

PaniniQA: Enhancing Patient Education Through Interactive Question Answering

Pengshan Cai^{*1}, Zonghai Yao^{*1}, Fei Liu², Dakuo Wang³, Meghan Reilly⁴, Huixue Zhou⁵
Lingxi Li¹, Yi Cao¹, Alok Kapoor⁴, Adarsha Bajracharya⁴, Dan Berlowitz⁶, Hong Yu^{1,4,6}

University of Massachusetts, Amherst, USA¹, Emory University, USA²,

Northeastern University, USA³, UMass Chan Medical School, USA⁴,

University of Minnesota, USA⁵, University of Massachusetts, Lowell, USA⁶

{pengshancai, zonghaiyao, lingxili, yicao, hongyu}@umass.edu,

fei.liu@emory.edu, d.wang@neu.edu,

Meghan.Reilly@umassmed.edu, zhou1742@umn.edu

{Alok.Kapoor, adarsha.Bajracharya}@umassmemorial.org,

Dan.Berlowitz@uml.edu

Abstract

A patient portal allows discharged patients to access their personalized discharge instructions in electronic health records (EHRs). However, many patients have difficulty understanding or memorizing their discharge instructions (Zhao et al., 2017). In this paper, we present *PaniniQA*, a patient-centric interactive question answering system designed to help patients understand their discharge instructions. PaniniQA first identifies important clinical content from patients' discharge instructions and then formulates patient-specific educational questions. In addition, PaniniQA is also equipped with answer verification functionality to provide timely feedback to correct patients' misunderstandings. Our comprehensive automatic & human evaluation results demonstrate our PaniniQA is capable of improving patients' mastery of their medical instructions through effective interactions.¹

1 Introduction

Limited patient understanding of their medical conditions can lead to poor self-care at home. Upon hospital discharge, physicians often provide discharge instructions to aid in patients' recovery and disease self-management (Federman et al., 2018). However, some patients may have difficulty understanding and memorizing instructions due to low health literacy, limited memory, or

an absence of supervision. For example, research shows that patients only retain a minimal amount of information from discharge instructions, with an immediate forgetting rate of up to 80% (Kessels, 2003; Richard et al., 2017). Further, when instructions are misinterpreted by patients, there is often a lack of corrective intervention. Limitations in a patient's understanding of their medical conditions hinder their prospects of recovery. It is imperative to investigate new methods of patient education to enhance health outcomes.

In this study, we explore a novel method inspired by *Dialogic Reading* (Whitehurst, 2002) to educate patients through interactive question-answering. Dialogic Reading actively involves patients in the learning process by following the P.E.E.R. sequence: Prompt, Evaluate, Expand, and Repeat, which enables patients to engage in a meaningful dialogue, further strengthening their understanding and retention of the material. As illustrated in Figure 1, our dialog agent asks questions about key aspects of discharge instructions and encourages patients to read and understand the instructions to provide accurate answers thoroughly.

Crafting questions that effectively meet educational objectives is challenging (Boyd-Graber and Börschinger, 2020; Dugan et al., 2022). A suitable question should be based on the patient's discharge instruction and aim to improve their understanding of health conditions, such as “*What was the probable cause of your chest pain?*”. Conversely, the question “*How does cardiac catheterization help treat a heart attack?*” illustrated in Figure 1, may exceed the education scope, as it is unanswerable

^{*}Indicates equal contribution.

¹Our data and code are released at <https://github.com/pengshancai/PaniniQA>.

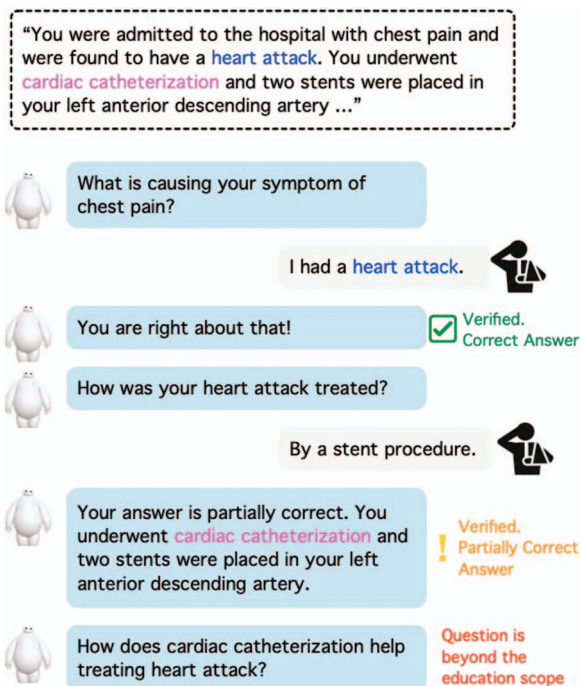


Figure 1: An illustration of PaniniQA, our interactive question-answering system for patient education. It generates questions from discharge instructions, helping patients understand their health conditions through interactive question answering. An answer verification module confirms correct responses or expands feedbacks on partially correct ones. The final turn shows a GPT-generated question. Its answer is absent from the discharge instruction and it is deemed inappropriate for patient education.

or requires knowledge beyond the provided discharge instruction. Such questions are considered unsuitable for patient education.

We introduce new question-generation methods that draw on the advancements of large language models (LLMs) (Brown et al., 2020; Ouyang et al., 2022; OpenAI, 2023). Utilizing OpenAI’s GPT-3.5 model, we generate informative questions from discharge instructions. Further, *we combine LLMs with medical event and relation extraction to constrain the model*, producing questions that target salient medical events identified in the discharge instructions. We create a new dataset with expert-annotated medical events and relations for discharge instructions from the MIMIC-III (Johnson et al., 2016a) database. While earlier efforts have annotated events that physicians would discuss during patient handoff (Pampari et al., 2018; Lehman et al., 2022), our focus is on identifying pairs of medical events with correlational or causal relationships. By pos-

ing questions about one event, we guide patients toward the other as potential answers.

Our system further incorporates an answer verification module to provide instant patient feedback. When patients give correct answers, the bot confirms them, reinforcing their understanding. If answers are incorrect or partially correct, the bot clarifies misunderstandings and provides additional information. Extensive automatic and human evaluations demonstrate the efficacy of our question-generation methods and show that PaniniQA holds great promise for promoting patient education. To summarize, our research contributions are as follows.

- ◊ We explore a new way of educating patients regarding their health conditions through interactive question-answering. Our approach aligns with the P.E.E.R. dialogic reading theory that promotes patients’ active participation in comprehending medical events.
- ◊ We compare questions generated using OpenAI’s GPT-3.5 model, our enhanced method with medical event extraction, and human-written questions tailored for patient education. We meticulously evaluated all questions, answers, and patients’ educational outcomes.
- ◊ Through comprehensive human evaluations, we demonstrate that PaniniQA holds promise for patient education. Future work includes controlling the difficulty of questions, prioritizing questions given patients’ health literacy, and enabling interactive learning of medical concepts.

2 Related Work

There is a growing need to improve patients’ understanding regarding their hospital experiences (Federman et al., 2018; Weerahandi et al., 2018; Kwon et al., 2022). Lack of understanding can result in non-adherence to discharge instructions and readmission to the hospital due to poor self-care at home. Previous research has attempted to generate hospital course summaries for patients using lay language (Di Eugenio et al., 2014; Acharya et al., 2018; Adams et al., 2021; Cai et al., 2022a; Hartman and Campion, 2022; Adams et al., 2022). This paper goes a step further by utilizing interactive question answering to communicate

essential medical events from discharge instructions to patients, thus enhancing their understanding and retention of the material.

Our proposed method differs from existing clinical question-answering studies in several aspects. Most clinical QAs are designed to satisfy individuals' information needs, with questions modeled after those that can be asked by physicians (Pampari et al., 2018; Jin et al., 2019; Raghavan et al., 2021; Lehman et al., 2022). These systems focus on improving the accuracy of their answers (Soni and Roberts, 2020; Rawat et al., 2020; Yue et al., 2020a,b). In contrast, our goal is to educate patients and prompt them with questions that will enhance patients' understanding of their doctors' visits. A successful QA system should be *comprehensive* and *exhaustive*, asking all relevant questions and prioritizing them based on the patient's medical history and health literacy.

Successful patient education requires effective questioning (Pylman and Ward, 2020). Particularly, question generation has been studied using template-based (Heilman and Smith, 2010; Chali and Hasan, 2015; Fabbri et al., 2020) and neural seq2seq models (Du and Cardie, 2017; Duan et al., 2017; Kim et al., 2018; Sultan et al., 2020; Shwartz et al., 2020). Instruction-tuned LLMs have demonstrated exceptional abilities in conversing with humans (Brown et al., 2020; Sanh et al., 2021; Ouyang et al., 2022; Chowdhery et al., 2022; Longpre et al., 2023). However, most research has been conducted using CommonCrawl, Wikipedia, and other generic texts. Considering the factuality issues of neural language models (Maynez et al., 2020; Pagnoni et al., 2021), question generation in the medical domain remains challenging.

Learning through conversation can improve education outcomes (Golinkoff et al., 2019; Zhang et al., 2020; Cai et al., 2022b; Yao et al., 2022a,a; Xu et al., 2022). *Dialogic Reading* (Whitehurst, 2002; Mol et al., 2008; Lever and Sénéchal, 2011) has demonstrated that engaging children in a guided conversation with parents while reading storybooks can significantly enhance their learning outcomes. While engaging physicians in high-quality conversations may not always be feasible, the use of question answering facilitated by a chatbot could be a valuable means of helping patients acquire a deeper understanding of their health conditions.

3 Question Answering in the GPT Era

LLMs such as ChatGPT have led to significant advancements in generative AI (Brown et al., 2020; Sanh et al., 2021; Chowdhery et al., 2022; Longpre et al., 2023; OpenAI, 2023; Wang et al., 2023a). Fine-tuning neural models on specific tasks often yields superior results. Furthermore, LLMs acquire emergent abilities through instruction tuning and reinforcement learning using human feedback (Ouyang et al., 2022). This allows them to generalize to new tasks effectively. Common human-LLM interactions include (a) *zero-shot prompting*, where users provide a prompt for the LLM to complete, and (b) *in-context learning*, where users give task examples and ask the LLM to solve a new case, potentially involving a multi-step reasoning process (Wei et al., 2022). In this study, we focus on zero-shot prompting to assess the LLM's ability to comprehend discharge instructions.

LLMs possess vast world knowledge, and their performance on knowledge-intensive tasks correlates with training data and model size (Bommasani et al., 2022). However, it remains unclear whether LLMs have enough domain knowledge to facilitate patient education. For example, GPT-3, with its 175 billion parameters, is trained on general data sources such as Common Crawl, WebText2, Books, and Wikipedia (Brown et al., 2020). Yet, the model still generates factually inconsistent errors within their output. Our study presents an initial evaluation of GPT models' potential in interactive patient education. Following the P.E.E.R. framework of dialogic reading, we employ GPT models to perform the following tasks:

Question Generation. We use OpenAI's GPT-3.5 model (`text-davinci-003`) to generate informative questions from a discharge instruction. The questions aim at helping patients understand crucial medical events. Our prompt is “*Generate N questions to help the patient understand crucial medical events in the above discharge instruction.*” Similar to a teacher designing exam questions, we anticipate the GPT model to produce a set of questions all at once rather than incrementally. The questions must collectively cover the salient events identified in the discharge instruction while minimizing redundancy.

Answer Verification. Useful feedback is essential for improving patient comprehension of the material. To perform this task, we prompt the GPT model with “*As a physician, your goal in the conversation is to help your patient better understand the discharge instructions before they leave the hospital.*” Utilizing OpenAI’s API, we also provide the original discharge instruction, interaction history, and current question-answer pair as key-value pairs for the model. We then instruct the model to “*verify if the patient’s answer is correct, incorrect, or partially correct, and generate a suitable response to improve the patient’s comprehension of this question.*” We empirically compared two GPT models, `text-davinci-003` and `gpt-3.5-turbo` (ChatGPT), and selected ChatGPT for answer verification as it is optimized for chat and generally produces higher quality responses.

4 Extracting Salient Medical Events

In this section, we present our question-answering system that emphasizes identifying salient medical events and their relations. We generate targeted questions using them and apply the same answer verification module described previously.

A typical discharge instruction includes **Visit Recap**, which recaps a patient’s clinical visit, including symptoms, diagnoses, treatments, and test results. Patients are expected to understand the relationships among these medical events, such as how the treatment *ERCP* relates to *cholangitis* as illustrated in Table 2 (top). **Detailed Instructions** include medication and aftercare instructions (bottom). They may be easy to understand but contain trivial details that patients may overlook, potentially hindering their self-care at home. We propose automatically extracting key medical events and relations from them (§4.1). Given their unique characteristics, we apply two distinct information extraction and question generation strategies for *Visit Recap* and *Detailed Instructions* to produce targeted questions (§4.2).

4.1 Event and Relation Identification

Key event and relation identification are conducted on **Visit Recap**. Event identification is framed as a *sequence labeling* task, where we assign a label to each token of the discharge note, representing its event type. We define 11 event

types in this study, detailed in Table 3, including symptoms, diseases, complications, tests, test goals/results/implications, procedures, medicines, treatment goals, and results. We fine-tune pre-trained sequence labeling models on our dataset, optimizing the cross-entropy loss of gold standard labels.

Relation identification is framed as a *sequence classification* task. We focus on binary relations consisting of two medical events. We evaluate all pairwise combinations of identified medical events as candidates, provided their event types align with the six event relations defined in Table 1. Special tokens are inserted before and after each identified event to indicate both its position and event type.² The sequence, enhanced with special tokens, is fed into a sequence classification model to predict a binary label, where 1 indicates a relation between the two events, and 0 otherwise. We fine-tune pre-trained sequence classification models on our dataset (§5) by optimizing the cross-entropy loss for gold-standard labels.

We perform key event identification on **Detailed Instructions** using a different tool, as they contain medication and aftercare specifics that patients might overlook. We use an existing high-performing medical NER system to extract medical entities.³ This model was pre-trained on the MACCROBAT dataset (Caufield et al., 2019) and can identify 84 biomedical entities within clinical narratives. We limit the model to identify 7 entity types: *Medicine Dosage*, *Medicine Frequency*, *Medicine Duration*, *Medication Name*, *Sign & Symptom*, *Diagnostic Procedure*, *Upcoming Appointment*. Relation identification is not performed on detailed instructions.

4.2 Question Generation

Visit Recap. We generate a question from each identified binary relation. Different relation types are mapped to specific questions using templates

²E.g., the sentence “*You were admitted for diverticulitis and treated with antibiotics*” was modified as “*You were admitted for <dsyn> diverticulitis </dsyn> and treated with <medi> antibiotics </medi>*, where the special tokens <dsyn> and </dsyn>” indicates the start and end position of this event, and *dsyn* reflects the event belongs to the category *Disease*.

³<https://pypi.org/project/Bio-Epidemiology-NER/>.

BINARY RELATION OF EVENTS	QUESTION TEMPLATES
Symptom / Disease	Q: What is the cause of your symptom [Symptom]? A: [Disease]
Test / Test goal	Q: What is the goal of test [Test]? A: [Test goal]
Test / Test result	Q: What is the result of test [Test]? A: [Test result]
Test / Test implication	Q: What does test [Test] imply? A: [Test implication]
Procedure (or Medicine) / Treatment goal	Q: What is the goal of treatment [Procedure] (or [Medicine])? A: [Treatment goal]
Procedure (or Medicine) / Treatment result	Q: What is the result of treatment [Procedure] (or [Medicine])? A: [Treatment result]

Table 1: Expert-written question templates are used to generate a question from each binary relation. This method enables us to create targeted questions about salient medical events. By posing questions about one event, we guide patients towards the other as potential answers. The placeholders are to be replaced with medical events detected from discharge instructions.

QG - VISIT RECAP
<i>You were found to have an infection of your bile ducts called cholangitis. You had a procedure called an ERCP where a stent was placed to relieve the obstruction ...</i>
Relation: ERCP (Procedure) – cholangitis (Treatment Goal)
Question: What is the goal of treatment ERCP ?
QG - DETAILED INSTRUCTIONS
<i>We made the following changes to your medication regimen: 1. We started you on a new medication called Toprol XL 25mg by mouth twice a day ...</i>
Event: twice a day (Medicine Frequency)
Question: How often should Toprol XL be taken?

Table 2: Question generation (QG) from a medical event (bottom) or a binary relation of events (top).

provided by physicians according to their domain knowledge (see Table 1). Using a template-based approach allows us to create questions targeting salient medical events. By asking questions about one event, we guide patients towards the other as potential answers.

Detailed Instructions. We generate a question for each identified medical entity by creating a fill-in-the-blank question, which is then converted into a natural language question using the GPT model. An example is shown in Table 2. Although cloze-style questions can serve educational purposes, we want to prevent patients from using string matching to find answers. Instead, natural language questions require patients to have a deeper understanding of the discharge note, thus fulfilling our education objective. When selecting medical entities as triggers, we prioritize four categories: *Medicine Dosage*, *Medicine Frequency*, *Medicine Duration*, and *Upcoming Appointment*, as they are informative and better guide patient comprehension. To convert a cloze-style question into a natural question, we provide this prompt to the GPT model: *[Fill-in-the-Blank Sentence]*

Generate a simple question targeting the blank in the above sentence.

5 Data Annotation

We seek to annotate discharge instructions from the MIMIC-III database (v1.4) (Johnson et al., 2016a) with key medical events that are important for patients to understand. MIMIC-III is a publicly available repository of de-identified health records of over 40,000 patients collected from the Beth Israel Deaconess Medical Center in Massachusetts. Our aim is to identify text snippets in discharge instructions that correspond to significant medical events, including *symptoms*, *diseases*, *test results*, and *treatments*. We annotate not only individual events but also their relationships. They are organized into a *hierarchy* as outlined in the schema shown in Table 3. Consistent with Lehman et al.’s (2022) approach, we utilize events and their relationships as *triggers* that prompt the generation of questions.

We recruited five medical experts to create a sizeable dataset. They are MD students at UMass Chan Medical School and have a high level of language proficiency. Each expert is given 150 discharge notes to annotate. It is possible to skip some notes due to low text quality. Annotators were also given detailed instructions and examples. We developed a web-based interface to facilitate the annotation process, which has been iteratively improved to meet the needs of this study. Due to budget constraints, we assign one annotator to each discharge note. In total, we completed 458 discharge notes with medical event annotations.

Our annotation consists of two phases. In the first phase, an expert selects text snippets from the discharge instruction corresponding to medical events that the patient needs to understand. Each

MEDICAL EVENTS & RELATIONS	# OF INSTANCES
E-1.1 Symptom	772
E-1.2 Disease	541
E-1.3 Complication	135
R-1.1 [Symptom] caused by [Disease]	323
E-2.1 Test	216
E-2.2 Test goal	17
E-2.3 Test result	208
E-2.4 Test implication	24
R-2.1 [Test] goal: [Test-Goal]	12
R-2.2 [Test] result: [Test-Result]	163
R-2.3 [Test] implication: [Test-Implication]	20
E-3.1 Treatment	
E-3.1.1 Procedure	359
E-3.1.2 Medicine	536
E-3.3 Treatment goal	86
E-3.4 Treatment result	239
R-3.1 [Treatment] goal: [Treatment-Goal]	287
R-3.2 [Treatment] result: [Treatment-Result]	237

Table 3: A hierarchy of salient medical events. We consider both medical events (E) and their binary relationships (R).

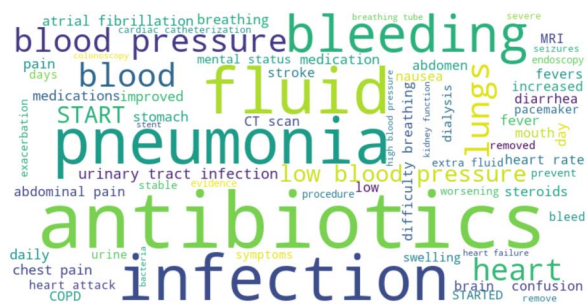


Figure 2: Word cloud demonstrating the most frequent medical terminologies and their frequency in our annotations. The sizes of the terminologies refer to their frequency in our dataset. These terminologies are identified from annotated medical events using SciSpacy.

snippet is assigned a coarse event category, such as a *medical issue*, *laboratory test*, *treatment*. The expert further refines it by assigning a fine-grained event type, resulting in a schema with 11 event types (Table 3). In the second phase, the expert identifies relationships between medical events using a set of 6 pre-defined relationships, such as “[*Symptom*] . . . *caused by* [*Disease*].” We show a distribution of medical events in Figure 2.

A key distinction between our work and earlier dataset curation efforts (Pampari et al., 2018; Yue et al., 2020b; Lehman et al., 2022) is that the ear-

lier efforts aim to annotate *questions* that physicians would ask during patient hand-off, which may be informal and unanswerable based on the discharge instruction. In contrast, our focus is on annotating *salient medical events* that are essential to patient’s understanding of their medical conditions.

We split our annotated data into train / validation / test splits, which contain 338 / 60 / 60 discharge instructions, respectively. For relation identification, we use the event pairs from the human-annotated relations as positive relations and all other medical event pairs of compliant types (e.g., the event pair types in Table 3) as negative relations. We collect all negative event pairs⁴ as negative cases. Overall, our medical relation dataset contains 2530 / 399 / 332 instances in the train / validation / test set, respectively; 28.7% instances are positive relations.

6 Evaluating Information Extraction

To improve LLMs’ ability to generate educationally effective questions for patient education, we designed an **Information Extraction (IE)** module (medical event/relation identification) to guide question generation. We report automatic evaluation results for different IE methods in this section.

6.1 IE Evaluation Settings

We fine-tune four pre-trained language models on our annotated dataset for key medical event and relation identification in Section 5. These models are obtained from HuggingFace: (1) BERT-large (Devlin et al., 2019); (2) BioBERT (Lee et al., 2020); (3) PubmedBERT (Gu et al., 2020); (4) ClinicalRoBERTa (Lewis et al., 2020); All four pre-trained models have the same scale of parameters (345 million). The later three language models were pre-trained on different bio-medical or clinical corpora, thus are better transferable to our patient education task due to the model’s level of medical knowledge (Sung et al., 2021; Yao et al., 2022b,c). The models are trained on a single RTX 6000 GPU with 24G memory. The average training time for the relation identification

⁴That is, no relationship exist between the event pair, in addition, the types of the two events are restricted by Table 1.

	Pretrained Model	P(%)	R(%)	F1(%)
Medical Events	Bert	31.38	44.58	36.83
	BioBert	40.43	51.63	45.35
	PubmedBERT	42.70	50.12	46.11
	ClinicalRoBERTa	44.28	54.03	48.67
Event Relations	Bert	57.48	75.31	65.21
	BioBert	73.41	80.37	76.73
	PubmedBERT	72.56	75.31	73.91
	ClinicalRoBERTa	74.28	82.27	78.07

Table 4: Results of fine-tuning four pretrained models on IE task: medical event extraction (Top) and event-relation identification (Bottom).

Medical Event	P (%)	R (%)	F1 (%)
Symptom	50.8	78.2	61.6
Disease	54.3	73.5	62.5
Complication	25.0	23.5	24.2
Test	65.9	82.8	73.4
Test goal	25.0	25.0	25.0
Test result	36.3	30.0	32.8
Test implication	16.6	16.6	16.6
Procedure	38.5	53.6	44.8
Medicine	42.3	42.3	42.3
Treatment goal	16.6	28.5	28.0
Treatment result	19.0	22.8	20.7
Overall	44.2	54.0	48.6

Table 5: Automatic evaluation results of medical event identification per category with ClinicalRoBERTa model.

model is around 20 minutes.⁵ For evaluation metrics, we report the model’s Micro-average precision, recall, and F-1 score.

6.2 IE Evaluation Results

The performance of four evaluated models is in Table 4. The results suggest that the models pre-trained with biomedical or clinical corpus show better performance than the naive BERT model. For both tasks, ClinicalRoberta achieves the best performance, so we report only this model’s performance in following category-wise performance analysis.

We further report more fine-grained results of the medical event extraction per category in Table 5, and the *Symptom*, *Disease*, *Test*, *Procedure*, and *Medicine* categories generally achieve better performance, as we suspect it is due to a more abundant training data. Table 6 shows the

⁵Due to data sparsity, when training both the medical event and relation identification models, we first explore the optimal hyper-parameter set using the validation set. We then combine the validation set into the train set to train our models.

Medical Event Relation	P(%)	R(%)	F1(%)
[Symptom] caused by [Disease]	81.25	79.59	80.41
[Test] goal: [Test-Goal]	100.0	60.0	75.0
[Test] result: [Test-Result]	61.54	92.31	73.85
[Test] implication: [Test-Implication]	57.14	66.67	61.54
[Treatment] goal: [Treatment-Goal]	81.82	81.82	81.82
[Treatment] result: [Treatment-Result]	70.59	85.71	77.42
Overall	74.28	82.27	78.07

Table 6: Event-Relation detection per category with ClinicalRoBERTa model.

fine-grained performance of event-relation identification per category. The F-1 scores of most relations are around 80%, implying fair performance. The relation *Test goal* achieves 100% in precision because our test set contains eight *Test goal* instances.

To explore the generalization ability of the model, we compare the model’s performance on seen and unseen medical events during training. Specifically, seen events are events that appear in the training set, while unseen events are not. We observed that 15.21% of the test instances are unseen medical events. For the medical-event extraction task, the F1 score of seen events is 49.36%, and the F1 score of unseen events is 44.82%. For the event-relation identification task, if both event in the event pair are seen events, the model achieves 78.72% in F1 score, otherwise the performance drop to 74.50%. This indicates the model only shows slight drop in performance when encountering unseen medical events during training.

7 Evaluating Patient Education

In order to comprehensively evaluate patient education outcome, We conducted human evaluations from both the patient’s and the physician’s perspectives, as well as a GPT-4 powered automatic evaluation. These evaluations focus on two main aspects: (1) The generated question’s quality of different models (GPT, GPT+IE, and human ground-truth); (2) The preference of different designs of the interaction experience (None of support, Raising Questions only, and Raising Questions and Verifying Answers).

7.1 Human Evaluation Settings

The goal of **physician evaluation** is to have human domain experts evaluate whether these

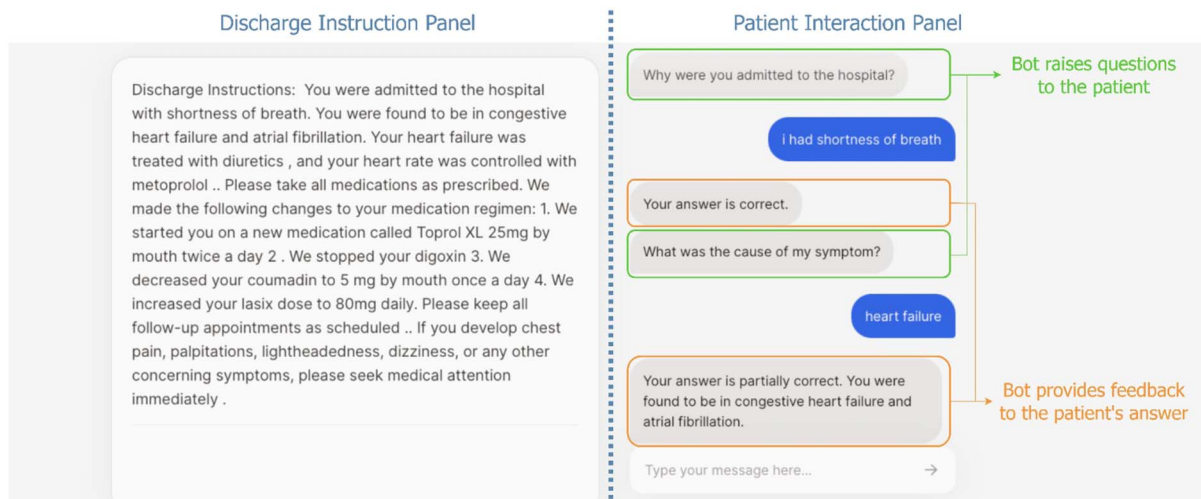


Figure 3: System UI of our human evaluation study: the left panel shows a discharge note (Condition *None*), and the right panel provides the question-answering interactions to the user via a chatbot. The bot can either present only questions (Condition *Q* in green boxes) or plus answer feedbacks (Condition *QA* in orange boxes).

machine-generated questions are comparable to the human-crafted questions or not. To do so, we recruited 3 medical practitioners⁶ and their tasks are to read the discharge instructions, and provide qualitative feedback on if these machine-generated questions are educationally effective to the patients; if not, how should they be improved.

The goal of **patient evaluation** is to have the general public users interact with and provide ratings on the different combinations of the question-generation models and the interaction designs. We also designed a post-experiment evaluation task (i.e., Cloze Test) to quantitatively measure their understanding outcome. We recruit 30 human evaluators to participate in our patient education experiment. All the evaluators have bachelor's degrees but do not have any medical education background. We present a screenshot of our patient evaluation user interface in Figure 3.

In our study, we have the following three options for the user interaction experience design:

1. Condition *None*: The evaluator only sees the discharge instruction, no question-answer interaction. This is today's baseline.
2. Condition *Q*: The evaluator reads the discharge instruction, and interact with the chatbot, which can only ask questions but do not to provide feedback to users' answers.
3. Condition *QA*: The evaluator reads the discharge instruction, and interact with the

⁶Two licensed physicians and one medical student with hospital internship experience.

chatbot, which can ask questions and provide answer feedback to the user.

The questions asked by the chatbots can come from following three sources:

1. *Human*: Expert-written questions based on discharge instructions. We ask an MD student to read each discharge instruction and write down all questions she would ask a patient about this discharge instruction for patient-education purposes.
2. *GPT*: We utilize GPT-3 model to generate a series of questions (at least four) directly from the discharge instruction. Specifically, we use the following prompt: *[Discharge Instruction] Generate at least four questions to help the patient understand crucial medical events in the above discharge instruction.*⁷
3. *GPT+IE*: Our question generation model enhances by the information extraction technique described in Section 4.

The average number of questions from approach *Human / GPT / GPT+IE* are 7.5 / 6.17 / 6.1. When combining the variety of interaction designs and

⁷We have tried a collection of prompts for the similar purpose, and do not observe significant differences in the quality of generated questions. We used the chosen prompt as it is naive to understand and leads to more succinct questions. Specifically, we instruct GPT-3 to generate at least four questions to benchmark against the least number of questions from the human annotator.

question-generation methods, there are five different conditions: (1) *None*; (2) *Q (Human)*; (3) *QA (GPT)*; (4) *QA (GPT+IE)*; (5) *QA (Human)*. We perform a within-subject experiment setup, where each of the 30 human evaluators should experience all five conditions using different discharge instructions. In total, we have 150 data points (30 per each condition). The order of the five conditions are shuffled so that each condition appears six times at each of the five orders.

7.2 Patient Evaluation Measurements

We use two measurements to evaluate patient’s educational outcome and preference.

1) **CLOZE TEST:** We recruited an MD student to identify 5–7 important medical events that she thinks the patient should be aware of, and replace them with blanks. We use these cloze tests as a post-study evaluation to ask each participant to try their best to fill in the blanks using their memory. The more blanks they fill in correctly, the better the patient’s education outcome is. We report the participant’s accuracy rate as the primary evaluation outcome.

2) **PREFERENCE RANKING:** We ask evaluators to rank their experience using the following four questionnaire items (Evaluators are allowed to rank two conditions as tied):

- *Coverage:* Does the conversation cover the cloze test in the evaluation?
- *Appropriateness:* Are the questions properly raised, and appropriate for patient education?
- *Education Outcome:* How do you think the learning experience improves your understanding of discharge instructions?
- *Overall:* How do you like the general learning experience considering the above aspects?

We report the Mean Reciprocal Rank (MRR) (Radev et al., 2002) of each model’s final ranking. Generally, a higher MRR value implies the evaluators have more preference over an approach.

7.3 GPT-4’s Automatic Evaluation Settings

Following recent practice of applying LLMs in evaluating dialogue tasks (Liu et al., 2023), we utilize GPT-4 as the evaluation model to automatically measure the quality of AI generated questions and feedback. Similar to patient evaluation in Section 7.2, we evaluate the quality

You are a physician who wants to evaluate how helpful an AI model is for educating patients. The model asks the patient questions, then verifies the patient’s answers, in order to help patients memorize their discharge instructions.

Four evaluation aspects for AI model’s question quality includes:
Coverage: Does the conversation cover the cloze test in the evaluation?

Question Appropriateness: Are the answers to the questions contained in the discharge instruction?

Education Outcome: Do you think the chatbot helps patients understand their discharge instructions?

Overall: How do you like the general experience with the chatbot considering the above aspects?

Two evaluation aspects of the AI model’s feedback includes:

Correctness: Are the responses from the AI model factually correct?

Education Potential: Do the AI model’s responses provide helpful information for educating patients?

5-point Likert scale:

1: very low rating

2: low rating

3: neutral or medium rating

4: higher rating

5: very highly rating

The patient’s discharge instructions: [*The Patient’s Discharge Instruction*] The conversation between the patient and the AI model: [*The Conversation History*]

Give the 5-point Likert scale of the AI model’s question quality (four aspects) and answer feedback (two aspects) one by one.

Return the scores as dictionary objects, adhering to the following structure: “Coverage”: . . . , “Question Appropriateness”: . . .

Please provide your response solely in the dictionary format without including any additional text.

Table 7: Prompt presented to GPT-4 for evaluating the quality of generated questions and answer verification feedback. GPT-4 is expected to output a score on each perspective directly.

of generated questions from the four perspective (i.e., *Coverage*, *Question Appropriateness*, *Education Outcome* and *Overall*). Additionally, we also evaluate the quality of AI models’ feedback from two perspectives, i.e., *Correctness* and *Education Potential*. Our prompt to the evaluation model is shown in Table 7. We collect the evaluation model’s responses and report the average score of each perspective.

7.4 Synthesized Dataset for Evaluation

Directly presenting real health records to LLMs or participants can lead to data privacy violation.⁸ Thus, we created 30 synthesized discharge instructions for our human evaluation study. We randomly sampled 30 hospital course notes (a part of EHR data) from the MIMIC-III database, and converted them into synthetic discharge instructions following a neural abstractive summarization method proposed by Cai et al. (2022a).

⁸<https://physionet.org/content/mimiciii/view-dua/1.4/>.

1	<p><i>Your symptoms improved and you were discharged to home with close follow-up with your primary care physician and an allergist ...</i></p> <p>(GPT) Q: When is your follow-up appointment with your primary care physician and allergist and what should you expect during these visits?</p>
2	<p><i>We also found that you have a condition called tracheobronchomalacia, which is a blockage of your airways. You had a stent placed in your airway to help keep it open ...</i></p> <p>(GPT) Q: How should you maintain the stent in your airway?</p>
3	<p><i>You were admitted to the hospital with fevers. You were found to have pneumonia, and you were treated with antibiotics ...</i></p> <p>(GPT+IE) Q1: What is the cause of your symptom fevers? Q2: What treatment is applied to disease pneumonia?</p>
4	<p><i>You were admitted to the hospital with swelling in your throat ... You were treated with steroids, benadryl, famotidine and epinephrine ...</i></p> <p>(GPT+IE) Q: What treatment is applied to disease swelling in your throat?</p>
5	<p><i>The following changes have been made to your medications: START Prednisone 40mg daily for 5 days ...</i></p> <p>(GPT+IE) Q: What is the recommended duration for taking Prednisone at 40mg daily?</p>

Table 8: Examples of the synthetic discharge instructions and generated questions.

Our physician collaborators reviewed these synthesized discharge instructions to ensure content validity and anonymity.

We then apply the various ways (human, GPT, GPT+IE) to create question-answer pairs for these anonymized synthesized data. We demonstrate some sampled discharge instructions and corresponding generated questions in Table 8.

7.5 Physician Evaluation Results

We interview three physician participants with the following questions: (1) Do you think the questions are effective for patients to understand the important info in the discharge instruction? If not, what questions would you ask? (2) How do you like the questions generated from GPT and GPT+IE?

Physician participants all believe that GPT-generated questions tend to target content that patients do not need to be aware of (e.g., asking why heart attack could cause chest pain is a medical-domain-specific knowledge not suitable for patient’s education). Sometimes the answers to the GPT-generated questions do not even exist in the discharge instruction. Take example 1 in Table 8, the question asks what the patient should expect in their follow-up visits, but this information is not mentioned in the discharge in-

struction. These qualitative findings may explain why GPT-generated questions’ are rated by patient participants as having a low accuracy score in the Cloze Test metric, as well as being ranked lower in Coverage, Appropriateness, and Education Outcome in Section 7.6.

Worth noting, in some cases where the answers are not in the discharge instructions, physician participants actually believe those questions could be useful for patient education. In example 2 in Table 8, although the discharge instruction does not contain information on *how to maintain the stent*, physicians still think it is a question they would ask their patients, as it would motivate patients to have better self-managed recovery activities.

For questions generated by GPT+IE, most questions were perceived by the physicians as appropriate (e.g., example 3). However, the GPT-IE may still generate improper questions due to errors in the medical event-relation identification. As shown in example 4, the information extraction model identifies the *symptom* “swelling in your throat” as a *disease*, which leads to improper questions.

Physician participants also suggested that some GPT-IE-generated questions lack language fluency. As shown in example 5, the generated question seems redundant and can be better rephrased as “How long do you need to take Prednisone?”

7.6 Patient Evaluation Results

We summarize the patient evaluation results in Figure 4. From the (a) Cloze Test chart, we observe that having a chatbot interact with patient participants (regardless of only with Q or with both QA) can indeed improve their performance over the baseline condition *None*, which suggests our proposed interactive question-answering design is a promising for patient education. In terms of whether having an answer feedback is helpful or not, the 92.7% accuracy of QA (Human) significantly outperforms the 80.6% accuracy performance of Q(Human), this implies the importance of validating patients’ answers and presenting feedback, thus we decided to always including an answer feedback when conducting further comparison analysis regarding the GPT v.s. GPT+IE question generation algorithms. The result shows that QA (GPT+IE) 88.3% achieves higher accuracy than QA (GPT) 74.1%.

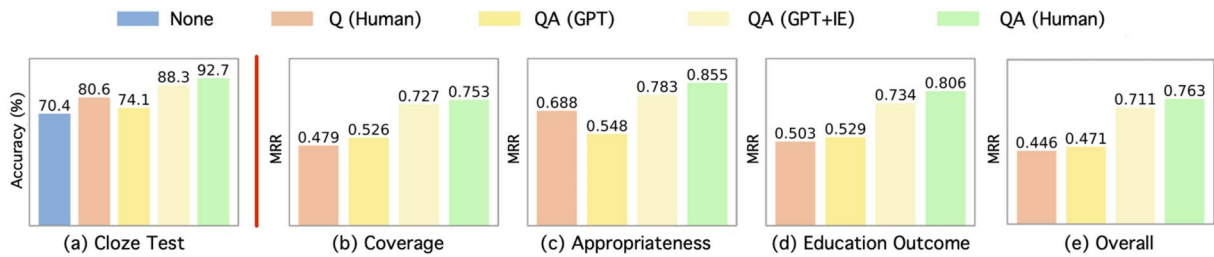


Figure 4: Patient evaluation results, including Cloze Test accuracy and evaluator rankings’ MRR scores across four categories (higher is better). The methods are represented with color-coding: *None*-blue, *Q (Human)*-orange, *QA (GPT)*-yellow, *QA (GPT+IE)*-beige, *QA (Human)*-green.

This demonstrates the improvement by applying enhancements to LLMs for patient education purposes.

The result related to Evaluator Ranking shows (plots (b, c, d, e) in Figure 4): (1) Considering the Overall ranking of three sets of questions using *QA* interactive approach, *Human* questions performs better than AI generated questions. This suggests machine-generated questions are still not comparable to human ones. (2) Comparing the three interactive approaches, we observe $QA(Human) \gg Q(Human) > None$, which is in line with the findings of Cloze Test. (3) In terms of Appropriateness, and Education outcome, *GPT* achieves the lowest ranking. According to our observation, many *GPT*-generated questions ask the evaluators about content not existing in the discharge instruction. As a result, evaluators think the questions are inappropriate and do not help patient education. (4) $QA(GPT+IE)$ has higher ranking in Coverage than $QA(GPT)$. This result is consistent with other recent discussions that incorporate the copying mechanism into LM or LLM by modifying the model structure, loss function, or prompting (Wang et al., 2023b; Chang et al., 2023; Ereemeev et al., 2023). $QA(Human)$ has higher ranking in Coverage than $Q(Human)$, despite they use the same questions. This suggests much benefit is provided to patients through the answer feedback interaction.

7.7 GTP-4’s Automatic Evaluation Results

In terms of question quality (as shown in Figure 5), we observe GPT-4’s evaluation scores generally follow the same pattern of patient evaluation results, where questions from $Q(Enhanced)$ are deemed better than $Q(Direct)$. In addition, we observe the scores of all approaches are close or higher than 4, this implies GPT-4 judges the generated questions are of good quality in four

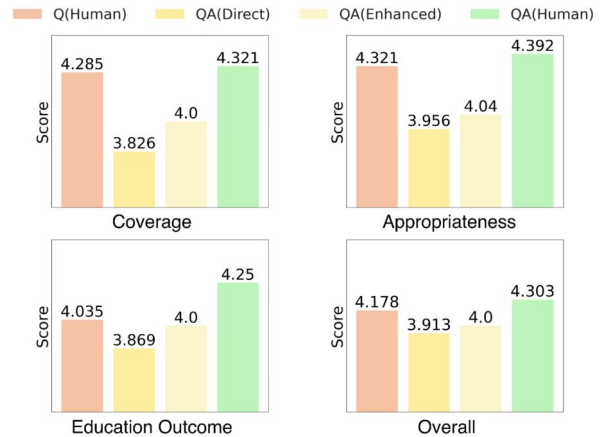


Figure 5: GPT-4’s evaluation scores for question quality.

perspectives. In terms of answer verification, as all interactive conditions all share the same verification method, we only present the average *Correctness* and *Education Potential* score. Specifically, GPT-4 gives 4.14 on *Correctness* and 4.01 on *Education Potential*. Both scores are above four, indicating GPT-4 judges feedback from our AI agents’ feedback as high quality.

7.8 Heuristic Evaluation of Conversation Log

We further conducted a heuristic evaluation to explore the deficiency of AI-generated responses and potential improvements. Specifically, we asked an MD student to evaluate the conversation log data of all patient participants.⁹ Overall, we collect 192 responses from 30 conversations between the participants mimicking patients and the AI model.

⁹The conversation logs are re-used from the patient evaluation described in section 7.6.

1	<p>Response helps patients review instructions accurately</p> <p>“The following changes were made to your medication regimen. 1. Please take HCTZ 25mg three times a day ...”</p> <p>Bot: What is the dosage of HCTZ?</p> <p>User: 25mg</p> <p>Bot: Your answer is partially correct. The dosage of HCTZ is 25mg three times a day.</p>
2	<p>Response provides information with extra educational value</p> <p>“We made the following changes to your medications: -STOPPED Verapamil -STARTED Guaifenesin for airway stent maintenance -INCREASED Nifedipine to 40 mg daily ...”</p> <p>Bot: What medicine is used to treat your disease low heart rate and high blood pressure?</p> <p>User: guaifensin and nifedipine</p> <p>Bot: Your answer is correct. Guaifenesin is used for airway stent maintenance, while Nifedipine is used to lower blood pressure.</p>
3	<p>Response is factually incorrect and confuses patient</p> <p>“We gave you medications to lower your blood pressure and your breathing improved ...”</p> <p>Bot: What is the result of the medication given to you during your hospital stay?</p> <p>User: my blood pressure decreased and my breathing improved</p> <p>Bot: The patient’s answer is partially correct. The medication helped to improve the patient’s breathing.</p>
4	<p>Response provides information with little educational value</p> <p>“The following changes were made to your medications: STOPPED Losartan CHANGED Lasix to 40mg by mouth twice daily ...”</p> <p>Bot: Should you continue to take Losartan?</p> <p>User: no, changed to lasix</p> <p>Bot: Your answer is correct.</p>

Table 9: Examples of chatbot’s answer feedback.

We ask our MD-background human evaluator to grade each of the AI model’s answer feedback, and we apply the same evaluation metric, i.e., **correctness** and **education potential** as introduced in Section 7.3. We apply binary coding, i.e., evaluator judge response as positive or negative. The positive rate for *Correctness* is 86.4%, and the positive rate for *Education Potential* is 74.1%. This suggests that most responses are factually correct and provide helpful information to patients.

Table 9 shows some examples of the answer feedback from the chatbot, and we have following design suggestions for future research to improve the quality of the answer feedback: (1) Most responses are helpful for patients in reviewing their discharge instructions (example 1). But, some responses are factually incorrect and may confuse patients. The AI model may state that the patient’s answer is incorrect or partially correct (example 3), while the patient’s response is actually completely correct. (2) While the responses are generally helpful, they still have a deficiency in providing sufficient and attentive responses in educating patients like a human physician. As shown

in example 4, a physician will provide more information about the distinctions between the two medications, including the specific diseases for which they are prescribed.

8 Limitations and Ethical Considerations

This study offers valuable insights, but with a few limitations we would like to note.

Biases. Large language models trained on vast amounts of text data can pick up biases present in data. For example, they may prefer certain questions related to aspirin or even associate certain health conditions with specific groups of people. They may also perpetuate misinformation and provide incorrect information. In addition, people who participated in our evaluation have different levels of language proficiency and medical background. These biases may be mitigated by enhancing model alignment with each individual’s background and health literacy level.

Broader Impacts. We have performed a preliminary study to educate patients on discharge instructions using interactive question answering. Although we evaluated our system using the MIMIC III dataset, which represents an intensive care unit setting, the system should be generalizable to other settings, including perioperative care (from preparation before the surgery to recovery after the surgery), cancer treatment, and chronic condition management. Our system may help patients receive customized information that is tailored to their individual needs and preferences.

Social Influence. Our system has two pillars. First, it is grounded in discharge notes, where we identify important medical events and their relationships that patients should know. Second, it serves an education purpose. For that, we explore the P.E.E.R. sequence to prompt the patient, evaluate, extend and ask them to repeat the answer to reinforce their understanding. Additionally, social influence strategies such as small talk, empathy, persuasion can be explored in the future to shape, reinforce, or change a patient’s behavior and promote engagement.

Privacy Implications. LLMs can present privacy concerns in patient education when health records are used, potentially violating the HIPPA

regulations. However, in this study, we handle data usage with great care. We conduct all experiments on *open-sourced* real patient data and present an approach to synthetic patient discharge notes. Each synthetic discharge note used in this study has been reviewed by physicians to ensure their validity. We strictly limit our API usage to synthetic data.

9 Conclusion

In this study, we present *PaniniQA*, a patient-centric interactive question answering system designed to help patients understand and memorize their discharge instructions. PaniniQA generates educational questions from discharge instructions after identifying salient medical events and event relations. LLMs with prompting is promising for question-answer generation, but sometimes hallucinating. Extensive evaluations highlight the importance of providing answer feedback.

Acknowledgments

The authors would like to express sincere gratitude to Center for Biomedical and Health Research in Data Sciences, UMass Lowell, which made this research possible. Hong Yu is supported in part by NIH R01DA056470 and 1R01AG080670, NSF IIS 2124126, and HSR&D 1I01HX003711-01A1. Fei Liu is supported in part by National Science Foundation grant IIS-2303678. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health, National Science Foundation, and Health Services Research & Development.

References

Sabita Acharya, Barbara Di Eugenio, Andrew Boyd, Richard Cameron, Karen Dunn Lopez, Pamela Martyn-Nemeth, Carolyn Dickens, and Amer Ardati. 2018. Towards generating personalized hospitalization summaries. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 74–82, New Orleans, Louisiana, USA. Association for Computational Linguistics. <https://doi.org/10.18653/v1/N18-4011>

Griffin Adams, Emily Alsentzer, Mert Ketenci, Jason Zucker, and Noémie Elhadad. 2021. What’s in a summary? Laying the groundwork for advances in hospital-course summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4794–4811, Online. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2021.naacl-main.382>, PubMed: 34179900

Griffin Adams, Han-Chin Shing, Qing Sun, Christopher Winestock, Kathleen McKeown, and Noémie Elhadad. 2022. Learning to revise references for faithful summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4009–4027, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2022.findings-emnlp.296>

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu

- Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2022. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.
- Jordan Boyd-Graber and Benjamin Börschinger. 2020. What question answering can learn from trivia nerds. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7422–7435, Online. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.acl-main.662>
- 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. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.
- Pengshan Cai, Fei Liu, Adarsha Bajracharya, Joe Sills, Alok Kapoor, Weisong Liu, Dan Berlowitz, David Levy, Richeek Pradhan, and Hong Yu. 2022a. Generation of patient after-visit summaries to support physicians. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6234–6247, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- Pengshan Cai, Hui Wan, Fei Liu, Mo Yu, Hong Yu, and Sachindra Joshi. 2022b. Learning as conversation: Dialogue systems reinforced for information acquisition. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4781–4796, Seattle, United States. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2022.naacl-main.352>
- J. Harry Caulfield, Yichao Zhou, Yunsheng Bai, David A. Liem, Anders O. Garlid, Kai-Wei Chang, Yizhou Sun, Peipei Ping, and Wei Wang. 2019. A comprehensive typing system for information extraction from clinical narratives. *medRxiv*, pages 19009118. <https://doi.org/10.1101/19009118>
- Yllias Chali and Sadid A. Hasan. 2015. Towards topic-to-question generation. *Computational Linguistics*, 41(1):1–20. https://doi.org/10.1162/COLI_a_00206
- Haw-Shiuan Chang, Zonghai Yao, Alolika Gon, Hong Yu, and Andrew McCallum. 2023. Revisiting the architectures like pointer networks to efficiently improve the next word distribution, summarization factuality, and beyond. *arXiv preprint arXiv:2305.12289*. <https://doi.org/10.18653/v1/2023.findings-acl.805>
- Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayanan Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff

- Dean, Slav Petrov, and Noah Fiedel. 2022. PaLM: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*. <https://doi.org/10.48550/arXiv.2204.02311>
- 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. <http://doi.org/10.18653/v1/N19-1423>
- Barbara Di Eugenio, Andrew Boyd, Camillo Lugaresi, Abhinaya Balasubramanian, Gail Keenan, Mike Burton, Tamara Goncalves Rezende Macieira, Jianrong Li, Yves Lussier, and Yves Lussier. 2014. PatientNarr: Towards generating patient-centric summaries of hospital stays. In *Proceedings of the 8th International Natural Language Generation Conference (INLG)*, pages 6–10, Philadelphia, Pennsylvania, U.S.A. Association for Computational Linguistics. <https://doi.org/10.3115/v1/W14-4402>
- Xinya Du and Claire Cardie. 2017. Identifying where to focus in reading comprehension for neural question generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2067–2073, Copenhagen, Denmark. Association for Computational Linguistics. <https://doi.org/10.18653/v1/D17-1219>
- 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. <https://doi.org/10.18653/v1/D17-1090>
- Liam Dugan, Eleni Miltsakaki, Shriyash Upadhyay, Etan Ginsberg, Hannah Gonzalez, DaHyeon Choi, Chuning Yuan, and Chris Callison-Burch. 2022. A feasibility study of answer-agnostic question generation for education. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1919–1926, Dublin, Ireland. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2022.findings-acl.151>
- Maksim Ereemeev, Ilya Valmianski, Xavier Amatriain, and Anitha Kannan. 2023. Injecting knowledge into language generation: A case study in auto-charting after-visit care instructions from medical dialogue. *arXiv preprint arXiv:2306.03652*. <https://doi.org/10.18653/v1/2023.acl-long.133>
- Alexander Fabbri, Patrick Ng, Zhiguo Wang, Ramesh Nallapati, and Bing Xiang. 2020. Template-based question generation from retrieved sentences for improved unsupervised question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4508–4513, Online. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.acl-main.413>
- Alex Federman, Erin Sarzynski, Cindy Brach, Paul Francaviglia, Jessica Jacques, Lina Jandorf, Angela Sanchez Munoz, Michael Wolf, and Joseph Kannry. 2018. Challenges optimizing the after visit summary. *International Journal of Medical Informatics*, 120:14–19. <https://doi.org/10.1016/j.ijmedinf.2018.09.009>, PubMed: 30409339
- Roberta Michnick Golinkoff, Erika Hoff, Meredith L. Rowe, Catherine S. Tamis-LeMonda, and Kathy Hirsh-Pasek. 2019. Language matters: Denying the existence of the 30-million-word gap has serious consequences. *Child Development*, 90(3):985–992. <https://doi.org/10.1111/cdev.13128>, PubMed: 30102419
- Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2020. Domain-specific language model pre-training for biomedical natural language processing. <https://doi.org/10.1145/3458754>
- Vince Hartman and Thomas R. Champion. 2022. A day-to-day approach for automating the hospital course section of the discharge summary. *AMIA Annual Symposium Proceedings*, 2022:216–225.

- 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.
- Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, Hong Kong, China. Association for Computational Linguistics. <https://doi.org/10.18653/v1/D19-1259>
- Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016a. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3(1):1–9. <https://doi.org/10.1038/sdata.2016.35>, PubMed: 27219127
- Roy P. C. Kessels. 2003. Patients’ memory for medical information. *Journal of the Royal Society of Medicine*, 96(5):219–222. <https://doi.org/10.1177/014107680309600504>, PubMed: 12724430
- Yanghoon Kim, Hwanhee Lee, Joongbo Shin, and Kyomin Jung. 2018. Improving neural question generation using answer separation. *CoRR*, abs/1809.02393.
- Sunjae Kwon, Zonghai Yao, Harmon S. Jordan, David A. Levy, Brian Corner, and Hong Yu. 2022. Medjex: A medical jargon extraction model with wiki’s hyperlink span and contextualized masked language model score. *arXiv preprint arXiv:2210.05875*. <https://doi.org/10.18653/v1/2022.emnlp-main.805>, PubMed: 37103473
- Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240. <https://doi.org/10.1093/bioinformatics/btz682>, PubMed: 31501885
- Eric Lehman, Vladislav Lialin, Katelyn Edelwina Legaspi, Anne Janelle Sy, Patricia Therese Pile, Nicole Rose Alberto, Richard Raymund Ragasa, Corinna Victoria Puyat, Marianne Katharina Taliño, Isabelle Rose Alberto, Pia Gabrielle Alfonso, Dana Moukheiber, Byron Wallace, Anna Rumshisky, Jennifer Liang, Preethi Raghavan, Leo Anthony Celi, and Peter Szolovits. 2022. Learning to ask like a physician. In *Proceedings of the 4th Clinical Natural Language Processing Workshop*, pages 74–86, Seattle, WA. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2022.clinicalnlp-1.8>
- Rosemary Lever and Monique Sénéchal. 2011. Discussing stories: On how a dialogic reading intervention improves kindergartners’ oral narrative construction. *Journal of Experimental Child Psychology*, 108(1):1–24. <https://doi.org/10.1016/j.jecp.2010.07.002>, PubMed: 20828708
- Patrick Lewis, Myle Ott, Jingfei Du, and Veselin Stoyanov. 2020. Pretrained language models for biomedical and clinical tasks: Understanding and extending the state-of-the-art. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 146–157. <https://doi.org/10.18653/v1/2020.clinicalnlp-1.17>
- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. Gpteval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.
- Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V. Le, Barret Zoph, Jason Wei, and Adam Roberts. 2023. The flan collection: Designing data and methods for effective instruction tuning. <https://doi.org/10.48550/arXiv.2301.13688>
- Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for

- Computational Linguistics. <https://doi.org/10.18653/v1/2020.acl-main.173>
- Suzanne E. Mol, Adriana G. Bus, Maria T. De Jong, and Daisy J. H. Smeets. 2008. Added value of dialogic parent–child book readings: A meta-analysis. *Early Education and Development*, 19(1):7–26. <https://doi.org/10.1080/10409280701838603>
- OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*. <https://doi.org/10.48550/arXiv.2203.02155>
- Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4812–4829, Online. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2021.naacl-main.383>
- Anusri Pampari, Preethi Raghavan, Jennifer Liang, and Jian Peng. 2018. emrQA: A large corpus for question answering on electronic medical records. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2357–2368, Brussels, Belgium. Association for Computational Linguistics. <https://doi.org/10.18653/v1/D18-1258>
- Stacey Pylman and Amy Ward. 2020. 12 tips for effective questioning in medical education. *Medical Teacher*, 42(12):1330–1336. <https://doi.org/10.1080/0142159X.2020.1749583>, PubMed: 32297833
- Dragomir R. Radev, Hong Qi, Harris Wu, and Weiguo Fan. 2002. Evaluating web-based question answering systems. In *LREC*. Citeseer.
- 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. <https://doi.org/10.18653/v1/2021.bionlp-1.7>
- Bhanu Pratap Singh Rawat, Wei-Hung Weng, So Yeon Min, Preethi Raghavan, and Peter Szolovits. 2020. Entity-enriched neural models for clinical question answering. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, pages 112–122, Online. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.bionlp-1.12>
- Claude Richard, Emma Glaser, and Marie-Thérèse Lussier. 2017. Communication and patient participation influencing patient recall of treatment discussions. *Health Expectations*, 20(4):760–770. <https://doi.org/10.1111/hex.12515>, PubMed: 27868327
- 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, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal V. 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 Févry, 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. *CoRR*, abs/2110.08207. <https://doi.org/10.48550/arXiv.2110.08207>
- Vered Shwartz, Peter West, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Unsupervised commonsense question answering with self-talk. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4615–4629, Online. Association for

- Computational Linguistics. <https://doi.org/10.18653/v1/2020.emnlp-main.373>
- Sarvesh Soni and Kirk Roberts. 2020. Evaluation of dataset selection for pre-training and fine-tuning transformer language models for clinical question answering. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5532–5538, Marseille, France. European Language Resources Association.
- Md Arafat Sultan, Shubham Chandel, Ramón Fernandez Astudillo, and Vittorio Castelli. 2020. On the importance of diversity in question generation for QA. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5651–5656, Online. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.acl-main.500>
- Mujeen Sung, Jinhyuk Lee, Sean Yi, Minji Jeon, Sungdong Kim, and Jaewoo Kang. 2021. Can language models be biomedical knowledge bases. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. <https://doi.org/10.18653/v1/2021.emnlp-main.388>
- Junda Wang, Zonghai Yao, Avijit Mitra, Samuel Osebe, Zhichao Yang, and Hong Yu. 2023a. UMASS_BioNLP at MEDIQA-chat 2023: Can LLMs generate high-quality synthetic note-oriented doctor-patient conversations? In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 460–471, Toronto, Canada. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.clinicalnlp-1.49>
- Yiming Wang, Zhuosheng Zhang, and Rui Wang. 2023b. Element-aware summarization with large language models: Expert-aligned evaluation and chain-of-thought method. *arXiv preprint arXiv:2305.13412*. <https://doi.org/10.18653/v1/2023.acl-long.482>
- Himali Weerahandi, Boback Ziaeeian, Robert L. Fogerty, Grace Y. Jenq, and Leora I. Horwitz. 2018. Predictors for patients understanding reason for hospitalization. *PLoS One*, 13(4):e0196479. <https://doi.org/10.1371/journal.pone.0196479>, PubMed: 29702676
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*. <https://doi.org/10.48550/arXiv.2201.11903>
- Grover J. Whitehurst. 2002. Dialogic reading: An effective way to read aloud with young children. <https://www.readingrockets.org/article/dialogic-reading-effective-way-read-aloud-young-children>.
- Ying Xu, Dakuo Wang, Mo Yu, Daniel Ritchie, Bingsheng Yao, Tongshuang Wu, Zheng Zhang, Toby Li, Nora Bradford, Branda Sun, Tran Hoang, Yisi Sang, Yufang Hou, Xiaojuan Ma, Diyi Yang, Nanyun Peng, Zhou Yu, and Mark Warschauer. 2022. Fantastic questions and where to find them: FairytaleQA – an authentic dataset for narrative comprehension. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 447–460, Dublin, Ireland. Association for Computational Linguistics. <http://doi.org/10.18653/v1/2022.acl-long.34>
- Bingsheng Yao, Dakuo Wang, Tongshuang Wu, Zheng Zhang, Toby Li, Mo Yu, and Ying Xu. 2022a. It is AI’s turn to ask humans a question: Question-answer pair generation for children’s story books. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 731–744, Dublin, Ireland. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2022.acl-long.54>
- Zonghai Yao, Yi Cao, Zhichao Yang, Vijeta Deshpande, and Hong Yu. 2022b. Extracting biomedical factual knowledge using pretrained language model and electronic health record context. *arXiv preprint arXiv:2209.07859*.
- Zonghai Yao, Yi Cao, Zhichao Yang, and Hong Yu. 2022c. Context variance evaluation of pretrained language models for prompt-based biomedical knowledge probing. *arXiv preprint arXiv:2211.10265*.
- Xiang Yue, Bernal Jimenez Gutierrez, and Huan Sun. 2020a. 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. <https://doi.org/10.18653/v1/2020.acl-main.410>
- Xiang Yue, Xinliang Frederick Zhang, Ziyu Yao, Simon Lin, and Huan Sun. 2020b. Cliniqg4qa: Generating diverse questions for domain adaptation of clinical question answering. *arXiv preprint arXiv:2010.16021*. <https://doi.org/10.1109/BIBM52615.2021.9669300>
- Yizhe Zhang, Siqu Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT: Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.acl-demos.30>
- Jane Y. Zhao, Buer Song, Edwin Anand, Diane Schwartz, Mandip Panesar, Gretchen P. Jackson, and Peter L. Elkin. 2017. Barriers, facilitators, and solutions to optimal patient portal and personal health record use: A systematic review of the literature. In *AMIA Annual Symposium Proceedings*, volume 2017, page 1913. American Medical Informatics Association.