Towards conversational assistants for health applications: using ChatGPT to generate conversations about heart failure

Anonymous ACL submission

Abstract

We explore the potential of ChatGPT (3.5-turbo and 4) to generate realistic dialogues centered on self-care strategies for African-American heart failure patients, a domain with limited specialized datasets. To achieve this, we implemented four progressively refined prompting strategies to simulate patient-health educator conversations. The conversations were generated for 3 key domains of food, exercise, and fluid intake, which are essential for heart failure self-care. Key evaluation metrics included the conversational quality of the dialogues, the empathy conveyed in the conversation, the health educator's consideration of patient-specific factors, and the impact of incorporating reasoning on the overall quality of the interactions.

1 Introduction

Heart failure, often referred to as congestive heart failure, occurs when the heart is unable to pump sufficient blood to meet the body's needs. Effective management of heart failure requires comprehensive self-care practices, including controlled salt intake, maintaining proper hydration, regular exercise, adherence to prescribed medications, and routine medical check-ups (Savarese and Lund, 2017). However, in the United States, individuals from minority communities (African American and Hispanic/Latino) often face worse outcomes and higher mortality rates than the general Caucasian population due to factors like genetic variations, healthcare access disparities, socioeconomic conditions, and lower health literacy levels (Nayak et al., 2020). Additionally, most self-care materials are designed for a white, educated demographic, lacking the cultural nuances needed to support minority patients (Barrett et al., 2019). This gap in resources contributes to poor self-care regimens in these communities. Adhering to proper medication, monitoring symptoms, regular exercise, and regulating diet/fluid intake can improve the quality of life and life expectancy for heart failure patients (Dickson and Riegel, 2009). By providing patients with personalized education, their understanding of self-care requirements can be enhanced, leading to better health outcomes and reducing readmissions (Di Eugenio et al., 2019). Our goal is to develop a conversational dialogue agent to assist African-American heart failure patients for them to ask questions about different self-care patients' needs. The dialog system will cater to the problems these patients might face - related to the disease, food, and other questions related to exercise and medicines. While previous systems like CARDIAC (Ferguson et al., 2009) have been designed to collect patient information and MyHeart (Alnosayan et al., 2014) and DIL (Moulik, 2019) have assisted in transitioning patients from hospital to home, there is currently no culturally sensitive dialog agent specifically designed to support minority communities in managing their heart failure self-care.

A significant challenge is the lack of real-world patient-language data from underrepresented communities. High-quality data is often costly to obtain, especially when no existing corpora are available. To address this, we collected a real-world dataset by recruiting three health educators to provide heart failure education to African American (AA) and Hispanic/Latino (H/L) patients. A total of 20 interactions were recorded, involving 18 AA patients and 2 H/L patients (Reference withheld). Initial analysis revealed that educators dominated the conversations, with patients contributing less. The key topics discussed during these sessions included exercise, fluid intake, symptom management, sleep, weight management, familial aspects, and salt intake.

Unlike the recorded interactions, where educators primarily asked the questions, the dialogue system being developed in this study is designed to encourage patient-initiated questions. However,

due to the lack of real-world data featuring patient-driven queries, we generated a synthetic dataset using ChatGPT. This synthetic data will supplement the real-world interactions to support the development of a patient-driven dialogue system. There is no conversational data and, more importantly, no data that is personalized for SDOH features. With the advancement of LLMs, we thought to examine different prompting strategies and evaluate whether they could be used for creating conversational datasets. Data augmentation (Yoo et al., 2021) is a promising strategy to address the pervasive issue of data scarcity in clinical natural language processing (NLP) across various medical domains. The advanced capabilities of state-of-the-art large language models (LMs) to generate coherent text provide new opportunities for synthetic text generation. Dataset is publicly available.[1]In this paper, we will look into various prompting techniques used to create an augmentation dataset. We used 4 different prompts to generate conversation.

- We started with generating simulated conversations based on different domains required for self-care of heart failure patients of food, exercise, and fluid intake.
- We introduced an additional prompt where the patients communicate using African American Vernacular English (AAVE) while the educator communicates in standard English.
- We prompted to integrate SDOH Features of the patients. We considered gender {male, female}, age {young, mid-age, old}, neighborhood {safe, unsafe} and socio-economic conditions {below poverty line, well to do}.
- We first prompted ChatGPT to generate reasoning given the SDOH features of the patient and then prompted ChatGPT again to generate conversations given the reasoning and SDOH features.

In summary, our main goal was to explore the potential of ChatGPT in generating simulated conversations when framed within the context of self-care for African American heart failure patients. Additionally, we aimed to determine whether ChatGPT could be used to create personalized dialogues based on individual patient characteristics and whether the quality of these conversations improves when using reasoning. Specifically, we focused on addressing the following key questions:

- Can ChatGPT generate relevant conversations for Heart Failure Self-Care?
- Can ChatGPT personalize conversations based on the Social Determinants of Health (SDOH) features of the patients?
- Can ChatGPT express empathy with the patients?
- Is having ChatGPT to generate reasoning before conversations more effective than directly generating the conversations?

2 Related Work

Health Education. Linguistic and cultural barriers can significantly impact patients' access to healthcare. As noted in (Handtke et al., 2019), language differences and varying health beliefs often prevent linguistically diverse patients from effectively engaging with healthcare services.

To overcome these challenges, approaches have been developed to improve patient education. The authors in (Mendu S, 2018) designed an interactive virtual patient educator to counsel Hispanic women about cervical cancer and human papillomavirus (HPV). Similarly, PaniniQA (Cai et al., 2023) helps patients understand discharge instructions through a question-answering system. One of the first and best well-known systems that provided information to patients, albeit as a summarizer, not as a dialogue system, is BabyTalk (Portet et al., 2009), which provided personalized summaries of neonatal intensive care data for their parents (and for healthcare providers as well).

Additionally, natural language processing (NLP) is being leveraged to create diabetes self-care corpus (Cunha et al., 2024), demonstrating the potential of AI and language technologies to enhance patient communication and health management.

Prompting Recent advancements in large language models (LLMs) have been driven by scaling up both model size and training data, resulting in improved performance and sample efficiency (Hoffmann et al., 2024; Brown et al., 2020). Researchers have explored various prompting techniques to enhance LLM capabilities, starting with few-shot prompting (Brown et al., 2020), followed by

[1] https://anonymous.4open.science/r/HF-Dataset

more advanced methods such as chain-of-thought prompting (Wei et al., 2022) and chain-of-thought with self-consistency (Wang et al., 2023). To address the remaining challenges, new approaches like tree-of-thought prompting have been introduced (Yao et al., 2023a), where each "thought" is a coherent language sequence representing an intermediate step toward problem-solving.

Reasoning is a crucial capability for complex problem-solving. A comprehensive overview of reasoning strategies in LLMs is provided in (Qiao et al., 2023), covering commonsense reasoning (Liu et al., 2022), mathematical reasoning (Wang et al., 2017), and symbolic reasoning (Khot et al., 2023). The ReACT framework (Yao et al., 2023b) further integrates reasoning with action in a unified task. This reasoning ability is especially critical in healthcare contexts, where accurate, informed decision-making is essential.

With the release of large-scale medical dialogue datasets, e.g., MedDialog (Zeng et al., 2020), MedDG (Xu et al., 2023), medical dialogue response generation attracts increasing attention. (Li et al., 2023) undertakes the task of enhancing and fine-tuning the LLaMa model with a dataset of approximately 100,000 patient-doctor dialogues. In (Guevara et al., 2024), the authors used LLM to extract SDOH features of housing, employment, transportation, parental status, relationship, and emotional support from the EHR data.

3 Methodology

3.1 Can ChatGPT generate relevant conversations for Heart Failure Self-Care?

By generating relevant conversations, we mean that ChatGPT should facilitate dynamic, two-way interactions between the patient and the health educator (HE) rather than a simple question-answer format where only the patient asks questions and the HE responds. In this two-way conversation, the HE should also ask follow-up questions to gather more information or to clarify the patient's condition (Walker and Whittaker, 1990). Furthermore, the conversation should not only provide answers to the patient's inquiries but also offer actionable and easily implementable advice, empowering the patient to manage their heart failure effectively.

3.2 Can ChatGPT personalize conversations based on the SDOH features of the patients?

SDOH are defined by the World Health Organization as "the conditions in which people are born, grow, live, work, and age...shaped by the distribution of money, power, and resources at global, national, and local levels" (Marmot and Wilkinson, 2005). These factors significantly influence health outcomes by affecting access to and the quality of medical care, playing a major role in health disparities. We examined whether ChatGPT can adjust its dialogue generation based on individual patient characteristics, focusing on four key social determinants of health (SDOH): age, gender, neighborhood, and socio-economic conditions. By varying these patient features, we generated conversations that reflect how these SDOH factors can significantly influence the dialogue.

3.3 Can ChatGPT express empathy with the patients?

Effective healthcare communication demands both factual accuracy and genuine concern for patients. We explored whether ChatGPT can recognize appropriate moments to express empathy during conversations.

3.4 Is having ChatGPT to generate reasoning before conversations more effective than directly generating the conversations?

Research, including (Yao et al., 2023b), has demonstrated that reasoning is not an innate capability of ChatGPT, and incorporating reasoning improves performance. Following this approach, we first generated the reasoning based on the patient's SDOH features and then, in the next step, generated the conversation using both the reasoning and the SDOH features, as detailed in section 3.5.4.

3.5 Approaches to Generate Simulated Conversations

In the absence of a specialized dataset focused on self-care strategies for African-American heart failure patients, we employed four progressively refined prompting strategies to generate simulated conversations, but before that, let's look into some additional instructions given in the prompt.

In our initial prompts to ChatGPT, even though we specified it to be a health educator, the model still advised the patient to consult a doctor. To address this, we added a specific instruction to avoid

phrases related to consulting a healthcare provider, as shown in Figure 2. Furthermore, in our earlier attempts, there was no defined structure for the generated conversations. The lack of a standardized format made it difficult to analyze and compare the conversations effectively. We anticipated that specifying a consistent format for the dialogue would lead to clearer and more organized outputs from the model. To achieve this, we introduced a specific conversation structure, where each exchange between the participants follows a predefined format: [speaker][utterance]. This structure explicitly separates the speaker from their utterance.

Recognizing that heart failure patients face significant challenges managing their health, we emphasized the need for empathetic communication from patient educators. Since many patients have lower health literacy compared to their general literacy (Paasche-Orlow et al., 2005), we instructed the model to present medical information in simple terms. Additionally, to accommodate different conversational styles, we experimented with various dialogue lengths, as some patients might prefer shorter exchanges while others engage in longer conversations. To maintain clarity and avoid overwhelming the patient, we also prompted the model to limit each educator's turn to 20 words, using plain English throughout (Figure 5). These adjustments were aimed at creating more realistic dialogues.

3.5.1 Approach 1- Generating Conversations According to Domain

Heart failure patients must manage various aspects of their health, including exercise, diet, water intake, medications, lifestyle adjustments, sleep, and overall self-care (Gupta et al., 2020). To explore how these areas could be addressed in patient-health educator conversations, we prompted ChatGPT to generate dialogues centered around different domains of heart failure care. By focusing on these domains, we aimed to generate conversations that could offer patients practical advice on how to manage these areas while allowing us to study the model's ability to generate relevant and domain-specific dialogue.

Heart failure patients are primarily advised to limit their salt intake rather than focusing on reducing calories or sugar (Gupta et al., 2020; Tayal et al., 2024). Additionally, they are encouraged to restrict their fluid intake, with no recommendation to increase water consumption. This contrasts with earlier advice given by ChatGPT, which suggested reducing sugar, unhealthy fats, and extra calories while also recommending an increase in water intake, as shown in Figure 1 along with limiting salt and alcohol intake.

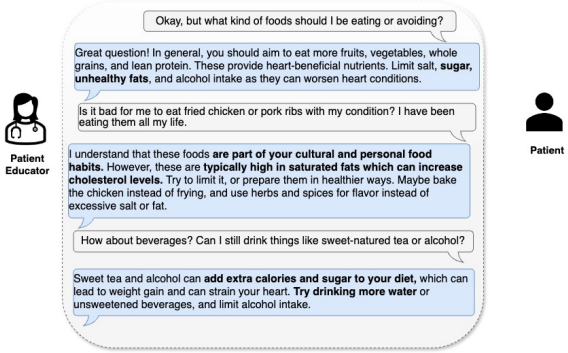


Figure 1: Excerpt of conversation where the patient is advised to limit calories and sugar along with salt intake and advised to drink more water instead of less.

3.5.2 Approach 2 - African American Vernacular English

One way to introduce cultural compatibility is through language. African American Vernacular English (AAVE) (Green, 2002), is a dialect of Standard American English (SAE) with defined syntactic-semantic, phonological and lexical features. Since African American patients may speak AAVE (it is important to note that not all AA's speak AAVE), we prompted ChatGPT so that the educator would speak in standard English while the patient communicated in AAVE (Figure 2). However, instead of keeping the educator's speech in standard English, we noticed instances where the educator also used AAVE, as seen in Figure 3.

You are a healthcare specialist focusing on heart failure. Your purpose is to answer heart failure patient questions and ask follow-up questions.
You should empathetically communicate medical information in a simple manner.
Simulate {lines} round conversation between African American heart failure patient and healthcare specialist where the patient asks for recommendations for {domain}.
The healthcare specialist uses the **Standard English** Version with 6th grade reading level, and the patient uses **African American Vernacular English** AAVE.
Scripts should be generated in the following format: [speaker] [utterance] of a patient and healthcare specialist.
(Do not use phrases such as " consult with your healthcare provider", assume that you are the healthcare provider)

Figure 2: Prompt for generating conversations where the patient converses in AAVE

This is problematic because non-AAVE speakers attempting to use AAVE with patients may come across as offensive and disrespectful. Educators

should be aware of cultural boundaries and sensitive to the linguistic preferences of African American patients. While they should be prepared to respectfully engage if the patient uses AAVE, they should not initiate or adopt the dialect themselves. Maintaining cultural awareness is key to fostering respectful and effective communication.

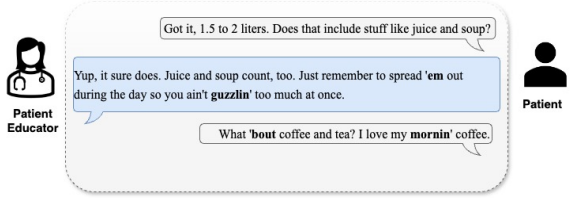


Figure 3: Excerpt of a conversation where both patient and educator are speaking AAVE

In Figure 4, it is acceptable for the patient to ask about cultural foods like watermelon, okra, and herbal teas, as this aligns with their personal interests or cultural background. When patients bring up these topics, it allows for a natural and open conversation that respects their perspective. However, if the HE had initiated this discussion, it might be perceived as introducing bias, especially if the conversation were based on assumptions about the patient's background or identity. In such cases, the HE might unintentionally appear to be stereotyping or making generalizations, which could compromise the neutrality and professionalism expected in healthcare settings.

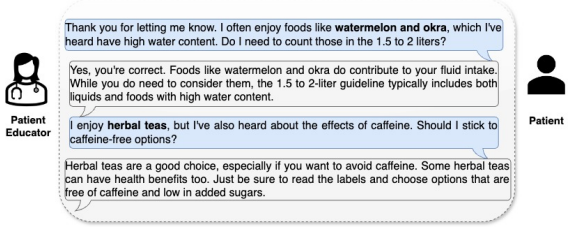


Figure 4: Excerpt of a conversation on fluid-intake where patient discusses cultural foods like watermelon, okra, herbal teas

3.5.3 Approach 3 - Integrating SDOH Features

Given the importance of SDOH, it is critical that generated conversations for healthcare applications reflect the diverse experiences of individuals based on their unique circumstances. To achieve this, our approach selectively differentiates key features such as gender {male, female}, age {young, middle age, old}, neighborhood {safe, unsafe}, and socioeconomic status {well-to-do, below poverty line} in simulated patient conversations (Figure 5). Although the list of SDOH-related factors we selected is not exhaustive, we focused on these key areas—gender, age, neighborhood, and socioeconomic conditions—because they heavily influence healthcare access and outcomes. By tailoring these dialogues to reflect specific SDOH features, we can more accurately capture the nuanced ways in which these factors influence patient-educator interactions. For example, the way a young, well-to-do

You are a healthcare educator focusing on heart failure.
Your purpose is to answer heart failure patient questions based on patient description.
You should empathetically communicate medical information in a simple manner.
Simulate {lines} round conversation between African American heart failure patient and healthcare educator where the patient asks for recommendations for {domain}.
Scripts should be generated in the following format: [speaker] [utterance] between patient and the healthcare educator.
Each educator turn should not be longer than 20 words and should use simple english.
(Do not use phrases such as " consult with your healthcare provider&quot, assume that you are the healthcare provider)

Patient Description:
gender: {gender}
socio-economic condition: {socio_economic}
neighborhood: {neighborhood}
age: {age}

Figure 5: Prompt for generating the conversation given the SDOH Features

patient discusses self-care strategies might differ significantly from the conversation of an elderly patient living in a low-income neighborhood. By adjusting these features in our simulated conversations, we aim to capture the diverse realities that patients face when managing their health.

3.5.4 Approach 4 - Generating Reasoning as an Intermediate Step

Building on the idea that reasoning enhances the capabilities of large language models (LLMs) (Wei et al., 2022; Yao et al., 2023b) and is essential for generating meaningful conversations, we introduced an intermediate step to first generate reasoning prior to generating the simulated conversation.

We approached conversation generation as a reasoning chain, using chaining (Wu et al., 2022) to divide the process into two phases. In the first phase, we prompted ChatGPT to analyze the patient's social determinants of health (SDOH) features to create logical reasoning. The reasoning process was segmented into six steps: (1) **Premise**, a tricky logical statement about how factors like socioeconomic status, cultural influences, and health-

You are a healthcare educator focusing on heart failure.

Patient Description:
Age: {age}
Gender: {gender}
Neighborhood: {neighborhood}
Socio-economic condition: {socio_economic}

Given the patient description, formulate reasoning for addressing inquiries from an African-American heart failure patient regarding {domain}.
Pay special attention to the patient's description for the reasoning.
Explain your reasoning in detail. The reason should follow this 6-line format.

Premise: <A tricky logical statement about the world, such as how socio-economic factors, cultural influences, and healthcare access impact heart failure management in African-American patients.>
Reasoning: <Break down and analyze the premise by exploring the possible scenarios, outcomes, and factors involved, while incorporating the patient's age, gender, neighborhood, and socio-economic condition into the analysis.>
Solution: <Provide a solution to the patient's inquiry that is actionable based on their description.>
Anticipate Barrier: <Identify potential barrier that may arise based on patient description>
Solve Barrier: <Propose ways to overcome the barrier>
Educate: <Educate patient about solution>

Figure 6: Prompt for generating the reasoning given the SDOH Features

care access affect heart failure management; (2) **Reasoning**, where the premise is analyzed with respect to the patient's SDOH features; (3) **Solution**, proposing an actionable plan; (4) **Anticipate Barrier**, identifying potential obstacles based on the patient's circumstances; (5) **Solve Barrier**, offering ways to overcome those obstacles; and (6) **Educate**, ensuring the educator not only addresses the patient's query but also provides further education.

You are a healthcare educator focusing on heart failure.
Your purpose is to answer heart failure patient questions.
You should empathetically communicate medical information in a simple manner.

Patient Description:
age: {age}
gender: {gender}
neighborhood: {neighborhood}
socio-economic condition: {socio_economic}

{reasoning}

Given the patient description and reasoning, simulate {lines} round conversation between African American heart failure patient and healthcare educator where the patient asks for recommendations for {domain}.
Scripts should be generated in the following format: [speaker] [utterance] between the patient and the healthcare educator.
Each educator's turn should not be longer than 20 words and should use simple English.
(Do not use phrases such as " consult with your healthcare provider&quot, assume that you are the healthcare provider)

Figure 7: Prompt for generating the conversation given the reasoning and the SDOH Features

This segmented approach, shown in Figure 6 and adapted from (Gupta et al., 2021), is designed for ChatGPT to focus on analyzing the patient's needs based on their SDOH features. For example, reasoning generated for a {gender: female, age: elderly, neighborhood: unsafe, socio-economic condition: below poverty line } in the exercise domain is shown in Figure 8a. It highlights appropriate

Premise: Socio-economic factors, cultural influences, and healthcare access impact heart failure management in African-American patients, leading to disparities in disease outcomes.

Reasoning: Considering the patient's age, gender, living in an unsafe neighborhood, and being below the poverty line, it is crucial to address the barriers she may face in accessing **safe and affordable exercise options.** Lack of resources, fear of safety, and limited access to healthcare facilities may hinder her ability to engage in physical activity for heart failure management.

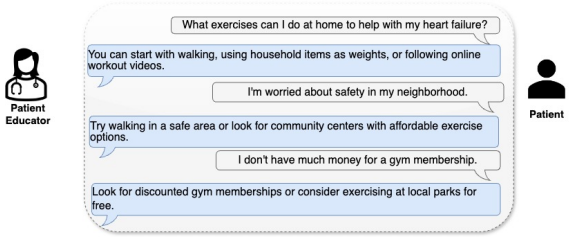
Solution: Encourage the patient to start with simple, **low-cost exercises** at home, such as walking in a safe area, **using household items as weights, or following online workout videos.** Emphasize the importance of consistency and starting slow to avoid injury.

Anticipate Barrier: The patient may **struggle to find a safe and affordable environment to exercise**, leading to inconsistency in her physical activity routine.

Solve Barrier: Provide resources **for community centers, local parks, or discounted gym memberships** that offer safe and affordable exercise options. **Encourage the patient to enlist the support of family members or friends to exercise together for added safety.**

Educate: Educate the patient on the importance of regular physical activity in managing heart failure, the benefits of exercise in improving heart function and overall health, and the impact of socio-economic factors on health disparities. Emphasize the need to prioritize her health and well-being despite the challenges she may face.

(a) A reasoning generated for patient {gender: female, age: elderly, neighborhood: unsafe, socio-economic condition: below poverty line } in the exercise domain



(b) Corresponding Conversation

Figure 8: Generated reasoning with Prompt 6 and conversation with Prompt 7

reasoning, such as suggesting low-cost exercises at home and addressing the difficulty of finding a safe and affordable environment by proposing community centers or local parks as alternatives.

In the second phase, we instructed the model to generate the conversation based on the generated reasoning. This ensured the dialogue aligned with the patient's SDOH features, resulting in more personalized and contextually appropriate conversations, as shown in Figure 8b. This two-step process improved the quality of the guidance offered by making it more relevant to the patient's specific needs, grounding the model's reasoning in factors like socio-economic conditions, cultural influences, and healthcare access.

4 Experiment and Evaluation

ChatGPT (versions 3.5-turbo and 4) was utilized to generate simulated conversations using four distinct approaches of Domain, AAVE, SDOH, and Reasoning. For each approach, conversations were created with varying numbers of conversation rounds {5, 10, 15} and across different domains {food, water, exercise}.

Additionally, in the SDOH and Reasoning ap-

proaches (Approaches 3 and 4), conversations were further varied by incorporating different SDOH features, such as gender {male, female}, age {young, mid-age, old}, neighborhood {safe, unsafe}, and socio-economic conditions {below the poverty line, well-to-do}. For the Domain and AAVE approaches (Approaches 1 and 2), conversations were also generated with varying temperature settings {0, 0.5, 0.7, 1.0}, which influenced the model's response variability.

In the Domain approach, additional topics like medication, sleep, and self-care were included alongside food, water, and exercise. In total, 144 conversations were generated for the Domain approach, 72 for AAVE, and 432 for the SDOH and Reasoning approaches, resulting in a rich and diverse set of simulated dialogues for evaluation.

Systematic human evaluation of the generated conversations was conducted, incorporating both qualitative and quantitative assessments. For the qualitative analysis, we distributed a questionnaire to 10 colleagues from our department who were not involved in the research. Their task was to review the conversations and provide feedback based on specific questions. Each participant was assigned three conversations per prompt, with a total of 7 respondents. As a result, we evaluated 21 conversations for each prompt. Details of the questionnaire can be found in Appendix B. For Quantitative analysis, refer to Appendix A

4.1 Can ChatGPT generate relevant conversations for Heart Failure Self-Care?

To assess the relevance of the simulated conversations, we used both quantitative and qualitative metrics. One key aspect was determining whether the conversations resembled meaningful two-way interactions or were simple question-and-answer exchanges. To measure this, the ratio of follow-up questions from the health educator (HE) to those from the patient was calculated (Follow-up Ratio). As shown in Table 1, the HE did not ask follow-up questions unless explicitly instructed to do so in the prompt, which was most evident in the AAVE approach (Approach 2). This finding highlights that the ChatGPT is not aware of the 2-way nature of the conversations.

To measure the balance of participation between the HE and the patient, the ratio of words spoken by the HE to those by the patient was calculated. Table 1 reveals that the HE consistently spoke more than the patient. However, when we limited responses to 20 words in approach 3 and 4, the answers became overly brief, offering little explanation or depth. One evaluator remarked that these responses felt vague and unsatisfying, with the lack of follow-up questions leaving the conversations incomplete and uninformative.

To further assess the relevance and appropriateness of the conversations, we asked evaluators to identify instances where the HE should have provided a different answer. Across all approaches, at least 20% of the conversations contained such instances. One evaluator emphasized the need for more concrete details, noting that the educators often focused only on positive examples (e.g., what to eat) without also mentioning what to avoid. For instance, a response like "This includes water, juice, and other beverages" was criticized for being too vague, as certain beverages, such as coffee, can cause dehydration.

In one conversation about fluid intake, a patient asked if they should be concerned about drinking too much water. Instead of addressing the patient's question directly, the HE responded with information about the risks of drinking water too quickly, missing the essence of the inquiry.

4.2 Can ChatGPT personalize conversations based on the Social Determinants of Health (SDOH) features of the patients?

To evaluate whether ChatGPT has the capacity to personalize the conversations based on the SDOH features, the evaluators were asked to qualitatively assess the generated conversations, specifically examining how well ChatGPT tailored the discussions to each of these SDOH features and whether the system appropriately adapted its responses based on the patient's unique context.

When evaluating simulated conversations for SDOH features in Approach 3, for gender 2/3 conversations were deemed appropriate, while one-third were considered very appropriate. For age, the ratings were as follows: 4.8% at level 1, 14.3% at level 3, 47.6% at level 4, and 33.3% at level 5. The neighborhood-related evaluations showed a similar trend, with 4.8% at level 1, 4.8% at level 2, 9.5% at level 3, 52.4% at level 4, and 28.6% at level 5. For socioeconomic status, 4.8% were rated at level 2, 9.5% at level 3, 33.3% at level 4, and 52.4% at level 5.

These results indicate that ChatGPT has some capacity to personalize conversations based on SDOH features, particularly in the higher rating levels.

However, the presence of lower ratings shows that there is still room for improvement in making conversations more consistently aligned with SDOH features.

4.3 Can ChatGPT express empathy with the patients?

To assess whether ChatGPT can express empathy, the evaluators were asked to qualitatively judge if the HE demonstrated concern toward the patients. The level of empathy remained consistent across the four approaches, with 25% of the conversations receiving a rating of 3 or lower. When asked for feedback, one evaluator noted that the educator failed to inquire about the challenges or obstacles the patients might encounter when trying to implement the given advice. In one conversation with a patient living in an unsafe neighborhood, the educator's response not only lacked empathy but also suggested a solution that reflected little understanding of the patient's reality.

For example: "[Patient] Can you recommend exercises that are safe for me to do in my neighborhood?

[Healthcare Educator] "Walking or cycling on safe streets can be good options for you."

This response was criticized for being tone-deaf, as it ignored the patient's specific concern about the safety of their neighborhood. One evaluator described the responses as robotic, lacking emotional depth and genuine empathy. The model displayed empathy primarily in positive situations (e.g., "That's fantastic!") but failed to convey understanding in negative contexts (e.g., "That's unfortunate, I understand it's tough"). Furthermore, the educators often did not ask about the difficulties patients might face in following the advice, nor did they acknowledge the practical barriers involved. As a result, the conversations felt cold and impersonal, making the educators seem disengaged and uninterested in the patient's experiences.

4.4 Is having ChatGPT to generate reasoning before conversations more effective than directly generating the conversations?

When providing the conversations for evaluation, we ensured that the patient features remained consistent for Approaches 3 and 4. This consistency allowed us to make meaningful comparisons based on the SDOH features. Although the generated conversations varied, maintaining the same patient characteristics ensured that we could effectively assess and compare the quality and relevance of the conversations across these approaches and evaluate whether generating the reasoning prior to generating the conversation more effective.. When evaluating the SDOH features for Approach 4 conversations, the results across different patient attributes were as follows: gender {4: 87.5%, 5: 12.5%}, age {3: 25%, 4: 62.5%, 5: 25%}, neighborhood {2: 12.5%, 4: 62.5%, 5: 25%}, and socio-economic {1: 12.5%, 2: 12.5%, 4: 62.5%, 5: 12.5%}.

When comparing these results to Approach 3 results in Section 4.2, it is evident that the incorporation of reasoning significantly improved the appropriateness of the conversations. The majority of conversations in Approach 4 received higher ratings, demonstrating the model's enhanced ability to engage in nuanced, context-sensitive interactions.

Generating reasoning before the conversations proved to be a meaningful enhancement. However, there remains room for further improvement.

5 Conclusion and Future Work

This study assessed the capabilities of ChatGPT (versions 3.5-turbo and 4) in generating simulated conversations related to self-care strategies for African-American heart failure patients. Simulated conversations were generated using four distinct prompts: Domain, African American Vernacular English (AAVE), Social Determinants of Health (SDOH), and Reasoning. Our findings highlight the critical role of prompt design, revealing that while ChatGPT can incorporate SDOH features and improve dialogue quality by generating reasoning prior to the conversation, further improvements are needed. Specifically, there is a clear need to refine the conversational style to make interactions feel more engaging and empathetic, an essential element in healthcare communication.

In the future, we aim to develop a task-oriented dialogue system tailored to the self-care needs of African-American heart failure patients, utilizing these simulated conversations. Key factors such as age, gender, neighborhood, and socio-economic conditions will be integrated into the to generate the educator's response.

6 Limitations and Ethics Statement

Our research focuses on exploring the dynamics of conversations generated by ChatGPT, utilizing simulated dialogues based on carefully designed prompts. The dataset is unique to each prompt,

showcasing the adaptability of ChatGPT in generating context-specific conversations.

In the qualitative evaluation, we engaged a small group of participants to assess the conversations. While the sample size was limited, this initial user study provided valuable insights into how ChatGPT handles dialogue generation in this context. Although the evaluators were not domain experts, their feedback highlighted key areas for future refinement. Moving forward, we plan to expand the evaluation process by involving a larger and more diverse group of reviewers. Each conversation will be reviewed by at least three individuals to ensure more robust and comprehensive feedback, which will help us enhance the system's performance and reliability.

References

Nagla Alnosayan, Edward Lee, Ala Alluhaidan, Samir Chatterjee, Linda Houston-Feenstra, Mercy Kagoda, and Wayne Dysinger. 2014. Myheart: An intelligent mhealth home monitoring system supporting heart failure self-care. In *2014 IEEE 16th International Conference on e-Health Networking, Applications and Services (Healthcom)*, pages 311–316.

Matthew Barrett, Josiane Boyne, Julia Brandts, Hans-Peter Brunner-La Rocca, Lieven De Maesschalck, Kurt De Wit, Lana Dixon, Casper Eurlings, Donna Fitzsimons, Olga Golubnitschaja, et al. 2019. Artificial intelligence supported patient self-care in chronic heart failure: a paradigm shift from reactive to predictive, preventive and personalised care. *Epma Journal*, 10:445–464.

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.

Pengshan Cai, Zonghai Yao, Fei Liu, Dakuo Wang, Meghan Reilly, Huixue Zhou, Lingxi Li, Yi Cao, Alok Kapoor, Adarsha Bajracharya, Dan Berlowitz, and Hong Yu. 2023. PaniniQA: Enhancing Patient Education Through Interactive Question Answering. *Transactions of the Association for Computational Linguistics*, 11:1518–1536.

Rossana Cunha, Thiago Castro Ferreira, Adriana Pagano, and Fabio Alves. 2024. A persona-based corpus in the diabetes self-care domain - applying a human-centered approach to a low-resource context. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1353–1369, Torino, Italia. ELRA and ICCL.

B Di Eugenio, R Cameron, A Boyd, K Lopez, P Martyn-Nemeth, C Dickens, A Ardati, and D Chattopadhyay. 2019. Designing self-care technologies for hf patients: a conceptual model. In *Conference on Human Factors in Computing Systems*, pages 12–16.

Victoria Vaughan Dickson and Barbara Riegel. 2009. Are we teaching what patients need to know? building skills in heart failure self-care. *Heart & Lung*, 38(3):253–261.

George Ferguson, James Allen, Lucian Galescu, Jill Quinn, and Mary Swift. 2009. Cardiac: An intelligent conversational assistant for chronic heart failure patient heath monitoring. In *2009 AAAI fall symposium series*.

Lisa J Green. 2002. *African American English: a linguistic introduction*. Cambridge University Press.

Marco Guevara, Shan Chen, Spencer Thomas, Tafadzwa L Chaunzwa, Idalid Franco, Benjamin H Kann, Shalini Moningi, Jack M Qian, Madeleine Goldstein, Susan Harper, et al. 2024. Large language models to identify social determinants of health in electronic health records. *NPJ digital medicine*, 7(1):6.

Itika Gupta, Barbara Di Eugenio, Devika Salunke, Andrew Boyd, Paula Allen-Meares, Carolyn Dickens, and Olga Garcia. 2020. Heart failure education of African American and Hispanic/Latino patients: Data collection and analysis. In *Proceedings of the First Workshop on Natural Language Processing for Medical Conversations*, pages 41–46, Online. Association for Computational Linguistics.

Itika Gupta, Barbara Di Eugenio, Brian D. Ziebart, Bing Liu, Ben S. Gerber, and Lisa K. Sharp. 2021. Summarizing behavioral change goals from SMS exchanges to support health coaches. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 276–289, Singapore and Online. Association for Computational Linguistics.

Oriana Handtke, Benjamin Schilgen, and Mike Mösko. 2019. Culturally competent healthcare – a scoping review of strategies implemented in healthcare organizations and a model of culturally competent healthcare provision. *PLOS ONE*, 14(7):1–24.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Oriol Vinyals, Jack W. Rae, and Laurent Sifre. 2024. Training compute-optimal

large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. 2023. Decomposed prompting: A modular approach for solving complex tasks. In *The Eleventh International Conference on Learning Representations*.

Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Preprint*, arXiv:2303.14070.

Jiacheng Liu, Alisa Liu, Ximing Lu, Sean Welleck, Peter West, Ronan Le Bras, Yejin Choi, and Hannaneh Hajishirzi. 2022. Generated knowledge prompting for commonsense reasoning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3154–3169, Dublin, Ireland. Association for Computational Linguistics.

Michael Marmot and Richard Wilkinson. 2005. *Social determinants of health*. Oup Oxford.

Gordon JR Mendu S, Boukhechba M. 2018. Design of a culturally-informed virtual human for educating hispanic women about cervical cancer.

Sanjoy Moulik. 2019. *DIL-A Conversational Agent for Heart Failure Patients*. Ph.D. thesis, The Claremont Graduate University.

A Nayak, AJ Hicks, and AA Morris. 2020. Understanding the complexity of heart failure risk and treatment in black patients. *Circulation: Heart Failure*, 13(8):e007264.

Michael K Paasche-Orlow, Ruth M Parker, Julie A Gazmararian, Lynn T Nielsen-Bohlman, and Rima R Rudd. 2005. The prevalence of limited health literacy. *Journal of general internal medicine*, 20(2):175–184.

François Portet, Ehud Reiter, Albert Gatt, Jim Hunter, Somayajulu Sripada, Yvonne Freer, and Cindy Sykes. 2009. Automatic generation of textual summaries from neonatal intensive care data. *Artificial Intelligence*, 173(7-8):789–816.

Shuofei Qiao, Yixin Ou, Ningyu Zhang, Xiang Chen, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, and Huajun Chen. 2023. Reasoning with language model prompting: A survey. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5368–5393, Toronto, Canada. Association for Computational Linguistics.

Gianluigi Savarese and Lars H Lund. 2017. Global public health burden of heart failure. *Cardiac failure review*, 3(1):7.

Anuja Tayal, Barbara Di Eugenio, Devika Salunke, Andrew D. Boyd, Carolyn A. Dickens, Eulalia P. Abril, Olga Garcia-Bedoya, and Paula G. Allen-Meares. 2024. A neuro-symbolic approach to monitoring salt content in food. In *Proceedings of the First Workshop on Patient-Oriented Language Processing (CL4Health) @ LREC-COLING 2024*, pages 93–103, Torino, Italia. ELRA and ICCL.

Marilyn Walker and Steve Whittaker. 1990. Mixed initiative in dialogue: An investigation into discourse segmentation. In *28th Annual Meeting of the Association for Computational Linguistics*, pages 70–78, Pittsburgh, Pennsylvania, USA. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Yan Wang, Xiaojiang Liu, and Shuming Shi. 2017. Deep neural solver for math word problems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 845–854, Copenhagen, Denmark. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Tongshuang Wu, Michael Terry, and Carrie Jun Cai. 2022. Ai chains: Transparent and controllable human-ai interaction by chaining large language model prompts. In *Proceedings of the 2022 CHI conference on human factors in computing systems*, pages 1–22.

Kaishuai Xu, Wenjun Hou, Yi Cheng, Jian Wang, and Wenjie Li. 2023. Medical dialogue generation via dual flow modeling. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6771–6784, Toronto, Canada. Association for Computational Linguistics.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik R Narasimhan. 2023a. Tree of thoughts: Deliberate problem solving with large language models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2023b. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*.

Kang Min Yoo, Dongju Park, Jaewook Kang, Sang-Woo Lee, and Woomyoung Park. 2021. GPT3Mix: Leveraging large-scale language models for text augmen-

tation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2225–2239, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang, Hongchao Fang, Penghui Zhu, Shu Chen, and Pengtao Xie. 2020. MedDialog: Large-scale medical dialogue datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9241–9250, Online. Association for Computational Linguistics.

A Quantitative Analysis

For the quantitative analysis, we assessed the instructions provided in the prompts, focusing on the number of conversation rounds and adherence to the correct format $[speaker][utterance]$. We calculated the percentage of conversations that followed the correct number of rounds (Round Adherence Ratio), as well as the ratio of conversations that adhered to the required format (Format Adherence Ratio). The results, presented in Table 1, show that both the models struggled to consistently follow even simple instructions.

	GPT3.5-turbo				GPT4			
	Domain	AAVE	SDOH	Reasoning	Domain	AAVE	SDOH	Reasoning
Round Adherence Rate	0.02	0.083	0.83	0.26	0	0.04	0.64	0.35
Follow-up ratio	0.4	0.52	0.003	0.007	0.01	0.85	0.003	0.02
Ratio of words	2.8	1.5	1.4	1.32	3.4	2.68	1.83	1.8
Format Adherence Rate	0	0.75	0.93	0.96	1.0	0	0	0.23

Table 1: Quantitave Results table

B Qualitative Questionnaire

- **General Questionnaire**
  - Was the health educator able to answer the patient's questions?
  - Was the advice given by the health educator actionable/ could easily be implemented by the patient?
  - Did the HE show concern toward patients? (If likert scale <=2, Why not)
  - Is there any question for which you think the health educator should give a different answer? (If likert scale <=2, Which instance)
- **SDOH Questions**
  - How appropriate was the conversation for each individual feature
- **Reasoning Questions**
  - Was the reasoning generated appropriate given the patient's sdoh features ?
  - Was the response generated according to the reasoning generated?