Learning to Ask Like a Physician

Eric Lehman[1,2,*], Vladislav Lialin[3], Katelyn Y. Legaspi[5], Anne Janelle R. Sy[4], Patricia Therese S. Pile[4], Nicole Rose I. Alberto[5], Richard Raymund R. Ragasa[5], Corinna Victoria M. Puyat[5], Isabelle Rose I. Alberto[5], Pia Gabrielle I. Alfonso[5], Marianne Taliño[6], Dana Moukheiber[1], Byron C. Wallace[7], Anna Rumshisky[3], Jennifer J. Liang[2,8], Preethi Raghavan[1,9], Leo Anthony Celi[1, 10], Peter Szolovits[1,2]

[1]MIT, [2]MIT-IBM Watson AI Lab, [3]University of Massachusetts Lowell, [4]UERM Memorial Medical Center [5]University of the Philippines, [6]ASMPH, [7]Northeastern University, [8]IBM Research, [9]Fidelity Investments, [10]Beth Israel Deaconess Medical Center

Abstract

Existing question answering (QA) datasets derived from electronic health records (EHR) are artificially generated and consequently fail to capture realistic physician information needs. We present **Di**scharge **S**ummary **C**linical **Q**uestions (DiSCQ), a newly curated question dataset composed of 2,000+ questions paired with the snippets of text (*triggers*) that prompted each question. The questions are generated by medical experts from 100+ MIMIC-III discharge summaries. We analyze this dataset to characterize the types of information sought by medical experts. We also train baseline models for trigger detection and question generation (QG), paired with unsupervised answer retrieval over EHRs. Our baseline model is able to generate high quality questions in over 62% of cases when prompted with human selected triggers. We release this dataset (and all code to reproduce baseline model results) to facilitate further research into realistic clinical QA and QG. [1]

1 Introduction

Physicians often query electronic health records (EHR) to make fully informed decisions about patient care (Demner-Fushman et al., 2009). However, D'Alessandro et al. (2004) found that it takes an average of 8.3 minutes to answer a single question, even when physicians are trained to retrieve information from an EHR platform. Natural language technologies such as automatic question answering (QA) may partially address this problem.

There have been several dataset collection efforts that aim to facilitate the training and evaluation of clinical QA models (Pampari et al., 2018; Yue et al., 2021; Raghavan et al., 2021; Kell et al., 2021). However, template-based (Pampari et al., 2018; Raghavan et al., 2021) and other kinds of automated generation (Yue et al., 2021) methods

His past medical history is significant for prostate cancer, benign prostatic hypertrophy, hypothyroidism, status post radiation for non Hodgkin's lymphoma, chronic painless hematuria, degenerative joint disease and history of a murmur.

(1) prostate cancer, benign prostatic hypertrophy
Date of diagnosis? Any interventions done (RT, surgery)?

(2) hypothyroidism
Maintenance medications?

Figure 1: Example of an annotated discharge summary section. The highlighted portion shows the "trigger" for the questions.

are by nature brittle and have limited evidence of producing questions that medical professionals ask.

Datasets such as emrQA (Pampari et al., 2018) and emrKBQA (Raghavan et al., 2021) attempt to simulate physician queries by defining templates derived from actual questions posed by physicians and then performing slot-filling with clinical entities. This method yields questions that are structurally realistic, but not consistently medically relevant. Yue et al. (2020) found that sampling just 5% of the emrQA questions was sufficient for training a model. They further note that 96% of the questions in a subsection of emrQA contain key phrases that overlap with those in the selected answer.

In follow-up work, Yue et al. (2021) provide a new dataset of 975 questions generated using a diverse question generation model with a human-in-the-loop and 312 questions generated by medical experts from scratch, with the caveat that they must be answerable on the given discharge summary. However, a random sample of 100 questions from the former reveals that 96% of the 975 questions were slot-filled templates directly from emrQA. A

* lehmer16@mit.edu

[1] https://github.com/elehman16/discq

Proceedings of the 4th Clinical Natural Language Processing Workshop, pages 74 - 86

separate random sample of 100 questions from the latter set reveals that 54% of the questions also use the same slot-filled templates from emrQA. Similarly, we find that 85% of the machine-generated questions and 75% of the human-generated questions contain the exact same key phrases as in the selected answer. Although Yue et al. (2020) does not discuss how they prompt physician questions, our analysis strongly suggests that even in the case of questions "written" by physicians, answer spans are likely identified in advance; this significantly constrains the set of questions a medical professional can ask.

To address this paucity of natural, clinically relevant questions, we collect queries that might plausibly be asked by healthcare providers during patient handoff (i.e., transitions of care). We use patient discharge summaries from the Medical Information Mart for Intensive Care III (MIMIC-III) English dataset (Johnson et al., 2016) to mimic the handoff process. We expect this process to produce more natural questions than prior work. We work with 10 medical experts of varying skill levels. We ask them to review a given discharge summary as the receiving physician in a patient handoff and record any questions they have as well as the piece of text within the discharge summary (trigger) that prompted the question. A sample of questions and corresponding triggers can be seen in Figure 1.

We train question trigger detection and question generation (QG) models on DiSCQ, paired with unsupervised answer retrieval over the EHR. Finally, we propose a new set of guidelines for human evaluation of clinical questions and evaluate the performance of our pipeline using these guidelines. Concretely, our contributions are summarized as follows:

- We work with 10 medical experts to compile DiSCQ, a new dataset of 2000+ questions and 1000+ triggers from over 100+ discharge summaries, providing an important new resource for research in clinical NLP.
- We demonstrate the dataset's utility by training baseline models for trigger detection and question generation.
- We develop novel guidelines for human evaluation of clinical questions. Our experiments show that widely used automated QG metrics do not correlate with human-evaluated question quality.

2 Related Work

2.1 Clinical Question Datasets

Clinical information retrieval, and in particular clinical question answering, is a challenging research task with direct potential applications in clinical practice. Several dataset collection efforts gather consumer health questions and pair them with answers from sources like WebMD and PubMED (Yu et al., 2007; Cao et al., 2011; Abacha and Zweigenbaum, 2015; Abacha et al., 2017; Zahid et al., 2018; Demner-Fushman et al., 2020; Savery et al., 2020; Zhu et al., 2020; Abacha et al., 2019). Likewise, Suster and Daelemans (2018) automatically generate 100,000+ information retrieval queries from over 11,000+ BMJ Case Reports. While these resources are helpful in testing a model's understanding and information retrieval ability on biomedical texts, these datasets consist of broad medical questions asked by the general population. Doctors will not only ask more specific and targeted questions, but also query the EHR to make fully informed decisions about patient care.

The number of publicly available QA datasets derived from EHR systems is quite limited due to the labor intensiveness and high skill requirement needed to create such a dataset. As mentioned previously, to help alleviate this dearth of clinical questions, Pampari et al. (2018) introduced emrQA, a QA dataset constructed from templatized physician queries slot-filled with n2c2 annotations.[2] Fan (2019) extended emrQA by explicitly focusing on "why" questions. Soni et al. (2019) introduced a novel approach for constructing clinical questions that can be slot-filled into logical-forms. Yue et al. (2021) applied an emrQA-trained question generation model paired with a human-in-the-loop to collect 1287 questions conditioned on and answerable from the given context.

In contrast, in our data collection process we do not restrict the medical expert to ask only questions answerable from a particular part of the discharge summary. This leads to more diverse and natural questions. Additionally, in DiSCQ each question is associated with a span of text that triggered the question.

2.2 Question Generation

Question Generation (QG) is a challenging task that requires a combination of reading comprehen-

[2] https://www.i2b2.org/NLP/DataSets/

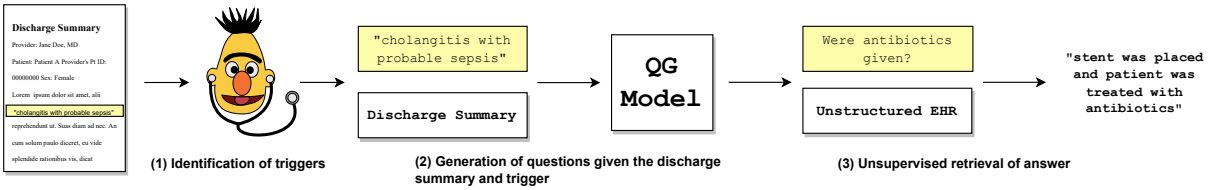

Figure 2: Schematic of the pipeline process used to generate and answer questions.

sion and text generation. Successful QG models may aid in education (Heilman and Smith, 2010; Du et al., 2017), creating dialogue systems or chatbots (Shang et al., 2015; Mostafazadeh et al., 2016; Shum et al., 2018), building datasets (Duan et al., 2017) or improving question answering models through data augmentation (Tang et al., 2017; Dong et al., 2019; Puri et al., 2020; Yue et al., 2021).

Most QG approaches can be broken down into either rule-based or neural methods. Rule-based approaches often involve slot filling templatized questions (Heilman and Smith, 2010; Mazidi and Nielsen, 2014; Labutov et al., 2015; Chali and Hasan, 2015; Pampari et al., 2018). While often effective at generating numerous questions, these methods are very rigid, as virtually any domain change requires a new set of rules. This problem is particularly important in medical QG, as different types of practices may focus on varying aspects of a patient and therefore ask different questions.

Compared to rule-based methods, sequence-to-sequence models (Serban et al., 2016; Du et al., 2017) and more recently transformer-based models (Dong et al., 2019; Qi et al., 2020; Lelkes et al., 2021; Murakhovs'ka et al., 2021; Luo et al., 2021) allow for generation of more diverse questions and can potentially mitigate the problem of domain generalization via large-scale pre-training (Brown et al., 2020) or domain adaptation techniques. We choose to train both BART (Lewis et al., 2020) and T0 (Sanh et al., 2021) models for the task of question generation due to their high performance and ability to generalize to new tasks.

3 DiSCQ Dataset

We work with 10 medical experts of varying skill levels, ranging from senior medical students to practicing MDs, to construct a dataset of 2029 questions over 100+ discharge summaries from MIMIC-III (Johnson et al., 2016).

3.1 Dataset Collection

The goal of our question collection is to gather questions that may be asked by healthcare providers during patient handoff (i.e., transitions of care). We use the patient discharge summary to simulate the handoff process,[3] where the discharge summary is the communication from the previous physician regarding the patient's care, treatment and current status. Annotators are asked to review the discharge summary as the receiving physician and ask any questions they may have as the physician taking over the care of this patient.

Annotators are instructed to read the discharge summary line-by-line and record (1) any questions that may be important with respect to the patient's future care, and, (2) the text within the note that triggered the question. This may mean that questions asked early on may be answered later in the discharge summary. Annotators are permitted to go back and ask questions if they feel the need to do so. To capture the annotators' natural thought processes, we purposely provide only minimal guidance to annotators on how to select a trigger or what type of questions to ask. We only ask that annotators use the minimum span of text when specifying a trigger.[4]

We also encourage all questions to be asked in whatever format they feel appropriate. This leads to many informal queries, in which questions are incomplete or grammatically incorrect (Figure 1). Further, we encourage all types of questions to be asked, regardless of whether they could be answered based on the EHR. We also allow the annotators to ask an arbitrary number of questions. This allows for annotators to skip discharge summaries entirely should they not have any questions.

3.2 Dataset Statistics

The trigger/question pairs are generated over entire discharge summaries. We instruct annotators

[3] We discard any records pertaining to neonatal or deceased patients.

[4] Instructions given to annotators will be available here.

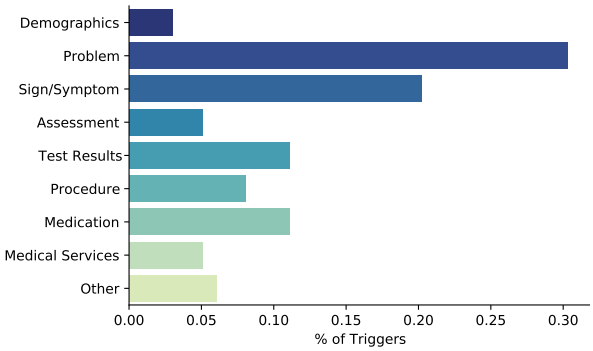

Figure 3: We randomly sample 100 gold triggers and have one of the authors, a physician, categorize the type of information that the trigger contains.

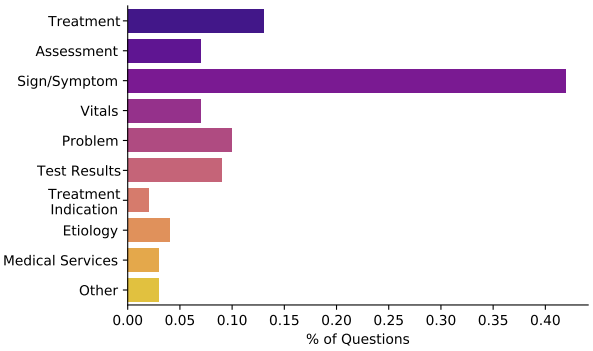

Figure 4: We randomly sample 100 questions and have one of the authors, a physician, categorize what type of information the question is asking for.

to select the minimum span that they used as the trigger to their question; this leads to triggers of length 5.0 ± 14.1 tokens. We additionally find that there are 1.86 ± 1.56 questions per trigger. As mentioned previously, we encourage our medical experts to ask questions however they feel most comfortable. This led to a wide variety in how questions were asked, with some entirely self-contained (46%), others requiring the trigger for understanding (46%), and some requiring the entire sentence containing the trigger to comprehend (8%).[5] We also observe that 59% of the bi-grams in our questions are unique (i.e., over half of all bi-grams that appear in one question are not seen in any other question), demonstrating the diversity of how our questions are asked (Table 1).

We additionally examine where in the discharge summary annotators tend to select triggers from. We find that a majority of triggers are selected from the `Hospital Course` (13%) and `History of Present Illness` (39%) sections. This is unsurprising, as these are the narrative sections of the note where the patient's history prior to admission and their medical care during hospitalization are described. Further, we find that a majority of triggers selected are either a `Problem` or `Sign/Symptom` (Figure 3). This aligns with our intuition, as clinicians are often trained to organize patient information from a problem-oriented perspective. Moreover, developing a differential diagnosis usually begins with gathering details of the patient's clinical presentation.

In Figure 4, we examine the types of information needs exhibited by our questions. We find that 83% and 80% of the questions categorized as `Sign/Symptom` and `Problem`, respectively, stem from the same category of trigger. `Sign/Symptom` questions generated from `Sign/Symptom` triggers are usually asking about associated symptoms (e.g., Trigger: *dysuria*; Question: *Any perineal rash or irritation?*) or additional details about the trigger (e.g., onset, timing). Similarly, `Problem` questions generated from `Problem` triggers are usually asking about associated comorbid conditions or additional details of a diagnosis (e.g., date of diagnosis, severity). We interestingly find that 62% of the `Treatment` questions and 56% of the `Test Results` questions are derived from triggers of type `Problem`. This can be attributed to diagnostic tests being used to monitor disease progression and treatment questions asking about how a problem is managed.

As a soundness check, we randomly sample 100 questions from our dataset and find that only 22% of them directly map to emrQA templates. Of the 22 that match, 17 of them map directly to `|problem|?` and `|test|?`. Additionally, we sample 100 questions to determine where a physician would hypothetically search the EHR should

Characteristics	emrQA	CliniQG4QA	DiSCQ
Total Articles	2,425	36	114
Total Questions	455,837	1287	2029
Questions / Article	187	35.8	17.8
Article Length	3828	2644	1481
Question Length	7.8	8.7	4.4
Unique Question Bi-grams	-	24%	59%
Physician Generated	0%	24%	100%
Indicates Question Motivation	No	No	Yes

Table 1: Comparison of emrQA, CliniQG4QA and our dataset. Question and article length scale given in tokens. Unique question bi-grams is given as a ratio.

[5] Based on a sample of 100 questions.

they choose to find the answers to these questions.[6] We find that one of the authors, a physician, would search external resources 3% of the time, the structured data 20% of the time and both the structured and unstructured data 21% of the time. The remaining 56% of questions would be answered solely from unstructured EHR data. This differs significantly from both emrQA and CliniQG4QA, in which all questions are answerable using unstructured EHR data.

As mentioned previously, we provide only minimal guidance on how to select a trigger or what type of question to ask, in order to capture the annotators' natural thought processes. The task is purposely presented in an open-ended fashion to encourage natural questions. This may lead to situations in which two annotators examining the same discharge summary focus on entirely different aspects of the patient. Such a scenario is likely to be common, as if most experts agree that a piece of information is important, then it would likely already be in the discharge summary. We can attempt to measure this variation between medical experts by calculating trigger level agreement in documents annotated by two different annotators (roughly 50% of discharge summaries in DiSCQ). We find a Cohen Kappa of 0.08.[7]

This lower agreement can be expected, as different spans can express the same information due to information redundancy in clinical notes. Furthermore, clinical reasoning is not a linear process; therefore, different triggers can lead to the same question. For example, an expression of elevated blood pressure (*"blood pressure of 148 to 162/45 to 54"*) and a diagnosis of hypertension (*"Hypertension"*) led two annotators to both ask about the patient's normal blood pressure range. We do not measure agreement of questions asked, as this is an inherently subjective task and questions are asked *because* of differences between medical experts.

4 Task Setup

We consider the task of generating questions that are relevant to a patient's care, given a discharge summary and a trigger. Afterwards, we attempt to find answers to these generated questions (Figure 2). We also examine model performance for when the trigger is not provided and must instead be predicted. The task of generating questions without triggers can be viewed similarly to answer-agnostic question generation. We take a similar approach to (Subramanian et al., 2018), in which we implement a pipeline system that first selects key phrases from the passage and then generates questions about the selected key phrases.

While a majority of past works attempt to ensure that the generated question is answerable (Nema et al., 2019; Pan et al., 2020; Wang et al., 2020a; Huang et al., 2021), we do not impose this constraint. In fact, we argue that the ability to generate unanswerable questions is necessary for real-world applications, as a question answering system should be able to identify such questions. These questions can be used as hard-negatives to train and calibrate QA systems.

5 Models

Pre-trained transformers have become ubiquitous in many natural language processing tasks (Devlin et al., 2019; Raffel et al., 2020; Sanh et al., 2021), including natural language generation (Lewis et al., 2020; Bao et al., 2020). Additionally, large-scale transformers have demonstrated the importance of parameter count for both upstream (Kaplan et al., 2020) and downstream tasks, especially in low-resource settings (Brown et al., 2020; Sanh et al., 2021). As these results were mainly shown in non-clinical general domains, we find it important to evaluate both medium-sized and large models.

We formulate trigger detection as a tagging problem, for which we fine-tune ClinicalBERT (Alsentzer et al., 2019). For question generation, we experiment with both BART (406M parameters) (Lewis et al., 2020) and T0 (11B parameters) (Sanh et al., 2021). Question generation is formulated as a conditional generation problem and modelled via a sequence-to-sequence approach. During evaluation, we use greedy sampling to produce generated text.

Reducing context size Due to memory constraints and the limited sequence length of pre-trained models, we only select the part of the discharge summary containing the trigger. This is done in two possible ways: (1) extracting the sentence[8] with the trigger or multiple sentences if a trigger spans across sentence boundaries or (2) extracting a chunk of size 512 containing the trigger in it. To check if this context is actually used by

[6] We use the same sample of 100 questions as before.

[7] This is calculated on a per-token level.

[8] Sentence splitting is performed using ScispaCy's `en_core_sci_md`.

the models we also fine-tune BART without extra discharge summary context (trigger text only).

Handling multiple questions 41% of the DiSCQ examples have multiple questions per trigger. Sometimes the questions depend on each other:

- *What meds was used? dosage? and route of administration?*
- *Any culture done? What were the findings?*

For this reason, we train and evaluate models in two different setups: split questions (by the `?`-symbol) and combined questions. While the split-questions format might be more comparable to pre-existing work, the combined-questions setting likely models more realistic behavior of medical professionals.

Prompting Schick and Schütze (2021) demonstrate that adding natural language instructions to the model input can significantly improve model quality. The area of prompting has recently gained widespread popularity (Liu et al., 2021) and has had particular success in low-supervision scenarios (Schick and Schütze, 2021). T0 (Sanh et al., 2021) is a fine-tuned T5 (Raffel et al., 2020) model trained on 64 datasets and prompts from the Public Pool of Prompts (Bach et al., 2022). Given a trigger and some context from the discharge summary, we fine-tune T0++ and BART with the following prompt: "`{context}` After reading the above EMR, what question do you have about "`{trigger}`"? Question:".

6 Results

We split 2029 questions into train (70%), validation (10%) and test (20%) sets[9] and fine-tune the models as described in Section 5. To evaluate trigger detection, we use token-level precision, recall and F1 score. For automated evaluation of question generation we use ROUGE-L (Lin, 2004), METEOR (Banerjee and Lavie, 2005) and BERTScore (Zhang et al., 2020) metrics. To monitor the diversity of generated questions, we measure the fraction of unique questions on the evaluation set. As the question generation task has high variability of plausible generations, the utility of automatic metrics is debatable due to poor correlation with human evaluation (Callison-Burch et al., 2006; Novikova et al., 2017; Elliott and Keller, 2014; Zhang et al., 2020; Bhandari et al., 2020). For this reason, we additionally perform human evaluation (Section 7).

[9]We use a document level split.

6.1 Trigger detection

As mentioned in Section 3, we collect triggers for each question asked. We train a simple ClinicalBERT model to predict whether or not each token-piece is a trigger. To ground these results, we additionally use ScispaCy Large (Neumann et al., 2019) to tag and classify all clinical entities as triggers. Results are shown in Table 2.

Model	Recall	Precision	F1
ScispaCy	0.186	0.033	0.056
ClinicalBERT	0.184	0.196	0.190

Table 2: Trigger detection results on the test set.

We see that our model exhibits poor performance likely due to the fact that there is low agreement between annotators about which spans to highlight when asking questions.

6.2 Question generation

Automated metrics for question generation experiments are available in Table 4. While generation diversity changes significantly between different models, ranging from 30% of unique questions to 79%, METEOR, ROUGE-L and BERTScore show very similar and low performance across the board.

However, upon observation, many of the generated questions seem reasonable (Table 3), suggesting that these metrics might not fit the task. We hypothesize that this is caused by two reasons: (1) the short length of our questions and (2) a high number of potentially reasonable questions that could be generated. As we observe during the data collection process, different annotators seem to ask different questions despite citing the same trigger. For these reasons, human evaluation (Section 7) might be a more appropriate approach for testing the quality of these models.

6.3 Answer Selection

In addition to identifying triggers and generating questions, we attempt to find answers to these questions. We only consider the unstructured portion of the EHR data. We train a ClinicalBERT model on emrQA augmented with unanswerable questions via negative sampling (Liang et al., 2022). Due to the question's frequent dependency on the trigger, given a trigger and a question, we prompt the model with the following text: "With respect to `{trigger}`, `{question}`?". We first query the remainder of the discharge summary that the question was generated from. If we are unable to find

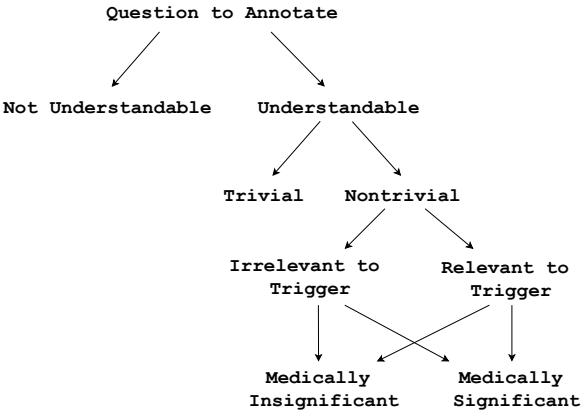

Figure 5: A breakdown of how questions are annotated.

an answer with probability above some threshold[10], we query the model on prior patient notes. We then select the highest probability span and expand it to a sentence level prediction. We always return a prediction even in cases where all sentences are equally unlikely to be the answer.

7 Human Evaluation

Human evaluation is still the most reliable way to compare generative models for diverse tasks like question generation. Common categories for question generation to consider are grammar, difficulty, answerability and fluency (Nema et al., 2019; Tuan et al., 2019; Wang et al., 2020b; Huang et al., 2021). However, not all of these categories are relevant to *clinical* question generation. We evaluate questions generated using our pipeline, as well as gold standard questions on the following four categories (binary scale):

Understandability Can an individual familiar with medical/clinical language understand the information needs expressed, even if the question is not a complete sentence or contains grammar/spelling errors?

Nontriviality Is the question unanswerable with respect to the sentence it was triggered/generated from? A question that would be considered *trivial* would be "Did the patient have a fever?" if the context presented was "The patient had a fever".

Relevancy to trigger Is the trigger or the sentence containing the trigger related to the question?

[10]This threshold was chosen manually by examining question-answer pairs on a validation set.

Clinical meaningfulness Will the answer to this question be helpful for further treatment of this patient or understanding the patient's current condition? Or alternatively, is it reasonable that a medical professional would ask this question given the provided context?

Annotations were divided evenly between medical experts. Each question is scored independently by two different annotators. However, due to time constraints, there are no discussions between annotators about their decisions. We also ensure that annotators did not receive discharge summaries that they had seen previously. Lastly, it is important to note that annotations were assigned blindly. Annotators were informed that they would be scoring both human and machine generated questions, but were not informed about (1) where the question was generated from (i.e., human or machine) and (2) the proportion of human:machine generated questions.

We score questions using the tree presented in Figure 5. In cases in which the question is both understandable and nontrivial, we additionally ask medical experts to determine whether or not the proposed answer fully answers, partially answers or is irrelevant to the question. Results can be seen in Table 5 and Table 6.

8 Discussion

We evaluate performance of both the best BART and T0 model with respect to ROUGE-L score. We select 400 questions generated from each model, half of which are generated with gold triggers and the other half with predicted triggers, as described in Section 6.1. Two medical experts score each question. Due to the subjective nature of the task, we find moderate agreement between annotators with respect to scoring questions ($\kappa = 0.46$) and scoring answer sufficiency ($\kappa = 0.47$). We use the "Satisfies All" column (i.e., satisfies all four human evaluation categories) to calculate agreement between questions.

Results show that the T0 model prompted with gold triggers successfully generates a high-quality question 62.5% of the time (Table 5). This model significantly outperforms BART when given gold-standard triggers. However, the performance significantly drops when the triggers are no longer provided. We find that T0 produces a large number of *trivial* questions when given a predicted trigger. More testing and investigation is needed to further

Context	Generated Question	Trigger Type	Question Type
Pt reports that he noticed a *right neck mass* last October	Size, outline (asymmetry), color, elevation, evolving?	sign/symptom	sign/symptom
She was also significantly *tachypneic*	were there interventions done to address this?	sign/symptom	treatment
According to Dr. <name>, she has had stable deficits for many years *without any flare-like episodes*.	How is her vision now?	assessment	sign/symptom
Her bicarb began to drop and she developed an *anion gap acidosis*	confusion? confusion? agitation? hand tremors? bounding pulses?	problem	sign/symptom

Table 3: Example T0 model generations, cherry-picked. This model examines single sentences and is trained with combined questions. Trigger phrases are *italicized*.

Model Type	Context	Split Qs	Unique Question Ratio	METEOR	BERTScore	ROUGE-L
BART	Trigger	N	0.301	3.6	0.856	10.2
BART	Trigger	Y	0.037	0.1	0.838	3.4
BART	Sentence	N	0.526	6.1	0.860	10.2
BART	Sentence	Y	0.468	7.8	0.858	12.0
BART	Chunk	N	0.741	7.9	0.861	11.9
BART	Chunk	Y	0.619	7.2	0.861	11.6
T0-11B	Sentence	N	**0.779**	3.9	0.861	11.9
T0-11B	Sentence	Y	0.410	**8.4**	**0.884**	12.2
T0-11B	Chunk	N	0.398	3.7	0.860	**12.4**
T0-11B	Chunk	Y	0.400	6.7	0.879	10.9

Table 4: Automated metrics for baseline models on the question generation task. *Sentence* and *Chunk* contexts include both the text surrounding the trigger and the trigger itself. *Trigger* context only includes trigger text. Split Qs means splitting multiple questions for a trigger into multiple examples (unique question ratio of these models should not be compared). Results given on dev set.

understand this large drop in performance, as we do not observe this same behavior with BART.

As human evaluation demonstrates, despite low automatic metric scores, both BART and T0 achieve reasonable success in generating coherent, relevant and clinically interesting questions. To evaluate if the automated metrics can capture the quality of generated questions, we calculate the Spearman's Rank Correlation Coefficient between human evaluation and automatic metrics. We find extremely low and statistically insignificant correlation for ROUGE-L (-0.09), METEOR (-0.04) and BERTScore (-0.04). This is unsurprising, as these automatic metrics are not designed to capture the categories we examine during human evaluation.

We also score the answers selected by our ClinicalBERT model trained on emrQA (Section 6.3). Interestingly, we find that of the answers the model successfully recovers, 44% are extracted from the remainder of the discharge summary used to generate the question. The remaining 56% come from nursing notes, Radiology/ECG reports and previous discharge summaries. However, for a majority of the questions, we are unable to recover a sufficient answer (Table 6). We sample 50 gold standard questions whose suggested answers were marked as invalid, in order to determine if this was due to the model's poor performance. We find that 36% of the questions do in fact have answers in the EHR, thus demonstrating the need for improved clinical QA resources and models.

9 Conclusion

We present **Di**scharge **S**ummary **C**linical **Q**uestions (DiSCQ), a new human-generated clinical question dataset composed of 2000+ questions paired with the snippets of text that prompted each question. We train baseline models for trigger detection and question generation. We find that despite poor performance on automatic metrics, we are

Model	Triggers	Understandable	Nontrivial	Relevant	Clinically Meaningful	Satisfies All
Gold	-	93.8%	86.0%	83.3%	82.3%	80.5%
BART	Gold	81.5%	59.8%	52.3%	54.8%	47.8%
T0	Gold	85.8%	72.3%	68.0%	66.5%	**62.5%**
BART	Predicted	78.3%	57.3%	49.3%	49.8%	41.8%
T0	Predicted	76.8%	49.0%	45.0%	44.5%	41.0%

Table 5: We present results of human evaluation on generated questions. Gold refers to questions generated by medical experts. We do not annotate whether or not a question is nontrivial, relevant and clinically meaningful if it is not understandable, thus lowering the number of questions that satisfy these categories.

Model	Triggers	Partially	Fully
Gold	-	15.0%	7.50%
BART	Gold	13.75%	7.75%
T0	Gold	11.5%	6.00%
BART	Predicted	14.5%	6.25%
T0	Predicted	9.75%	3.25%

Table 6: Percent of the time that the answer retrieved by our model partially answers and fully answers the question.

able to produce reasonable questions in a majority of cases when given triggers selected by medical experts. However, we find that performance significantly drops when given machine predicted triggers. Further, we find that baseline models trained on emrQA are insufficient for recovering answers to both human and machine generated questions. Our results demonstrate that existing machine learning systems, including large-scale neural networks, struggle with the tasks we propose. We encourage the community to improve on our baseline models. We release this dataset and our code to facilitate further research into realistic clinical question answering and generation here.

10 Acknowledgements

This work was supported and sponsored by the MIT-IBM Watson AI Lab. The authors would like to thank Sierra Tseng for feedback on a draft of this manuscript, as well as Melina Young and Maggie Liu for their help in designing some of the figures.

References

Asma Ben Abacha, Eugene Agichtein, Yuval Pinter, and Dina Demner-Fushman. 2017. Overview of the medical question answering task at trec 2017 liveqa. In *TREC*.

Asma Ben Abacha, Yassine Mrabet, Mark E. Sharp, Travis R. Goodwin, Sonya E. Shooshan, and Dina Demner-Fushman. 2019. Bridging the gap between consumers' medication questions and trusted answers. *Studies in health technology and informatics*, 264:25–29.

Asma Ben Abacha and Pierre Zweigenbaum. 2015. Means: A medical question-answering system combining nlp techniques and semantic web technologies. *Inf. Process. Manag.*, 51:570–594.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Stephen H. Bach, Victor Sanh, Zheng-Xin Yong, Albert Webson, Colin Raffel, Nihal V. Nayak, Abheesht Sharma, Taewoon Kim, M Saiful Bari, Thibault Fevry, Zaid Alyafeai, Manan Dey, Andrea Santilli, Zhiqing Sun, Srulik Ben-David, Canwen Xu, Gunjan Chhablani, Han Wang, Jason Alan Fries, Maged S. Al-shaibani, Shanya Sharma, Urmish Thakker, Khalid Almubarak, Xiangru Tang, Dragomir Radev, Mike Tian-Jian Jiang, and Alexander M. Rush. 2022. Promptsource: An integrated development environment and repository for natural language prompts.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Hangbo Bao, Li Dong, Furu Wei, Wenhui Wang, Nan Yang, Xiaodong Liu, Yu Wang, Songhao Piao, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2020. Unilmv2: Pseudo-masked language models for unified language model pre-training. In *ICML*.

Manik Bhandari, Pranav Narayan Gour, Atabak Ashfaq, Peng fei Liu, and Graham Neubig. 2020. Re-

evaluating evaluation in text summarization. *ArXiv*, abs/2010.07100.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of Bleu in machine translation research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 249–256, Trento, Italy. Association for Computational Linguistics.

Yonggang Cao, F. Liu, Pippa M Simpson, Lamont D. Antieau, Andrew S. Bennett, James J. Cimino, John W. Ely, and Hong Yu. 2011. Askhermes: An online question answering system for complex clinical questions. *Journal of biomedical informatics*, 44 2:277–88.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom B. Brown, Dawn Xiaodong Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. 2021. Extracting training data from large language models. In *USENIX Security Symposium*.

Yllias Chali and Sadid A. Hasan. 2015. Towards topic-to-question generation. *Computational Linguistics*, 41(1):1–20.

Donna D'Alessandro, Clarence Kreiter, and Michael Peterson. 2004. An evaluation of information-seeking behaviors of general pediatricians. *Pediatrics*, 113:64–9.

Dina Demner-Fushman, Wendy Chapman, and Clement Mcdonald. 2009. What can natural language processing do for clinical decision support? *Journal of biomedical informatics*, 42:760–72.

Dina Demner-Fushman, Yassine Mrabet, and Asma Ben Abacha. 2020. Consumer health information and question answering: helping consumers find answers to their health-related information needs. *Journal of the American Medical Informatics Association : JAMIA*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation.

Xinya Du, Junru Shao, and Claire Cardie. 2017. Learning to ask: Neural question generation for reading comprehension. *CoRR*, abs/1705.00106.

Nan Duan, Duyu Tang, Peng Chen, and Ming Zhou. 2017. Question generation for question answering. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 866–874, Copenhagen, Denmark. Association for Computational Linguistics.

Desmond Elliott and Frank Keller. 2014. Comparing automatic evaluation measures for image description. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 452–457, Baltimore, Maryland. Association for Computational Linguistics.

Jungwei Fan. 2019. Annotating and characterizing clinical sentences with explicit why-QA cues. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 101–106, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Michael Heilman and Noah A. Smith. 2010. Good question! statistical ranking for question generation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 609–617, Los Angeles, California. Association for Computational Linguistics.

Qingbao Huang, Mingyi Fu, Linzhang Mo, Yi Cai, Jingyun Xu, Pijian Li, Qing Li, and Ho-fung Leung. 2021. Entity guided question generation with contextual structure and sequence information capturing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):13064–13072.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035.

Jared Kaplan, Sam McCandlish, T. J. Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeff Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *ArXiv*, abs/2001.08361.

Gregory Kell, Iain Marshall, Byron Wallace, and Andre Jaun. 2021. What would it take to get biomedical QA systems into practice? In *Proceedings of the 3rd*

Workshop on Machine Reading for Question Answering, pages 28–41, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Igor Labutov, Sumit Basu, and Lucy Vanderwende. 2015. Deep questions without deep understanding. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 889–898, Beijing, China. Association for Computational Linguistics.

Eric P. Lehman, Sarthak Jain, Karl Pichotta, Yoav Goldberg, and Byron C. Wallace. 2021. Does bert pretrained on clinical notes reveal sensitive data? *ArXiv*, abs/2104.07762.

Adam D. Lelkes, Vinh Q. Tran, and Cong Yu. 2021. Quiz-style question generation for news stories.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Jennifer J. Liang, Eric Lehman, Ananya S. Iyengar, Diwakar Mahajan, Preethi Raghavan, Cindy Y. Chang, and Peter Szolovits. 2022. Towards generalizable methods for automating risk score calculation. In *Proceedings of the 21st SIGBioMed Workshop on Biomedical Language Processing*. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing.

Hongyin Luo, Seunghak Yu, Shang-Wen Li, and James R. Glass. 2021. Cooperative learning of zero-shot machine reading comprehension. *ArXiv*, abs/2103.07449.

Karen Mazidi and Rodney D. Nielsen. 2014. Linguistic considerations in automatic question generation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 321–326, Baltimore, Maryland. Association for Computational Linguistics.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California. Association for Computational Linguistics.

Lidiya Murakhovs'ka, Chien Sheng Wu, Tong Niu, Wenhao Liu, and Caiming Xiong. 2021. Mixqg: Neural question generation with mixed answer types. *ArXiv*, abs/2110.08175.

Preksha Nema, Akash Kumar Mohankumar, Mitesh M. Khapra, Balaji Vasan Srinivasan, and Balaraman Ravindran. 2019. Let's ask again: Refine network for automatic question generation. *ArXiv*, abs/1909.05355.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. Scispacy: Fast and robust models for biomedical natural language processing. *ArXiv*, abs/1902.07669.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252, Copenhagen, Denmark. Association for Computational Linguistics.

Anusri Pampari, Preethi Raghavan, Jennifer Liang, and Jian Peng. 2018. emrqa: A large corpus for question answering on electronic medical records.

Liangming Pan, Yuxi Xie, Yansong Feng, Tat-Seng Chua, and Min-Yen Kan. 2020. Semantic graphs for generating deep questions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1463–1475, Online. Association for Computational Linguistics.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Raul Puri, Ryan Spring, Mohammad Shoeybi, Mostofa Patwary, and Bryan Catanzaro. 2020. Training question answering models from synthetic data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5811–5826, Online. Association for Computational Linguistics.

Weizhen Qi, Yu Yan, Yeyun Gong, Dayiheng Liu, Nan Duan, Jiusheng Chen, Ruofei Zhang, and Ming Zhou. 2020. ProphetNet: Predicting future n-gram for sequence-to-SequencePre-training. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2401–2410, Online. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Preethi Raghavan, Jennifer J Liang, Diwakar Mahajan, Rachita Chandra, and Peter Szolovits. 2021. emrKBQA: A clinical knowledge-base question answering dataset. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 64–73, Online. Association for Computational Linguistics.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Stella Biderman, Leo Gao, Tali Bers, Thomas Wolf, and Alexander M. Rush. 2021. Multitask prompted training enables zero-shot task generalization.

Max E. Savery, Asma Ben Abacha, Soumya Gayen, and Dina Demner-Fushman. 2020. Question-driven summarization of answers to consumer health questions. *Scientific Data*, 7.

Timo Schick and Hinrich Schütze. 2021. It's not just size that matters: Small language models are also few-shot learners. *ArXiv*, abs/2009.07118.

Iulian Vlad Serban, Tim Klinger, Gerald Tesauro, Kartik Talamadupula, Bowen Zhou, Yoshua Bengio, and Aaron C. Courville. 2016. Multiresolution recurrent neural networks: An application to dialogue response generation. *CoRR*, abs/1606.00776.

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural responding machine for short-text conversation. *CoRR*, abs/1503.02364.

Heung-Yeung Shum, Xiaodong He, and Di Li. 2018. From eliza to xiaoice: Challenges and opportunities with social chatbots.

Sarvesh Soni, Meghana Gudala, Daisy Zhe Wang, and Kirk Roberts. 2019. Using fhir to construct a corpus of clinical questions annotated with logical forms and answers. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 2019:1207–1215.

Sandeep Subramanian, Tong Wang, Xingdi Yuan, Saizheng Zhang, Adam Trischler, and Yoshua Bengio. 2018. Neural models for key phrase extraction and question generation. In *Proceedings of the Workshop on Machine Reading for Question Answering*, pages 78–88, Melbourne, Australia. Association for Computational Linguistics.

Simon Suster and Walter Daelemans. 2018. Clicr: a dataset of clinical case reports for machine reading comprehension. In *NAACL*.

Duyu Tang, Nan Duan, Tao Qin, and Ming Zhou. 2017. Question answering and question generation as dual tasks. *CoRR*, abs/1706.02027.

Luu Anh Tuan, Darsh J Shah, and Regina Barzilay. 2019. Capturing greater context for question generation.

Bingning Wang, Xiaochuan Wang, Ting Tao, Qi Zhang, and Jingfang Xu. 2020a. Neural question generation with answer pivot. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9138–9145.

Liuyin Wang, Zihan Xu, Zibo Lin, Haitao Zheng, and Ying Shen. 2020b. Answer-driven deep question generation based on reinforcement learning. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5159–5170, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Hong Yu, Minsuk Lee, David R. Kaufman, John W. Ely, Jerome A. Osheroff, George Hripcsak, and James J. Cimino. 2007. Development, implementation, and a cognitive evaluation of a definitional question answering system for physicians. *Journal of biomedical informatics*, 40 3:236–51.

Xiang Yue, Bernal Jimenez Gutierrez, and Huan Sun. 2020. Clinical reading comprehension: A thorough analysis of the emrQA dataset. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4474–4486, Online. Association for Computational Linguistics.

Xiang Yue, Xinliang Frederick Zhang, Ziyu Yao, Simon Lin, and Huan Sun. 2021. Cliniqg4qa: Generating diverse questions for domain adaptation of clinical question answering.

M. A. H. Zahid, Ankush Mittal, Ramesh Chandra Joshi, and Gowtham Atluri. 2018. Cliniqa: A machine intelligence based clinical question answering system. *ArXiv*, abs/1805.05927.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert.

Ming Zhu, Aman Ahuja, Da-Cheng Juan, Wei Wei, and Chandan K. Reddy. 2020. Question answering with long multiple-span answers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3840–3849, Online. Association for Computational Linguistics.

A Appendix

A.1 Model and Metric Implementation

To run BART and T0, we make use of the Huggingface implementations (Wolf et al., 2019). We additionally calculate automated metrics for question generation using Huggingface. For calculating Cohen Kappa, precision, recall, and F1 score, we use sklearn (Pedregosa et al., 2011).

A.2 Model Hyperparameters

We use a majority of the default settings provided by the Huggingface library (Wolf et al., 2019). However, we do experiment with varying learning rates (2e-5, 2e-4, 3e-4, 4e-4), warm up steps (100, 200), and weight-decay (0, 1e-6, 1e-3, 1e-1). For the best BART model, we find that using a learning rate of 2e-4, warm up steps of 200, and weight decay of 1e-6 led to the best model. For the T0 model, we find that using a learning rate of 3e-4, running for 100 warmup steps and using a weight-decay of 0.1 led to the best performance. We run for 50 epochs on the BART model and 30 epochs on the T0 model. We use the best epoch with respect to evaluation loss. In our dev set evaluation, we use a beam search width of 5. We use a gradient accumulation step of 32 and 16 for our BART model and T0 model, respectively,

A.3 GPUs and Run Time

For the BART models, we run on 4 GeForce GTX TITAN X. Due to the limited size of these GPUs, we only use a batch size of 1 per GPU. The BART style models take roughly 8 hours to finish training.

For the T0 models, we train using eight V100 GPUs. We set batch size to be 2 per GPU. These models take roughly 24 hours to train.

A.4 Risk of Patient Privacy

We will release our code and data under MIMIC-III access. Carlini et al. (2021) warns against training large-scale transformer models (particularly ones for generation) on sensitive data. Although MIMIC-III notes consist of deidentified data, we will not release our model weights to the general public. With respect to the trigger detection system, there is less risk in releasing the model weights, as BERT has not been pretrained with generation tasks (Lehman et al., 2021). We caution all follow up work to take these privacy concerns into account.