Conversational Assistants to support Heart Failure Patients: comparing a Neurosymbolic Architecture with ChatGPT

Anonymous ACL submission

Abstract

Conversational assistants are becoming more and more popular, including in healthcare, partly because of the availability and capabilities of Large Language Models. There is a need for controlled, probing evaluations with real stakeholders which can highlight advantages and disadvantages of more traditional architectures and those based on generative AI. We present a within-group user study to compare two versions of a conversational assistant that allows heart failure patients to ask about salt content in food. One version of the system was developed in-house with a neurosymbolic architecture, and one is based on ChatGPT. The evaluation shows that the in-house system is more accurate, completes more tasks and is less verbose than the one based on ChatGPT; on the other hand, the one based on ChatGPT makes fewer speech errors and requires fewer clarifications to complete the task. Patients show no preference for one over the other.

1 Introduction

Conversational assistants in the healthcare domain are as old as Natural Language Processing (NLP), since in 1966 ELIZA was already playing the role of a psychiatrist (Weizenbaum, 1966). They have proliferated in more recent years with the availability of datasets and machine learning approaches, even before the rise of Large Language Models (LLMs). The comprehensive survey in (Valizadeh and Parde, 2022) provides an in-depth analysis of these diverse healthcare-oriented dialogue systems, examining them from a computational perspective and highlighting their varied end-users.

Traditional task-oriented dialog systems are typically assessed using metrics such as Slot Filling F1-Score and Intent Recognition Accuracy for Natural Language Understanding (NLU), as well as Slot Accuracy and Joint Goal Accuracy for Dialog State Tracking (DST) (Budzianowski et al., 2018). However, task performance is only one aspect of dialogue system evaluation, as already demonstrated by Paradise (Walker et al., 1998), a framework that links user satisfaction to task success and interaction costs.

Evaluation of dialogue systems based on LLMs raises additional concerns: LLMs do not operate within rigid task boundaries, making it difficult to apply standard task-oriented evaluation metrics. Furthermore, LLMs often lack transparency regarding their data sources and may fail to reliably follow user prompts, raising significant concerns in high-stakes domains where accuracy and accountability are critical (Chowdhury et al., 2023). These limitations are especially critical in practical, real-world applications when facilitating medical conversations. In such cases, defining and assessing what constitutes "good" performance becomes far more complex and nuanced (Moor et al., 2023).

As a result, human evaluation remains the gold standard for assessing medical dialog systems (Chowdhury et al., 2023). Human evaluators can provide insights into subjective qualities such as coherence, informativeness, and user satisfaction—factors that are difficult to capture with automated metrics alone. We present a user study in which we compare one traditional task-oriented dialog system (ToDS) in healthcare with one based on LLMs.

The study focuses on African American patients with heart failure, aiming to assist them in managing their salt intake by providing information about the salt content in various foods. Heart failure patients must meticulously monitor and reduce their salt intake: African American individuals are more prone to heart failure (Nayak et al., 2020), have a higher sensitivity to salt, and face challenges like lack of access to healthy foods . Furthermore, in (Gupta et al., 2020), the authors show that African American patients with heart failure often discuss salt and food during heart failure educational sessions, indicating a significant interest and need for information in this area. By providing a tool that facilitates easy access to information about salt content in foods, we aim to empower patients to make healthier dietary choices, thereby addressing a critical aspect of managing heart failure.

The study compares two dialog systems—an in-house Neuro-Symbolic System (HFFood-NS) with a ChatGPT-based system (HFFood-GPT)—using a within-subject design. We conduct intrinsic (task performance) and extrinsic analyses (Sparck Jones and Galliers, 1995) using pre- and post-interaction surveys, to evaluate the 2 systems with African-American patients while hospitalized. By assessing how real patients, rather than typical study participants like Mechanical Turk workers or students, discuss food and prefer to receive information, we establish relevance and value of the intervention with real stakeholders.

2 Related Work

ToDS. Task-oriented Dialog Systems (ToDS.) typically follow a pipeline approach with four main components: natural language understanding (NLU) (Chen et al., 2016), dialog state tracking (DST) (Zhong et al., 2018), dialog manager (Su et al., 2016), and natural language generation (NLG) (Chen et al., 2019). The NLU module interprets user input and represents the dialog state as slots (e.g., location, price range) to be filled during the conversation. DST monitors unfilled slots to inform the dialog manager, which decides the next action. This action is then passed to the NLG module to generate the system's response.

Early systems (Young et al., 2013) used carefully designed action spaces to manage dialog states which were later replaced by neural networks (Lei et al., 2018; Peng et al., 2021). While the pipeline approach integrates domain-specific knowledge and slot-filling methods, it often requires additional human labeling.

Dialog Systems in Healthcare. ToDS have seen a significant rise in the healthcare sector (Valizadeh and Parde, 2022). These systems are developed for a wide array of diseases, including patient education (Cai et al., 2023; Gupta et al., 2020), heart failure (Moulik, 2019; Gupta et al., 2020), mental disorders (Ali et al., 2020), public anxiety (Wang et al., 2020), cancer (Belfin et al., 2019) etc. Their applications extend to several areas, including disease diagnosis (Wei et al., 2018) and health coaching (Zhou et al., 2022), among others.

Patient Centeredness. (Gupta et al., 2020; Salunke et al., 2023) highlight the development of a dialogue agent tailored to the self-care needs of heart failure patients, leveraging insights from educational sessions. (Kearns et al., 2020) investigates the use of the Wizard of Oz (WOZ) technique to create a persona-based health counseling dialogue dataset. Recent advancements have also incorporated Large Language Models (LLMs) to address patient inquiries (Chowdhury et al., 2023), with a strong emphasis on safety. To overcome the limitations in medical knowledge inherent to LLMs, (Li et al., 2023) focuses on enhancing and fine-tuning the LLaMa model using a dataset of approximately 100,000 patient-doctor dialogues. (Knoll et al., 2022) used a user-centered approach to iteratively improve their medical note generation model with user feedback conducted via semi-structured interviews. User studies are useful for collecting real-world evaluations in domains lacking specific use-case data (Walker et al., 1998; Riveiro and Thill, 2021).

3 Two dialogue system architectures

3.1 HFFood-NS

The first system, HFFood-NS, is a neuro-symbolic conversational system adapted from (Reference Withheld). As there is no conversational dataset related to food salt content, we created a template-based conversational dataset. We utilized the USFDC (U.S. Food Data Central) (USFDC, 2022) dataset, which provided detailed food descriptions along with their nutrient values. We annotated the dataset based on food slots of *food*, *cook*, *type*, *foodweight*, and *metric*. Full details of creating the dataset can be found in (Reference Withheld). Initially, we trained an end-to-end dialog system, but for the user study, we only used its DST part.

3.1.1 End-to-End dialog system

We used the end-to-end dialog system PPTOD (Plug-and-Play Task-Oriented Dialogue System) (Su et al., 2022) to train the dialog system. PPTOD, a T5-based model, excels in in-context learning by employing customized prompts for specific tasks.

Our experiments revealed that fine-tuning a transformer model to predict salt content alone is insufficient (see Table 8). This is due to two key reasons. First, large pre-trained language models (PLMs) like GPT-3 and T5 (Brown et al., 2020;

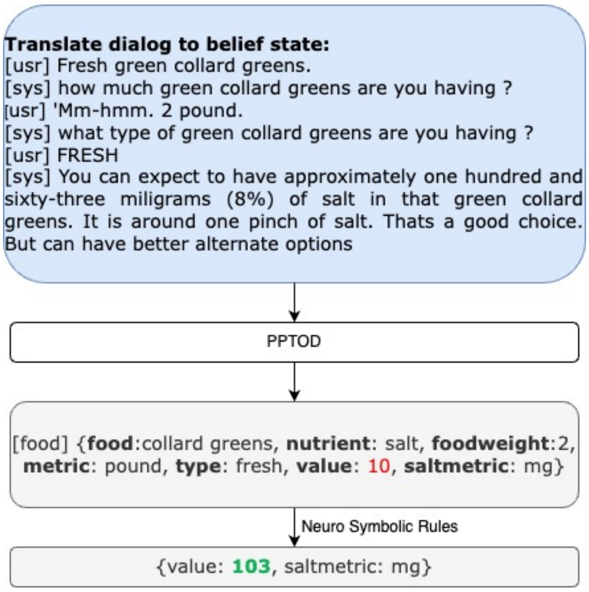


Figure 1: HFChat-NS with Interaction: As the value in DB was available for 100 gms, while the user asked about 2 pounds, it fetched the appropriate value and calculated the value for 2 pounds.

Raffel et al., 2020) occasionally make calculation errors, particularly as mathematical operations in equations grow more complex (Wei et al., 2022). Second, salt is a multi-valued slot, and its value changes over slight changes in food preparation method or quantity.

To address these challenges, we integrated the PPTOD model with neuro-symbolic rules. These rules enable the retrieval of accurate salt values from a database and perform mathematical calculations for specific food weights, allowing the system to handle non-standard food quantities effectively. This integration significantly enhances system performance, achieving a 20% improvement in joint goal accuracy across different dataset sizes (as shown in Table 7). These results demonstrate that combining pre-trained language models with neuro-symbolic rules achieves better accuracy.

3.2 Final Model- HFFood-NS

For the user-study version of HFFood-NS, several measures were implemented to modify the system to deploy in a patient-centric application. First, only the Dialogue State Tracking (DST) module of NS-PPTOD was used for simplicity and reliability. Secondly, if a slot remains unfilled, the system attempts to query the user about it up to two times. Third, the system responses were template-based responses which were created to train the model (Section 3.1.1). Fourth, since the USFDC dataset contains detailed information and users often give less precise food descriptions, the system selects the first food item that meets all minimum required keywords to account for the multi-valued nature of food items when providing salt values.

Finally, we carefully design the final system response to effectively represent salt values. We represent salt values using two formal units—milligrams (mg) and percentage (%)—commonly found on food labels, and we include an informal measure - pinches/dashes. We also compared the salt value to the recommended daily intake of 2000mg. We structure the final response to be motivational by categorizing salt content into four ranges (<5%, <20%, <50%, <100%, and others) and crafting a motivational message. Figure 1 shows an example of a user-study interaction and the application of the neuro-symbolic rule.

3.3 HFFood-GPT

We used assistant GPT to create HFFood-GPT as it was easier to integrate with the UI. Since our goal was to educate patients without providing medical advice, we prompted it (the prompt is provided in Appendix B) to not give health advice or suggest consulting a professional for dietary guidance. We named the system Sodium Scout and prompted to analyze the salt contents in foods. It advises that foods exceeding 20% of this intake are not recommended, while those below 5% are favorable choices. We enabled the code interpreter and retrieval features to enable data access. This allowed the assistant to retrieve information from the USFDC dataset, which was also used to create the HFFood-NS dataset. We instructed it not to mention the dataset with patients or search the web for information.

To compare HFFood-GPT with our in-house system, HFFood-NS, we prompted GPT-4 (OpenAI et al., 2024) to function as similarly as possible to HFFood-NS. We prompted Sodium Scout to ask clarification questions about food type, cooking method, and portion size, limiting the questions to one question at a time. Additionally, to address GPT-4's tendency to produce lengthy responses, we prompted it to keep the responses under 40 words.

4 User Study

SetUp We recruited 23 African American (AA) patients, aged between 18 and 89 years, who could speak English and had a history of heart failure. All participants were compensated for their time. The study was conducted while they were hospitalized, in their hospital bedroom.

A total of 20 participants (13 Males and 7 females, ages 18-89 - $\mu = 58.75$, $\sigma = 14.32$) completed the study which took around 6 months. Proper IRB guidelines were followed. If the participants met the inclusion criteria and were willing to participate in the study, they were provided with an informed-consent document and pre-survey questions. Each participant interacted with the two versions of the dialog agent (DA) one after the other in a randomized manner to reduce recall bias.

To help the participants think of questions, we prompted the participants to think about foods they ate for breakfast, lunch, or dinner. Participants then asked questions about the same food items to both dialog systems. This was done to ensure that the participants did not run out of questions for the second interaction and to make it easier to compare the 2 DAs. However, the interactions were not identical because each system asked different clarification questions, which led the conversations in different directions.

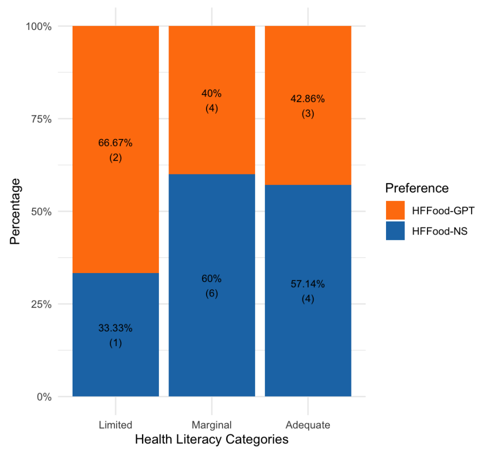
In this study, all interactions with the systems were conducted orally - the dialog systems were speech-based. This included obtaining oral consent and administering both the pre-and post-survey questions verbally. The decision to use oral communication was made to accommodate participants who might be unwell, or be connected to medical equipment, ensuring a more accessible and comfortable experience.

Conducting a user study with hospitalized patients comes with unique challenges and considerations. For instance, Patient 15's partner, who was the primary caretaker responsible for managing food, was also present for the session and together asked questions. Patient 17, despite being visually impaired, had no difficulty participating in the study as the system was speech-based. One patient had a tracheostomy tube, and another was pregnant, highlighting the diversity of participants. Additionally, three patients initially faced some challenges in understanding the systems due to their pain but eventually became comfortable in using it.

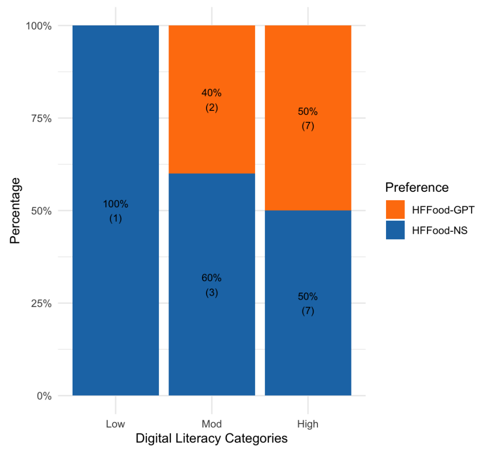
UI For the experimenters, we designed a UI to control the interactions. The UI was minimalistic and designed for ease of use, featuring two radio buttons to select the system and a button to indicate whether the system was recording or listening. For the UI, we utilized Gradio (Abid et al., 2019), a framework provided by Hugging Face, as it offers a simple and effective solution.

The UI was not "always listening" and could only listen or speak one at a time. It was done to prioritize patients' health and accommodate the presence of nurses and doctors during interactions. We named the two systems Lion and Shark so that the participants could easily recall the names of the systems being used. The two names were chosen as they represent a well-known, powerful animal.

We utilized OpenAI technologies for both Text-to-Speech (TTS) and Speech-to-Text (STT) functionalities in the study. For Speech-to-Text, we employed Whisper-1 (Radford et al., 2023), and for Text-to-Speech, we used TTS-1. We recorded audio and collected transcripts for error analysis.



(a) Health Literacy Vs Preference



(b) Digital Health Literacy Vs Preference

Figure 2: Comparison of user preferences categorized by health literacy and digital health literacy levels, with numbers in brackets indicating the number of users.

4.1 Survey Questions

4.1.1 PreSurvey Questionnaire

We assessed participants' health literacy and digital literacy using self-reported measures to better understand their abilities to access and process health information. The PreSurvey Questionnaire can be assessed in Appendix C.

Health Literacy was measured using the **BRIEF Health Literacy Screening Tool** (Alabama Department of Public Health, n.d.). This tool consists of 6 questions that evaluate individuals' ability to "obtain, process, and understand basic health information and services needed to make appropriate decisions about their health."

Digital Literacy was measured using the **eHealth Literacy Scale (eHEALS)** (Norman and Skinner, 2006). This tool consists of 8 questions that assess participants' perceived ability to find, evaluate, and apply electronic health information to address health-related issues.

4.1.2 PostSurvey Questionnaire

After users interacted with both systems, we asked them to evaluate their experience through a questionnaire. The Post-Survey Questionnaire is available in Appendix D.

We asked users about the following aspects of the two systems: how easy it was to understand the answers, the usefulness of the questions, and how easy the questions were to understand. We also asked which system they preferred, the most challenging aspect of using them, their likelihood of using the system in the future, whether the system was helpful, and if they had any additional feedback.

5 Evaluation

5.1 Intrinsic Evaluation

We used the USFDC dataset as a baseline to evaluate the accuracy of the two systems. We treated HFFood-GPT as a task-oriented dialogue system by categorizing the provided values into defined slots (Food, Cook, Type, FoodWeight, and metric) similar to HFFood-NS.

Table 1 summarizes the analysis of task completion and accuracy for both systems. **Task completion** is defined as the system providing a value, whether correct or incorrect, while **accuracy** refers to the system providing the correct salt value. Speech errors were excluded from the accuracy calculation. For task completion, we considered cases where the model provided a salt value (either as a specific number or a range) rather than categorical descriptors as observed in HFFood-GPT.

	HFFood-NS	HFFood-GPT
Avg No of turns	3.6	3
Avg Processing Time	6.7	11.4
Avg Words of the System	14.5	54.5
Avg Retries	2	1.7
Avg WER	.483	.41
Task Completion	84%	62%
Accuracy	37%	24%

Table 1: Intrinsic Evaluation: Comparing 2 systems

Evaluating HFFood-GPT proved particularly challenging due to its black-box nature. To maintain consistency, we evaluated it similarly to HFFood-NS by comparing the provided salt value to the first food item in the USFDC database that satisfied all minimum slot values. However, it remains unknown which data HFFood-GPT actually accesses or the process it follows to calculate its final answers.

Although HFFood-GPT demonstrated lower task completion accuracy compared to HFFood-NS, it achieved higher slot accuracy. Table 2 and Table 3 present the slot accuracy analysis for HFFood-NS and HFFood-GPT, respectively. We classified the incorrect slots into Speech Errors and Partial Speech Errors (P-SE), which likely occurred due to the noisy hospital environment, patients' accents, or the overall conditions in a hospital setting.

Although we used the same Text-to-Speech and Speech-to-Text systems for both the systems, HFFood-NS recorded a higher word error rate (WER) (Morris et al., 2004) than HFFood-GPT (Table 1). HFFood-GPT's incorrect slot errors resulted exclusively from Speech Errors and Partial Speech Errors.

	Correct	Incorrect	Speech Error	P-SE
Food	86	6	4	1
Cook	39	21	11	2
Type	44	31	17	2
Foodweight	29	56	18	0
Metric	27	60	20	0

Table 2: HFFood-NS Slot Accuracy Analysis: incorrect includes Speech Error and Partial Speech Error (P-SE)

	Correct	Incorrect	Speech Error	P-SE
Food	94	7	3	4
Cook	46	5	5	0
Type	58	15	9	6
FoodWeight	59	5	5	0
Metric	59	5	5	0

Table 3: HFFood-GPT Slot Accuracy Analysis: incorrect includes Speech Error and Partial Speech Error (P-SE)

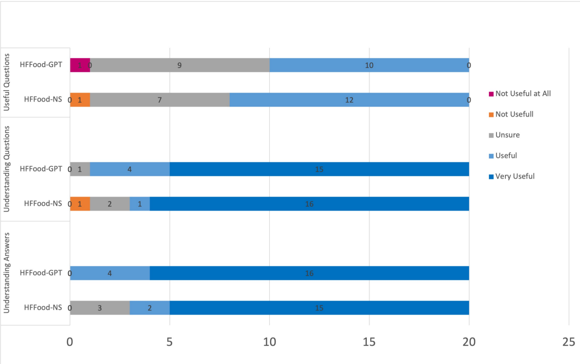


Figure 3: Post-survey: extrinsic evaluation of the two systems on the usefulness and understanding of answers and questions.

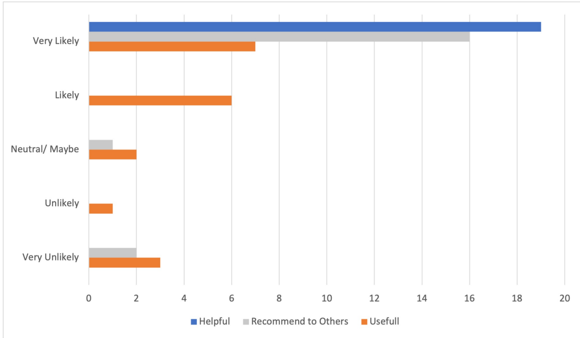


Figure 4: Post Survey Analysis: How users perceived the conversational system

5.2 Extrinsic Evaluation: User Perceptions

Figure 3 shows how participants rated the two systems in the post-survey questionnaire, focusing on Understanding Answers, Understanding Questions, and Useful Questions. We can see that more than 50% of the people find all the aspects useful.

Out of 20 patients, 11 preferred HFFood-NS, while 9 favored HFFood-GPT. Since this is a within-group study, there is a potential for recall bias; however, we did not observe any evidence of it. In 10 cases where HFFood-NS was tried first, 7 users favored it, while in 10 cases where HFFood-GPT was tried first, 6 users favored it. But, the preference for the first system was not statistically significant, as indicated by a Pearson correlation (r = 0.3, p = 0.19).

A comparison of preferences based on health and digital health literacy, as assessed through the pre-survey questionnaire, is presented in Figure 2. The numbers are too small to draw any definite conclusions.

When we asked the users about the reasons for their preferences, those who favored HFFood-NS highlighted its precise and to-the-point answers, faster flow, , and concise responses. On the other hand, users who preferred HFFood-GPT appreciated the ease of understanding its questions, clear and detailed explanations, and better-formulated questions.

While some users found no issues with either system and felt they understood them well, others reported challenges. These included the systems being repetitive, difficulty understanding the questions, and uncertainty about how to phrase their own questions.

Would patients use the DA in their daily life? Table 4 highlights that all users found the system helpful, with most indicating they would recommend it to others. The majority found both systems useful. However, data is only available for 19 patients, as the post-survey questions for patient P5 were interrupted due to a scheduled procedure.

Qualitative measures of salt content. In the post-survey, 55% of users (11 participants) preferred receiving information about salt content in informal terms, such as pinches or dashes. Only 20% (4 participants) preferred formal units like milligrams or grams, while the rest chose percentages or daily values. In the pre-survey, participants reported measuring salt informally by sprinkling, using pinches, or judging by eye. These informal methods differ from the formal values, such as milligrams or percentage of daily value, that appear on food labels.

6 Discussion: comparison between the two systems

To understand why one system was preferred over the other, we applied a mixed modeling approach, treating both the system and participants as random effects. In this analysis, we incorporated participant and system features. For participant features, we considered the Health and Digital Literacy Categories as well as the First System used. For system features, we analyzed metrics such as word error rate, number of turns, number of words, average response time, number of retries, task completion rate, and accuracy. This approach aimed to identify the key factors influencing user preference between the two systems but, unfortunately, did not find any significant factors.

Table 4 summarises the 2 systems, highlighting the pros and cons by comparing performance, Design and Usability, reliability, and flexibility.

	HFFood-NS	HFFood-GPT
Task completion	✓	✗
Accuracy	✓	✗
Slot Accuracy	✗	✓
Less Speech Error	✗	✓
Less Processing Time	✓	✗
More Constrained	✓	✗
Error Analysis	✓	✗
Reliable	✓	✗
Predictable	✓	✗
Handling Complex query	✗	✓
Gave Options	✗	✓
Fluent	✗	✓
Concise	✓	✗
Create easily with less time	✗	✓

Table 4: Pros and Cons of HFFood-NS and HFFood-GPT comparing on performance, Design and Usability, reliability and flexibility.

6.1 HFFood-NS

One significant advantage of employing a neuro-symbolic system combined with a traditional ToDS, rather than directly prompting a LLM, lies in the ability to conduct more thorough and precise error analysis. This capability is crucial, particularly in patient-centric systems. By utilizing this approach, we were able to accurately pinpoint the areas where our system fell short.

Table 5 provides a detailed breakdown of the error analysis performed on HFFood-NS. This analysis includes multiple overlapping categories, which highlight the nuanced nature of errors encountered. By systematically addressing these issues, we can iteratively develop a more robust and reliable dialogue system.

	HFFood-NS
Missed Slot	27
Wrong Food Identified	9
System Error	8
Internet	6
No Data Fetched	5
Wrong Math	5
Food Not in USFDC	4
Complex Query	3
Missed Slot not in USFDC	2
Wrong Food Fetched	1

Table 5: Through and precise error analysis HFFood-NS

Moreover, neuro-symbolic rules gave us the ability to add fail-safe to the system where when the model was not able to infer the slot for food weight and metric, it would assume 100g, thereby increasing the task completion rate.

Moreover, having greater control over the system provided significant advantages. Firstly, it ensured that the system remained aligned with its primary goal, which HFFood-GPT struggled to maintain consistently. For example, in one instance, HFFood-GPT asked a clarification question about the color of the bell pepper being used—a detail that is considered irrelevant when determining the salt amount. This level of control helped minimize distractions and kept the system focused on delivering accurate and goal-oriented responses.

Secondly, ensuring consistency in responses by providing values in milligrams (mg) made the system more user-friendly and reliable. In contrast, HFFood-GPT showed some inconsistency in its responses, with 10% of values given in grams (g), 10% as categorical descriptors (e.g., low, high, significant), 8% as ranges instead of precise values, and 2% in teaspoons (tsp) and rest in mg. This inconsistency could pose challenges for patients, who may overlook that 1 gram is equivalent to 1000 milligrams due to varying levels of numerical literacy (Dickson and Riegel, 2009). Maintaining consistency in responses is a crucial step to improving accessibility and ensuring users can easily interpret the information provided.

6.2 HFFood-GPT

HFFood-GPT handled diverse user descriptions effectively, interpreting non-standard inputs and maintaining smooth conversations. Patients described their food portions in various unconventional ways. Many reported portion sizes in terms of servings, such as "for 6 people" or "serving size for 1," instead of using standard units like ounces, grams, or pounds. Two patients used unique measurements, such as "spoonfuls" and "forkfuls," while two of them relied on size comparisons, like describing a hotdog as "the size of a cellphone" or quantifying food weight as "3 inches of bacon". HFFood-GPT demonstrated flexibility in handling these inputs, whereas HFFood-NS struggled with slot identification, as reflected in Table 2, where most errors occurred in food weight and metric slots.

When HFFood-GPT was unable to determine the salt amount in a food item directly, it attempted to gather the recipe and estimate the salt content based on the salt levels of its ingredients. While this approach was sometimes helpful, there were instances where the system asked overly specific questions, such as the exact weight or brand of each ingredient, which could be burdensome for users.

In some cases, the brand names were so widely recognized that the food was identified by the brand itself, such as "Chick-fil-A" or "Hurricane Whopper." In these situations, HFFood-GPT successfully deduced the food item without requiring additional descriptive details, showcasing its ability to handle commonly branded food references effectively.

However, due to the unpredictable nature of HFFood-GPT, the system sometimes specifically asked for food weight in grams or ounces. For example, in the conversation shown in Table 6, when asked about weight in grams, the user was not aware of it. It was good that HFFood-GPT tried to educate about it. Users may not always know these exact measurements and are more likely to provide quantities in terms they are comfortable with.

	Turn
System	Thank you for specifying the preparation method. Now, could you please tell me the portion size of the grilled turkey sausage in grams?
User	5 milligrams.
System	It seems there might be a misunderstanding with the portion size. Typically, food portion sizes are measured in grams (g) and would be much larger than 5 milligrams (mg). Could you please provide the portion size again in grams? For reference, a common serving size for sausage might be around 50 to 150 grams.
User	100 milligrams
System	It looks like we're still not quite accurate on the portion size. When we're talking about turkey sausage, the realistic portion size is usually around 100 grams (g), not milligrams (mg). One gram is equal to 1000 milligrams, so a portion size of turkey sausage might be 100 grams (which is 100,000 milligrams), for instance. Could you clarify the portion size in grams for the turkey sausage?

Table 6: Excerpt of a HFFood-GPT Conversation

There were 8% instances where HFFood-GPT attempted to assume the slots instead of asking a clarification question. While this approach sometimes aligned with user intent, the goal should be to avoid making assumptions and instead provide users with the choice. For example, in one case, the user mentioned "chicken breast," and despite a potential speech error, the system correctly identified it because chicken breast is the most common way to consume chicken. However, if the user had been asking about "chicken tenders," this assumption would have resulted in an incorrect response.

When building HFFood-GPT, we set specific rules to constrain its behavior. We limited responses to 40 words, but as shown in Table 1, the system averaged 54.5 words per response. We also instructed it to ask only one question per turn, but it violated this rule in about 7% of its turns by asking multiple questions at once.

We had prompted HFFood-GPT to only refer to the database provided and not refer it to the users. However, in 38.5% of system utterances, HFFood-GPT referenced the DB and said that the food item was not in the referenced DB and asked for more clarification. This could be one of the many reasons users did not prefer HFFood-GPT, as it could be off-putting. When we asked users how they would like to improve the system, their primary suggestions were to expand the knowledge base, include more ingredients, and add a broader range of food items to make the system more comprehensive and user-friendly.

GPT performed well by offering users options, such as saying, *"That sounds delicious! Could you tell me if you're using fresh or dried spaghetti?"* which happened in around 30% of system turns. This showed that HFFood-GPT had knowledge about different foods. However, in some cases, even after the option was chosen, HFFood-GPT responded that it lacked the information in the provided database. This behavior frustrated users because GPT relied on its own knowledge instead of the dataset, compromising the system's consistency and reliability.

7 Conclusion and Future Work

We conducted a user study with 20 African-American Heart Failure hospitalized patients. Using a within-subject design, we compared two dialogue systems: an in-house NeuroSymbolic System (HFFood-NS) and a ChatGPT-based system (HFFood-GPT). The evaluation utilized both intrinsic and extrinsic measures, and while neither system emerged as a clear winner, the study highlighted key differences between the two.

The evaluation revealed that HFFood-NS is more accurate, completes more tasks, and provides concise responses compared to HFFood-GPT. On the other hand, HFFood-GPT makes fewer speech errors, requires fewer clarifications to complete tasks, and handles complex queries more effectively.

The widespread use of large language models (LLMs), like ChatGPT, often lacks scrutiny, raising concerns in healthcare settings. Greater control is needed, as relying solely on prompting is not enough. Neuro-symbolic methods, which offer greater transparency, reliability, and explainability, should be further explored and integrated into future systems.

Moving forward, we aim to develop hybrid conversational systems that combine the strengths of both systems.

8 Limitations and Ethics Statement

While we believe user studies/ human evaluation are the best methods to assess medical dialogue systems, they have limitations. Our study recruited only 20 patients, which is a relatively small sample size, and it is difficult to get significant results.

To build the two systems, we relied on the USDA Food Composition Database (USFDC). While this database is a standard reference, it is neither fully exhaustive nor completely accurate. For instance, some foods, such as items from local restaurant chains or specific snack brands, are absent from the USFDC database. Additionally, in conversations about fried foods, users often referred to air fryer cooking methods, which were not accounted for in the database.

References

Abubakar Abid, Ali Abdalla, Ali Rahman, and James Zou. 2019. Gradio: Hassle-free sharing and testing of ml models in the wild. *arXiv preprint arXiv:1906.02569*.

Alabama Department of Public Health. n.d. Brief: Health literacy screening tool. Accessed: 2025-02-14.

Mohammad Rafayet Ali, Seyedeh Zahra Razavi, Raina Langevin, Abdullah Al Mamun, Benjamin Kane, Reza Rawassizadeh, Lenhart K Schubert, and Ehsan Hoque. 2020. A virtual conversational agent for teens with autism spectrum disorder: Experimental results and design lessons. In *Proceedings of the 20th ACM international conference on intelligent virtual agents*, pages 1–8.

RV Belfin, AJ Shobana, Megha Manilal, Ashly Ann Mathew, and Blessy Babu. 2019. A graph based chatbot for cancer patients. In *2019 5th international conference on advanced computing & communication systems (ICACCS)*, pages 717–721. IEEE.

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.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium. Association for Computational Linguistics.

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.

Wenhu Chen, Jianshu Chen, Pengda Qin, Xifeng Yan, and William Yang Wang. 2019. Semantically conditioned dialog response generation via hierarchical disentangled self-attention. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy. Association for Computational Linguistics.

Yun-Nung (Vivian) Chen, Dilek Z. Hakkani-Tür, Gökhan Tür, Jianfeng Gao, and Li Deng. 2016. End-to-end memory networks with knowledge carryover for multi-turn spoken language understanding. In *Interspeech*.

Mohita Chowdhury, Ernest Lim, Aisling Higham, Rory McKinnon, Nikoletta Ventoura, Yajie He, and Nick De Pennington. 2023. Can large language models safely address patient questions following cataract surgery? In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 131–137, Toronto, Canada. Association for Computational Linguistics.

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.

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.

William R. Kearns, Neha Kaura, Myra Divina, Cuong Vo, Dong Si, Teresa Ward, and Weichao Yuwen. 2020. A wizard-of-oz interface and persona-based methodology for collecting health counseling dialog. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI EA '20, page 1–9, New York, NY, USA. Association for Computing Machinery.

Tom Knoll, Francesco Moramarco, Alex Papadopoulos Korfiatis, Rachel Young, Claudia Ruffini, Mark Perera, Christian Perstl, Ehud Reiter, Anya Belz, and Aleksandar Savkov. 2022. User-driven research of medical note generation software. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 385–394, Seattle, United States. Association for Computational Linguistics.

Wenqiang Lei, Xisen Jin, Min-Yen Kan, Zhaochun Ren, Xiangnan He, and Dawei Yin. 2018. Sequicity: Simplifying task-oriented dialogue systems with single sequence-to-sequence architectures. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia. Association for Computational Linguistics.

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.

Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M Krumholz, Jure Leskovec, Eric J Topol, and Pranav Rajpurkar. 2023. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265.

Andrew Cameron Morris, Viktoria Maier, and Phil D Green. 2004. From wer and ril to mer and wil: improved evaluation measures for connected speech recognition. In *Interspeech*, pages 2765–2768.

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.

Cameron D. Norman and Harvey A. Skinner. 2006. eHEALS: The eHealth Literacy Scale. *Journal of Medical Internet Research*, 8:e27.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Baolin Peng, Chunyuan Li, Jinchao Li, Shahin Shayandeh, Lars Liden, and Jianfeng Gao. 2021. Soloist: Building task bots at scale with transfer learning and machine teaching. *Transactions of the Association for Computational Linguistics*, 9.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

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

Maria Riveiro and Serge Thill. 2021. "that's (not) the output i expected!" on the role of end user expectations in creating explanations of ai systems. *Artificial Intelligence*, 298:103507.

Devika Salunke, Carolyn Dickens, Olga Garcia-Bedoya, Paula Allen-Meares, Barbara Di Eugenio, and Andrew Dallas Boyd. 2023. Learnings from pilot testing hfchat, an interactive dialogue agent that can assist black/african american and hispanic/latino patients with heart failure. *Journal of Cardiac Failure*, 29(4):563.

Karen Sparck Jones and Julia R Galliers. 1995. *Evaluating natural language processing systems: An analysis and review*. Springer Science & Business Media.

Pei-Hao Su, Milica Gašić, Nikola Mrkšić, Lina M. Rojas-Barahona, Stefan Ultes, David Vandyke, Tsung-Hsien Wen, and Steve Young. 2016. On-line active reward learning for policy optimisation in spoken dialogue systems. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Berlin, Germany. Association for Computational Linguistics.

Yixuan Su, Lei Shu, Elman Mansimov, Arshit Gupta, Deng Cai, Yi-An Lai, and Yi Zhang. 2022. Multi-task pre-training for plug-and-play task-oriented dialogue system. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4661–4676, Dublin, Ireland. Association for Computational Linguistics.

USFDC. 2022. Us food data central.

Mina Valizadeh and Natalie Parde. 2022. The AI doctor is in: A survey of task-oriented dialogue systems for healthcare applications. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6638–6660, Dublin, Ireland. Association for Computational Linguistics.

Marilyn A Walker, Diane J Litman, Candace A Kamm, and Alicia Abella. 1998. Evaluating spoken dialogue agents with paradise: Two case studies. *Computer Speech & Language*, 12(4):317–347.

Jinping Wang, Hyun Yang, Ruosi Shao, Saeed Abdullah, and S. Shyam Sundar. 2020. Alexa as coach: Leveraging smart speakers to build social agents that reduce public speaking anxiety. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–13, New York, NY, USA. Association for Computing Machinery.

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.

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

Joseph Weizenbaum. 1966. Eliza—a computer program for the study of natural language communication between man and machine. *Commun. ACM*, 9(1):36–45.

Steve Young, Milica Gašić, Blaise Thomson, and Jason D. Williams. 2013. Pomdp-based statistical spoken dialog systems: A review. *Proceedings of the IEEE*, 101.

Victor Zhong, Caiming Xiong, and Richard Socher. 2018. Global-locally self-attentive encoder for dialogue state tracking. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia. Association for Computational Linguistics.

Yue Zhou, Barbara Di Eugenio, Brian Ziebart, Lisa Sharp, Bing Liu, Ben Gerber, Nikolaos Agadakos, and Shweta Yadav. 2022. Towards enhancing health coaching dialogue in low-resource settings. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 694–706, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

A NS-PPTOD Evaluation Results

B HFFood-GPT prompt

To develop HFFood-GPT, we used zero-shot prompting on GPT-4.

Train Size	Epochs	Joint Accuracy	
		PPTOD	NS-PPTOD
100	6	55.56	73.08
300	4	51.92	72.8
500	6	58.75	83.2
1000	6	58.53	85.2

Table 7: Increase in Joint Accuracy when using NS-PPTOD compared to PPTOD across different training sizes

	Train Size	Epochs	Inform	Success	BLEU
PPTOD	100	8	71.43	0	24.99
NS-PPTOD	100	-	88.90	77.80	22.50
PPTOD	300	7	75.00	5.00	34.30
NS-PPTOD	300	-	81.50	63.00	26.90
PPTOD	500	9	82.86	2.86	29.81
NS-PPTOD	500	-	74.50	58.10	28.90
PPTOD	1000	7	93.50	2.70	29.00
NS-PPTOD	1000	-	85.90	71.70	30.00

Table 8: Increase in performance when using NS-PPTOD compared to PPTOD.

Sodium Scout greets users warmly and helps analyze the salt content in various foods. It methodically asks users about the ***food type, cooking method, and portion size, one question at a time, to accurately determine the salt content.*** *Using values from a provided JSON file, Sodium Scout calculates the estimate salt content and compares it to the recommended daily intake of 2000mg. It advises that foods exceeding 20% of this intake are not recommended, while those below 5% are favorable choices.* ***Sodium Scout refrains from giving health advice and suggesting from consulting a professional for dietary guidance.*** *The interface is friendly and straightforward. It focuses on informing users about salt levels in their meals with clarity, ensuring to ask only one question per turn. Answers are* ***kept under 40 words****, and it only searches the data provided in the JSON file.* ***Users do not know about the data file, so don't discuss it. Only focus on information related to food and their salt amount. Do not look for information on the web.***

C Pre-Survey Questions

- eHealth Literacy Questions
- Brief Health Literacy Screening Questions
- Do you pay attention to salt in your food [Yes, No, Maybe] How do you measure it?

D Post-Survey Questions

- 1. How easy was it to understand the answers you received from [insert system name: Lion/Shark]? Please rate from 1 to 5, where 5 is very easy to understand and 1 is very difficult to understand.
- Did you think [insert system name: Lion/Shark] asked useful questions? (a. Mostly Yes, b. Yes, c. Mostly No,d. No)
- How easy was it to understand questions from [insert system name Lion/Shark]? Please rate from 1 to 5, where 5 is very easy to understand and 1 is very difficult to understand.
- How would you like to receive information about the salt content in your food. Would you prefer to see it as milligrams, as a percentage of your daily value, or in some other way? Please share your preference.
- Which system, Lion or Shark, do you prefer?
- In your own words, what was the hardest aspect of using the two systems?
- How likely are you to use such a system? (Rate 1 to 5 for each system: 1-Very unlikely to use, 2: Unlikely to use, 3: Neutral, 4: Likely to use, 5: Very likely to use) a. Why? b. If the participant's response is yes, i. How often would you use it? ii. Would you recommend it to others? (Yes/No)
- Was the system helpful? a. How would you improve the system? (If there is a preferred one), b. How would you improve the systems? (If there is **not** a preferred one)
- Do you have any additional feedback or comments?