanding. In Proceedings of Interspeech, 2016.

Zhiyu Chen, Harini Eavani, Yinyin Liu, and William Yang Wang. Few-shot NLG with pre-trained language model. 2019.

Jai Chintagunta, Namit Katariya, Xavier Amatriain, and Anitha Kannan. Medically aware gpt-3 as a data generator for medical dialogue summarization. Machine Learning for Healthcare, 2021.

Li Deng. Deep Learning: From speech recognition to language and multimodal processing. In APSIPA Transactions on Signal and Information Processing (Cambridge University Press), 2016.

Li Deng and Yang Liu, editors. Deep Learning in Natural Language Processing. Springer, Singapore, 2018.

Li Deng and Doug O’Shaughness. Speech Processing: A Dynamic and Optimization-Oriented Approach. Marcel Dekker Inc., 2003.

Li Deng and Dong Yu. Deep Learning: Methods and Applications. NOW Publishers, 2014.

Bhuwan Dhingra, Lihong Li, Xinjun Li, Jianfeng Gao, Yun-Nung Chen, Faisal Ahmed, and Li Deng. End-to-end reinforcement learning of dialogue agents for information access. arXiv preprint arXiv:1609.00777, 2016.

Nan Du, Kai Chen, Anjuli Kannan, Linh Tran, Yuhui Chen, and Izhak Shafran. Extracting symptoms and their status from clinical conversations. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2019a.

Nan Du, Mingqiu Wang, Linh Tran, Gang Li, and Izhak Shafran. Learning to infer entities, properties and their relations from clinical conversations. CoRR, 2019b.

Seppo Enarvi, Marilisa Amoia, Miguel Del-Agua Teba, Brian Delaney, Frank Diehl, Stefan Hahn, Kristina Harris, Liam McGrath, Yue Pan, Joel Pinto, Luca Rubini, Miguel Ruiz, Gagandeep Singh, Fabian Stemmer, Weiyi Sun, Paul Vozila, Thomas Lin, and Ranjani Ramamurthy. Generating medical reports from patient-doctor conversations using sequence-to-sequence models. In Proceedings of the First Workshop on Natural Language Processing for Medical Conversations, pages 22–30. Association for Computational Linguistics, July 2020.

A. L. Gorin, G. Riccardi, and J. H. Wright. How may i help you? Speech Communication, 23:113–127, 1997.

Dilek Hakkani-Tür, Gokhan Tur, Asli Celikyilmaz, Yun-Nung Chen, Jianfeng Gao, Li Deng, and Ye-Yi Wang. Multi-domain joint semantic frame pars-
MEDCOD: A Medically-Accurate, Emotive, Diverse, and Controllable Dialog System

using bi-directional rnn-lstm. In Proceedings of Interspeech, pages 715–719, 2016.

X.D. He, L. Deng, and A. Acero. Why word error rate is not a good metric for speech recognizer training for the speech translation task? In Proc. ICASSP, Prague, 2011.

Xiaodong He and Li Deng. Speech-centric information processing: An optimization-oriented approach. In IEEE, 2013.

Geoffrey Hinton, Li Deng, Dong Yu, George Dahl, Abdel rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara Sainath, and Brian Kingsbury. Deep neural networks for acoustic modeling in speech recognition. IEEE Signal Processing Magazine, 29(6):82–97, 2012.

Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. A simple language model for task-oriented dialogue. CoRR, 2020.

Harold Hotelling. Analysis of a complex of statistical variables into principal components. Journal of educational psychology, 24(6):417, 1933.

Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry Heck. Learning deep structured semantic models for web search using clickthrough data. ACM International Conference on Information and Knowledge Management (CIKM), 2013.

Xuedong Huang and Li Deng. An overview of modern speech recognition. In Handbook of Natural Language Processing, Second Edition, Chapter 15, 2010.

Anitha Kannan, Richard Chen, Vignesh Venkataraman, Geoffrey J. Tso, and Xavier Amatriain. COVID-19 in differential diagnosis of online symptom assessments. CoRR, abs/2008.03323, 2020.

Hao-Cheng Kao, Kai-Fu Tang, and Edward Y. Chang. Context-aware symptom checking for disease diagnosis using hierarchical reinforcement learning. In AAAI, 2018.

Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. CTRL: A conditional transformer language model for controllable generation. CoRR, 2019.

Sopan Khosla, Shikhar Vashisht, Jill Fain Lehman, and Carolyn Penstein Rosé. Medfilter: Improving extraction of task-relevant utterances through integration of discourse structure and ontological knowledge. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020, pages 7781–7797. Association for Computational Linguistics, 2020.

Jiwei Li, Michel Galley, Chris Brockett, Georgios P. Spithourakis, Jianfeng Gao, and Bill Dolan. A persona based neural conversational model. ACL, 2016.

Shuai Lin, Pan Zhou, Xiaodan Liang, Jianheng Tang, Ruihui Zhao, Ziliang Chen, and Liang Lin. Graph-evolving meta-learning for low-resource medical dialogue generation. CoRR, 2020.

Zachary C. Lipton, Xiujun Li, Jianfeng Gao, Lihong Li, Faisal Ahmed, and Li Deng. Efficient dialogue policy learning with bbq-networks. arXiv.org, 2016.

Guanxiong Liu, Tzu-Ming Harry Hsu, Matthew B. A. McDermott, Willie Boag, Wei-Hung Weng, Peter Szolovits, and Marzyeh Ghassemi. Clinically accurate chest x-ray report generation. CoRR, 2019.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. What makes good in-context examples for gpt-3?, 2021.

G. Mesnil, Y. Dauphin, K. Yao, Y. Bengio, L. Deng, D. Hakkani-Tür, X. He, L. Heck, G. Tur, and D. Yu. Using recurrent neural networks for slot filling in spoken language understanding. IEEE Transactions on Audio, Speech, and Language Processing, 23(3):530–539, 2015.

Randolph A. Miller and Fred E Masarie Jr. Quick medical reference (qmr): An evolving, microcomputer-based diagnostic decision-support program for general internal medicine. Proceedings of the Annual Symposium on Computer Application in Medical Care, pages 947–948, Nov 1989.

Ashwin Paranjape, Abigail See, Kathleen Kenealy, Haojun Li, Amelia Hardy, Peng Qi, Kaushik Ram Sadagopan, Nguyen Minh Phu, Dilara Soylu, and Christopher D. Manning. Neural generation meets real people: Towards emotionally engaging mixed-initiative conversations. CoRR, abs/2008.12348, 2020.
R. C. Parker and R. A. Miller. Creation of realistic appearing simulated patient cases using the INTERNIST-1/QMR knowledge base and interrelationship properties of manifestations. *Methods Inf Med*, 28(4):346–351, Nov 1989.

Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2(11):559–572, 1901.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. Towards empathetic open-domain conversation models: A new benchmark and dataset. *arXiv preprint arXiv:1811.00207*, 2018.

Murali Ravuri, Anitha Kannan, Geoffrey J. Tso, and Xavier Amatriain. Learning from the experts: From expert systems to machine learned diagnosis models. *Machine Learning for Health Care*, 2018.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

Robin Rojowiec, Benjamin Roth, and Maximilian C Fink. Intent recognition in doctor-patient interviews. In *LREC*, 2020.

Alexander Rudnicky and Wei Xu. An agenda-based dialog management architecture for spoken language systems. In *IEEE Automatic Speech Recognition and Understanding Workshop*, volume 13, page 17, 1999.

Sai P. Selvaraj and Sandeep Konam. Medication regimen extraction from clinical conversations. *CoRR*, abs/1912.04961, 2019.

A. Simpson and N. M Eraser. Black box and glass box evaluation of the sundial system. *Third European Conference on Speech Communication and Technology*, 1993.

Hagen Soltau, Mingqiu Wang, Izhak Shafran, and Laurent El Shafey. Understanding medical conversations: Rich transcription, confidence scores & information extraction. *CoRR*, 2021.

G. Tur and L. Deng. *Intent Determination and Spoken Utterance Classification, Chapter 4 in Book: Spoken Language Understanding*. John Wiley and Sons, New York, NY, 2011.

Gokhan Tur, Li Deng, Dilek Hakkani-Tür, and Xiaodong He. Towards deeper understanding: Deep convex networks for semantic utterance classification. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5045–5048. IEEE, 2012.

Gokhan Tur, Asli Celikyilmaz, Xiaodong He, Dilek Hakkani-Tür, and Li Deng. Deep learning in conversational language understanding. In *Chapter 2 of Book: Deep Learning in Natural Language Processing*, 2018.

Oriol Vinyals and Quoc Le. A neural conversational model. *arXiv preprint arXiv:1506.05869*, 2015.
Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*, 2019.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. Dialogpt: Large-scale generative pre-training for conversational response generation. In *ACL, system demonstration*, 2020.
Appendix A. Simulated Clinical Case

{
    "id": 0,
    "age": [
        "young adult (18 to 40 yrs)"
    ],
    "gender": [
        "male"
    ],
    "RFE": [
        "abdominal fullness sensation+"
    ],
    "findings": [
        "diarrhea, chronic+",
        "abdominal pain, recurrent attacks+",
        "lactose intolerance-",
        "gluten intolerance-",
        "marijuana use+",
        "relieved by hot shower+",
        "worse in the morning-",
        "vomiting, recurrent+",
        "intermittent++",
        "anxiety+",
        "sweating increase+",
        "thirst increase+",
        "weight loss+",
        "epigastric abdominal pain-",
        "chronic (> 4 weeks)+"
    ]
}

Table A.1: A clinical case simulated from KB.

Appendix B. Question Paraphrasing using GPT-3

| Finding  | GPT-3 generated questions |
|----------|----------------------------|
| anxiety  | Are you anxious?           |
|          | Do you have anxiety?       |
|          | Have you been experiencing any anxiety? |
|          | Are you feeling nervous or anxious? |
| back pain| Do you feel pain in your back? |
|          | Is your back hurting?       |
|          | Does your back hurt?        |
|          | Are you experiencing pain in your back? |

Table B.1: Examples of question paraphrases generated by GPT-3.

Algorithm 1 Emotional Addition Extraction

Input: Default question $Q_d$, Edited question $Q_e$

1: Initialize $splits = \text{split } Q_e \text{ on punctuation}$
2: Initialize $scores = \text{empty array}$
3: for $s \in splits$ do
4:     $scores[s] \leftarrow FuzzyMatchScore(s, Q_d)$
5: end for
6: $i_q \leftarrow \text{argmax}(scores)$
7: return $splits[:i_q]$ (Return everything preceding the (most likely) question)

Appendix C. Details of Emote dataset construction

We generated the emote dataset using an in-house Doctors Edits dataset, which contains 3600 instances of medical professional-edited questions and their preceding medical professional and patient messages. Medical professional-edited questions are templated questions from KB and then were subsequently edited by the professionals based on the context of the conversation.

These edits are typically done to impart additional emotion to the text (emote phrase), although some of the edits are made for more pragmatic reasons (e.g. improve question readability). To extract the emote phrase, we use a simple heuristic method based on the assumption that the edited question is of the form: [emote phrase] KB question [additional information], e.g. "Oh I’m sorry to hear that. Do you have flushing? That is, do your arms feel warmer than usual?. We simply split the edited question on punctuation and identify which part of the split question is closest to the KB question by fuzzy string matching\(^2\); once the most similar question section is identified, the emote phrase is returned as the preceding sub-string to this section. (Algorithm 1 in Appendix for details). The accuracy of this simple algorithm for our task was evaluated manually and shown to achieve 99.4% accuracy within our limited domain of conversation. Table C.1 presents example emote language phrases corresponding to the emote control codes.

The dataset has a class imbalance towards None, indicating it is often not necessary to emote, and when one is emoting, Affirmative is the most common.

\(^2\) https://github.com/seatgeek/fuzzywuzzy
Control code | Emote language
--- | ---
affirmative | Thanks for the input
Okay
I see
Got it
empathy | Sorry about that
That’s concerning
Okay, I’m sorry to hear
apology | I am sorry for asking
I apologise if this is personal
I am sorry for asking if it sounds personal but may I know

Table C.1: Examples of emote codes and example sentences that are mined from real world medical conversations. See text for details

Appendix D. Emotion Classifier Performance

|                | Precision | Recall | F1  | Support |
|----------------|-----------|--------|-----|---------|
| None           | 0.90      | 0.99   | 0.94| 557     |
| Affirmative    | 0.93      | 0.56   | 0.70| 127     |
| Empathy        | 0.72      | 0.72   | 0.72| 18      |
| Apology        | 0.83      | 0.83   | 0.83| 6       |
| Accuracy       | -         | -      | 0.90| 708     |
| Macro Avg      | 0.85      | 0.78   | 0.80| 708     |
| Weighted Avg   | 0.90      | 0.90   | 0.89| 708     |

Table D.1: Classification scores for Emotion Classifier on human-augmented Emote dataset.

Appendix E. Dissecting Emotion Classifier: Emotion Attribution

As we mentioned in §3.1, we used logistic regression as the final classifier for the emote classifier, where:

$$\text{logits}(p_i) = \sum_j M_{ij}^T x_j + b_i,$$  \hspace{1cm} (1)

where \(p_i\) is the probability of the \(i^{th}\) control code and index \(j\) represents the input source (previous question, previous patient response, target finding), \(M_{ij}\) is the learned coefficient vector corresponding to \(i^{th}\) class and \(j^{th}\) source, \(x_j\) is SBERT embedding vector corresponding to \(j^{th}\) input source, after dimensionality reduced by PCA, and \(b\) are learned biases.

We analyzed how conversational context affects predicted emotion. Using eq. 1, we compute the contribution of each input source \(j\) for each output control code \(i\) by looking at the individual summands \(M_{ij}^T x_j\). The biases for each of the four classes are the following: None = 3.27, Affirmative = 0.88, Empathy = -1.66, Apology = -2.49.

Figure E.1: Contribution of input source to emote classification.

Figure E.1 shows the mean contribution of each input source to the predicted output control code \((i^* = \arg \max_i p)\). We find that the apology and empathy are strongly predetermined by a single input source. For the apology class main contributor is the previous question, this is consistent with expectation (i.e. we apologize for asking personal and embarrassing questions). For the empathy class the main contributor is the previous response, which again matches intuition that we show empathy if the patient response involves significant negative sentiment. For affirmative class, all three types of inputs are taken into account, with previous question has the largest contribution. Intuitively, this makes sense that we may prioritize responding to the previous answer based on not only what the exchange has been so far, but also what we are going to ask next. For the none class, none of the input sources contribute significantly, this is consistent with the none class being driven by the bias (3.27!) – when the model doesn’t predict an emote it is because none of the input sources contributed to an alternative class.
Appendix F. Generalization of MEDCOD’s NLG Component

One of the key benefits provided by a neural language generator like DialoGPT over a simple lookup-based system is its ability to generalize to new unseen findings; the broad world knowledge learned through pretraining gives the model an ability to generate appropriate questions for new findings. Fig. F provides qualitative results on our NLG model for various ‘types’ of input findings.

| Category                     | Finding                  | Generation                                      |
|------------------------------|--------------------------|-------------------------------------------------|
| Unseen semantically          | knuckle pain             | Do you have any pain in your knuckles?          |
| meaningfully findings        | spider bite              | Have you been bitten by a spider recently?      |
|                              | fell off motorbike       | Have you recently fallen off a motorbike?       |
|                              | broken arm               | Is your arm broken?                             |
| Unseen fake finding          | dysparatosis             | Do you suffer from dysparatosis?                |
|                              | dysparatosis (temp=0.9)  | Have you been diagnosed with dysparatosis?      |
| Misc. findings               | eaten too much cheese    | Have you recently eaten too much cheese?         |
| showcasing                   | lost my phone            | Have you lost your phone?                       |
| general world &              | slept through my alarm   | Did you miss your alarm?                        |
| grammar knowledge            |                          |                                                 |
|                              | (temp = 0.9)             | Have you been having trouble sleeping?          |

Table F.1: Examples of MEDCOD generating presence questions for unseen findings (temperature=0.6 unless specified). The model is able to generate a question medically consistent with the input finding, despite never seeing it during training, showing the NLG model’s ability to generalize. When generating for a finding which it can’t get semantic information from (e.g., a fake disease name), it defaults back to generic "Do you have [disease]" style questions, unless we use a higher temperature in which case we get unexpected results. The model can also generate questions for non-medical findings (e.g., *lost my phone*); in these cases it again can generate a semantically consistent question with the "finding", although generations lose accuracy when temperature is raised.
Appendix G. Qualitative comparison of generated questions by three MEDCOD Instantiations

| Model          | Finding              | Emote     | Generated Responses (multiple invocations)                                      |
|----------------|----------------------|-----------|----------------------------------------------------------------------------------|
| **Expert**     | recurrent headache   | affirmative | *Do you have headaches that come and go often?*                                 |
|                |                      |           | *Do you have headaches that come and go often?*                                 |
|                | generalized weakness | empathy   | *Does your entire body feel weak?*                                              |
|                |                      |           | *Does your entire body feel weak?*                                              |
| **MEDCOD-no-Emote** | recurrent headache | affirmative | *Do you have headaches that occur on a regular basis?*                        |
|                |                      |           | *Are your headaches recurring?*                                                  |
|                | generalized weakness | empathy   | *Are you feeling weak in your entire body?*                                     |
|                |                      |           | *Is your entire body weak?*                                                      |
| **MEDCOD**     | recurrent headache   | affirmative | *Thanks for the input. Does your headache keep coming back?*                    |
|                |                      |           | *Okay. Is your headache recurrent?*                                              |
|                | generalized weakness | empathy   | *Sorry to know that. Are you weak all over?*                                    |
|                |                      |           | *That’s worrisome. Do you experience weakness all over your body?*               |

Table G.1: Qualitative comparison of generated questions by the three NLG models. Note the lack of diversity in the **Expert** model and lack of emotion in the **MEDCOD-no-Emote** model. The questions generated by **MEDCOD** show the greatest combination of emotion and variation. For brevity, not shown is the previous turn of the conversation that is also used as input.
Appendix H. Full Conversation Comparison

| MEDCOD                  | Expert                      |
|-------------------------|-----------------------------|
| **Complaint: headache** |                             |
| **Curai**               | **Curai**                   |
| Are you experiencing a | Do you have a severe        |
| severe headache?        | headache?                   |
| **Patient**             | **Patient**                 |
| Yes, and it’s very     | Yes, and it’s very          |
| painful!                | painful!                    |
| **Curai**               | **Curai**                   |
| I am sorry to hear that.| Do you have a headache that |                             |
| Is your headache        | is                           |
| mostly concentrated     | concentrated around your    |
| around your forehead?   | forehead?                   |
| **Patient**             | **Patient**                 |
| Yes it is               | Yes it is                   |
| **Curai**               | **Curai**                   |
| I hear you. Do you     | Do you have headaches that  |
| suffer from headaches   | come and go?                |
| that come and go?       | **Patient**                 |
| **Patient**             | No, not usually             |
| **Curai**               | **Curai**                   |
| Is your headache mainly | Is your headache mainly in  |
| located in your         | your temple area on the     |
| temple area on the side | side of your head behind   |
| of your head behind     | your eyes?                  |
| your eyes?              | **Patient**                 |
| Yes most of the time    | Yes most of the time        |
| **Curai**               | **Curai**                   |
| Okay, I’m sorry to     | Do you have a headache that |
| hear. Does your headache| is only occur on one side   |
| only occur on one side  | of your head?               |
| of your head?           | **Patient**                 |
| **Patient**             | No it’s on both sides       |
| **Curai**               | **Curai**                   |
| Does your headache      | Would you describe your     |
| pulsate or throb?       | headache as throbbing?      |
| **Patient**             | **Patient**                 |
| Yes, often              | Yes, often                  |
| **Curai Current Question**| **Curai Current Question**  |
| Does light seem to      | Does light hurt your eyes?  |
| hurt your eyes?         |                             |

Figure H.1: MEDCOD and Expert system used for mock patient case. Note the more natural conversation flow with MEDCOD due to its emotive responses and variation in question framing.
Appendix I. GPT-3 Invocations

Following are some prompts given to GPT-3 to generate new paraphrases for the **progressive paralysis** finding. Bold is the generated response by GPT-3. These are used to generate the following five distinct paraphrases:

- Are you noticing a progressive weakness on one side of your body?
- Have you noticed a gradual weakening of one side of your body?
- Have you noticed a weakness on one side of your body that seems to be getting worse?
- Has your weakness on one side of your body been getting progressively worse?
- Is one side of your body becoming increasingly weaker than the other?
MEDCOD: A Medically-Accurate, Emotive, Diverse, and Controllable Dialog System

Rephrase the question asking if the patient has the given symptom

Symptom: epigastric pain [PRESENT]. Question: Do you have pain in your upper middle abdomen, just below your breast bone? => Do you have pain just beneath your breast bone in the middle of your upper abdomen?

Symptom: anosmia [PRESENT]. Question: Are you experiencing a decreased sense of smell? => Is your sense of smell impaired?

Symptom: crushing chest pain [PRESENT]. Question: Do have chest pain that feels like someone is crushing your chest? => Are you experiencing a crushing sensation around your chest with your chest pain

Symptom: parkinsonism [PRESENT]. Question: Have you ever been diagnosed with parkinsonism? => Do you have a history of parkinsonism?

Symptom: lower extremity pain [PRESENT]. Question: Does your leg hurt? => Are you experiencing pain in your leg?

Symptom: nocturia [PRESENT]. Question: Do you have to urinate frequently even at night, waking up two or more times to urinate? => During the night, do you wake up several times to urinate?

Symptom: chronic productive cough [PRESENT]. Question: Have you had a cough lasting for more than 8 weeks that is wet or brings up phlegm? => Has your cough been wet or bringing up phlegm for more than eight weeks?

Symptom: exertional chest pain [PRESENT]. Question: Do you experience chest pain that is worse with exertion, such as when walking or doing other physical activity? => Do you notice your chest pain gets worse with physical activity such as running?

Symptom: substernal chest pain [PRESENT]. Question: Do you have any pain in the center of your chest behind your sternum [breastbone]? => Is there any pain behind the sternum [breastbone] in the center of your chest?

Symptom: atherosclerosis [PRESENT]. Question: Have you been diagnosed with atherosclerosis? => Do you suffer from atherosclerosis?

Symptom: progressive paralysis [PRESENT]. Question: Have you noticed increasing weakness on one side of your body? => Are you noticing a progressive weakness on one side of your body?

(a) First invocation

Rephrase the question asking if the patient has the given symptom

Symptom: nocturia [PRESENT]. Question: Do you have to urinate frequently even at night, waking up two or more times to urinate? => During the night, do you wake up several times to urinate?

Symptom: hemiplegia [PRESENT]. Question: Do you have any weakness on one side of your body? => Does one side of your body seem to be weaker than the other?

Symptom: pain relieved with food [PRESENT]. Question: Do you have pain that is relieved by food? => Does eating food ease the pain?

Symptom: anosmia [PRESENT]. Question: Are you experiencing a decreased sense of smell? => Is your sense of smell impaired?

Symptom: recurrent abdominal pain [PRESENT]. Question: Have you had the repeated episodes of your abdominal pain over the last 3 months? => Has your abdominal pain occurred multiple times over the last 3 months?

Symptom: altered mental status [PRESENT]. Question: Do you have any impaired consciousness? => Do you feel that your mental state is impaired?

Symptom: parkinsonism [PRESENT]. Question: Have you ever been diagnosed with parkinsonism? => Do you have a history of parkinsonism?

Symptom: allergen exposure [PRESENT]. Question: Have you been around anything you are allergic to? => Have you recently come into contact with anything you are allergic to?

Symptom: muscle weakness [PRESENT]. Question: Do you feel like your muscles are abnormally weak? => Are you experiencing abnormally weak muscles?

Symptom: aphasia [PRESENT]. Question: Do you have difficulty speaking or understanding language? => Are you finding it hard to speak or understand language?

Symptom: progressive paralysis [PRESENT]. Question: Have you noticed increasing weakness on one side of your body? => Have you noticed a gradual weakening of one side of your body?

(b) Second invocation

Figure I.1: Individual GPT-3 Prompts for the same finding, showcasing the prompt engineering and diversity of output.