

Conversational Assistants to Support Patients with Heart Failure: Comparing a Neurosymbolic Architecture with GPT

Anuja Tayal¹, Devika Salunke², Barbara Di Eugenio¹, Paula G Allen-Meares³,
Eulalia P Abril⁴, Olga Garcia-Bedoya³, Carolyn A Dickens³, Andrew D. Boyd²

¹ Department of Computer Science ² Department of Biomedical and Health Information Sciences

³ Department of Medicine ⁴ Department of Communications

University of Illinois Chicago, IL, USA

{atayal4, dsalun2, bdieugen, pameares, eulalia, ogarciab, cdickens, boyda}@uic.edu

Abstract

Conversational assistants are becoming increasingly popular, including in healthcare, partly due to the availability and capabilities of Large Language Models. There is a need for controlled, probing evaluations with real stakeholders, which can highlight the 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 patients with heart failure to ask about the salt content in food. One version of the system was developed with a neurosymbolic architecture, and another is based on GPT. Our objective in evaluating the two dialogue systems was not only to compare task performance but also to gain insights from real stakeholders. Results indicate that the two systems complement each other, highlighting the promise of a hybrid approach that leverages the strengths of both systems.

Keywords: Evaluation, User-study, Medical Dialog Systems, Neuro-Symbolic systems

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 (TODS) are primarily evaluated using automated metrics (Budzianowski et al., 2018). However, as demonstrated early on by the PARADISE framework (Walker et al., 1998), task performance is only one aspect of evaluating a dialogue system.

LLMs (OpenAI et al., 2024; Touvron et al., 2023; Brown et al., 2020; DeepSeek-AI et al., 2025) offer strong generalization and adaptability across a wide range of diverse tasks. Yet, these very capabilities also make their evaluation increasingly complex (Zhou et al., 2024; Singh et al., 2024). Unlike traditional systems, 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 (Ji 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 (Walker et al., 1998; Ji et al., 2023).

As a result, human evaluation remains the gold standard for assessing medical dialog systems (Zhou et al., 2024). Human evaluators can provide insights into subjective qualities such as coherence, informativeness, and user satisfaction—factors that are difficult to capture with automated metrics.

We present a user study comparing a traditional TODS in healthcare with one based on LLMs. We aim to derive meaningful insights by deploying the two dialog systems in a patient-centric environment. The study focuses on African American (AA) patients with heart failure, aiming to assist them in managing their salt intake by providing information about the salt content in various foods. Patients with heart failure must meticulously monitor and reduce their salt intake: AA individuals are more prone to heart failure (Nayak et al., 2020), and have a higher sensitivity to salt. Many areas in large cities where AA tend to live are lower-income and lack access to healthy food. Furthermore, in (Gupta et al., 2020), the authors show that AA patients with heart failure often discuss salt and food during educational sessions, indicating a significant interest in this area.

The study compares two dialog systems—a Neuro-Symbolic System (HFFood-NS) with a GPT-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 sur-

veys to evaluate the two systems with AA patients while they are 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 the relevance and value of the intervention with real stakeholders.

The patient-centered perspective provides a deeper understanding of how dialogue systems operate, revealing their practical strengths and limitations, as well as opportunities for improvement.

The evaluation focuses on comparing the two systems, highlighting their advantages and disadvantages. The results show that the two systems complement each other, underscoring the potential of a hybrid approach that leverages the strengths of both LLMs and neuro-symbolic systems. This direction is particularly promising for healthcare dialogue systems, where more control is needed.

2. Related Work

Task-oriented dialog systems (TODS) are designed to accomplish a specific task within a limited number of turns and typically follow a pipeline structure. While effective, the pipeline systems rely on human-labeled data, motivating End-to-End models that learn directly from data. Within healthcare (Valizadeh and Parde, 2022), TODS have gained traction for patient education (Cai et al., 2023), including on heart failure (Moulik, 2019; Gupta et al., 2020), mental health (Ali et al., 2020), anxiety (Wang et al., 2020).

The introduction of the transformer architecture (Vaswani et al., 2017) enabled training of language models (LMs) such as GPT (Brown et al., 2020) and T5 (Raffel et al., 2020) which improved generalization. As models scaled into LLMs, they exhibited capabilities like reasoning (Qiao et al., 2023), but became a closed box. Moreover, they continued to struggle with tasks such as mathematical reasoning (Nogueira et al., 2021; Qian et al., 2023) and were prone to generating hallucinations. Subsequent research focused on enhancing control, interpretability, and reasoning through advanced prompting (Brown et al., 2020; Yao et al., 2023) and neuro-symbolic approaches.

Neuro-Symbolic Methods combine the generalization strengths of neural networks with the structure and interpretability of symbolic reasoning. For instance, (Romero et al., 2021) introduced symbolic representations into GPT-2 outputs to enhance structural awareness. DILOG (Zhou et al., 2020) leveraged inductive logic programming to learn dialogue policies from limited data, enabling zero-shot transfer. Similarly, (Arabshahi et al., 2021) showcased how multi-hop and commonsense reasoning

can be incorporated into dialogue systems.

Evaluation of TODS Evaluating the true capabilities of dialogue systems remains challenging and typically involves automated metrics and human evaluation. Automated metrics assess individual components—intent classification accuracy, entity F1 for NLU, joint goal accuracy for Dialogue State Tracking, and BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and METEOR (Banerjee and Lavie, 2005) for NLG. End-to-end evaluation commonly uses Inform and Success Rate (Budzianowski et al., 2018).

However, task success alone does not reflect user satisfaction, as highlighted by the PARADISE framework (Walker et al., 1998). Consequently, human evaluation remains the gold standard, especially in medical domains (Yeh et al., 2021; Deriu et al., 2021; Zhou et al., 2024), where subjective factors like coherence and informativeness are crucial. User studies are also often used to collect real-world data in domains lacking specific use-case datasets (Knoll et al., 2022). (Knoll et al., 2022) adopted a user-centered approach to iteratively refine a medical note generation model through feedback. As conducting a user study is costly, researchers are exploring other approaches like user simulation (Yun et al., 2025; Park et al., 2023). With the rise of LLMs, there is growing interest in using them as automated judges (Zheng et al., 2023), though their reliability remains debated.

3. Two dialogue system architectures

We compare and evaluate the two dialog systems—Neuro-Symbolic System (HFFood-NS) and the GPT-based system (HFFood-GPT). Both systems are designed to help patients with heart failure inquire about salt content in foods. The patients initiate the conversation by inquiring about the salt content in food, and the system responds by asking clarification questions to accurately determine the sodium amount¹. Both systems utilize the U.S. Food Data Central (USFDC) dataset (USFDC, 2022), a publicly available resource from the U.S. Department of Agriculture (USDA) that provides culturally diverse food entries with detailed descriptions and nutrient information.

3.1. HFFood-NS

HFFood-NS is a task-oriented dialog system adapted from our previous work (Tayal et al., 2024). We annotated the USFDC dataset with five key slots—*food*, *cook*, *type*, *foodweight*, and *metric*—and constructed a template-based conversational dataset comprising 87k dialogues. Using

¹We use the terms sodium and salt interchangeably.

this dataset, we trained a T5-small (Raffel et al., 2020) based end-to-end TODS dialogue system through few-shot learning. Although the model accurately identified slot values, it achieved only a 2% success rate in computing salt values (Wei et al., 2022) due to the multi-valued nature of salt, which varies with preparation method and portion size. To overcome these limitations, we integrated the model with neuro-symbolic rules. This integration enabled the system to retrieve precise salt values from the USFDC database, perform mathematical calculations for specific food weights, and effectively handle non-standard food quantities.

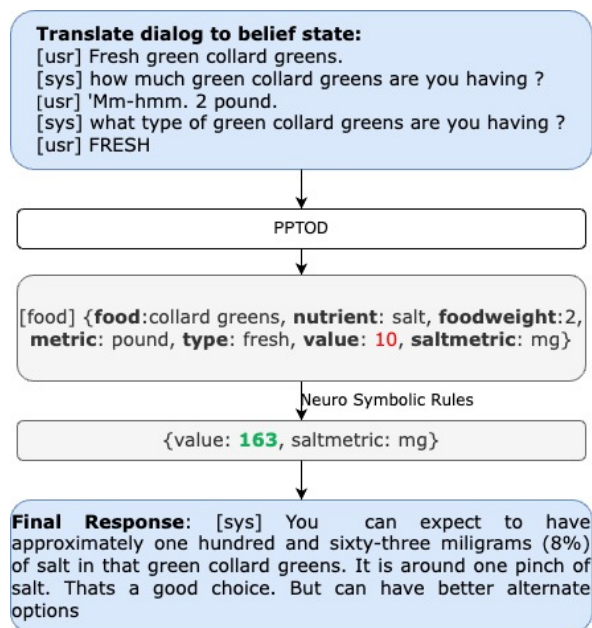


Figure 1: **HFFood-NS model interaction with a patient:** As the database contained salt values for 100 grams of collard greens, while the user inquired for 2 pounds, the neuro-symbolic system retrieved the appropriate data and computed the correct value (163).

To deploy HFFood-NS in a patient-centric setting, we modified our previous approach from (Tayal et al., 2024). For simplicity and reliability, only the DST module was used with template-based system responses. If a slot remains unfilled, the system attempts to query the user up to two times. We designed the final system response to clearly convey salt values and encourage healthier options. The salt content was expressed in milligrams (mg), percentage (%), and informal units (pinches/dashes) to explore patient preferences (Dickson and Riegel, 2009; Nayak et al., 2020; Tayal et al., 2024). The salt value was compared to the 2000 mg daily limit and categorized into four ranges (<5%, <20%, <50%, <100%, and others) with motivational message. ² Figure 1 shows an example of a user-

²This was based on the discussions with clinician

study interaction and the application of the neuro-symbolic rule.

3.2. HFFood-GPT

GPT-4 (OpenAI et al., 2024) was used with zero-shot prompting to create HFFood-GPT. Assistant GPT (Purpose-built AI that uses OpenAI’s models and calls tools) was used to create HFFood-GPT. To keep the two systems aligned and comparable, it was prompted to analyze the salt contents in foods and ask clarification questions about food type, cooking method, and portion size, limiting the questions to one question at a time. It advises that foods exceeding 20% of salt intake are not recommended, while those below 5% are favorable choices. We provided the USFDC dataset, by enabling the retrieval tool. Since our goal was to educate patients without providing medical advice, it was prompted not to give health advice or suggest consulting a professional for dietary guidance. Additionally, to address GPT-4’s tendency to produce lengthy responses (Tayal et al., 2025; Ji et al., 2023), we prompted it to keep the responses under 40 words. It was instructed not to mention the dataset or search the web for information. The prompt is shown in Appendix A.

4. User Study

Setup We approached approximately 70–80 hospitalized AA patients aged 18–89 years who spoke English and had a history of heart failure. Of the 23 recruited patients, 20 patients (13 males and 7 females, aged 18-89, $\mu = 58.75$, $\sigma = 14.32$) completed the study, which in its entirety took approximately 6 months. Proper institutional guidelines were followed. If patients met the inclusion criteria and were willing to participate in the study, they were provided with an informed consent document. This was a within-group study, meaning each participant interacted with both versions of the dialog agent (DA) in a randomized manner, one after the other, to reduce recall bias. Upon completion, the patients received a \$20 gift card.

Patients were prompted to ask each DA about the salt content of 5 food items. To help participants think of questions, we prompted them to consider the foods they had eaten for breakfast, lunch, or dinner. Participants then asked each DA questions about the same 5 food items. This was done to ensure that the participants did not run out of questions for the second interaction and to facilitate easier comparison between the two DAs. However, the interactions were not identical because each system asked different clarification questions, which led the conversations in different directions. At the

collaborators.

end of the user study, approximately 200 conversations were collected (20 patients × 5 food items × 2 models), which are being made publicly available following a proper de-identification process.³

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. We used OpenAI's Whisper-1 (Radford et al., 2023) for Speech-to-Text and TTS-1 (Zeng et al., 2020) for Text-to-Speech, recording audio and transcripts for error analysis.

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 they asked questions together. 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, which affected their speech. Additionally, three patients initially faced some challenges due to pain, but eventually became comfortable and could participate in the experiment.

The User Interface (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 HuggingFace.

The UI was not "always listening" and could either listen or speak at a time, ensuring patient comfort and accommodating nurses and doctors during interactions. We named the two systems *Lion* and *Shark* so that participants could easily recall the names of the systems being used. The two names were chosen because they represent a well-known and powerful animal, and at the same time mask the nature of the technology behind the UI.

4.1. Survey Questions

The patients were administered a pre- and post-survey questionnaire. The pre-survey evaluated participants' health literacy and digital health literacy, while the post-survey assessed the performance of the two systems, as well as participants' preferences and likelihood of using them in the future.

³<https://github.com/anujatayal/HFChat-NS-GPT>

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 B.

Health Literacy was measured using the **BRIEF Health Literacy Screening Tool** (Alabama Department of Public Health, n.d.). This tool consists of 4 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 the patients interacted with both systems, we asked them to evaluate the systems and their experience through a questionnaire. The Post-Survey Questionnaire is available in Appendix C.

We asked the patients 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 such a system in the future, whether the system was helpful, and if they had any additional feedback.

5. Evaluation

This section presents a comprehensive evaluation of the two systems using both intrinsic and extrinsic evaluation methods. The intrinsic evaluation focuses on the performance metrics—task accuracy and task completion—while the extrinsic evaluation assesses the patient's perception of the system.

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 TODS by categorizing the provided values into defined slots (Food, Cook, Type, FoodWeight, and metric) similar to HFFood-NS. Since the USFDC dataset contains detailed information, and users often provide less precise food descriptions, we use the first food item that meets all the minimum required keywords to account for the multi-valued nature of food items when providing salt values. However, it remains unknown

which data HFFood-GPT accesses or the process it follows to calculate its final answers.

Table 1 summarizes the analysis of task completion and accuracy for both systems. **Task completion** is defined as the system providing a salt value, whether correct or incorrect. HFFood-GPT provided salt values in specific numbers, ranges, and categorical values (e.g., low, high, significant). However, we excluded categorical values as they were too vague and could not be used to calculate task accuracy. Moreover, patient educators (Gupta et al., 2020) and existing guidelines emphasize the use of precise numerical values, even when approximate (e.g., pinch, etc.). Task accuracy refers to the system providing the correct salt value and is calculated with respect to task completion. We report two types of task accuracy: **Task Accuracy System** and **Task Accuracy Net**. The Task Accuracy System evaluates the system’s performance based only on the slots whose values were recognized accurately, i.e., excluding slots with speech errors; this ensures the results reflect only the system’s capabilities, as the speech-to-text and text-to-speech components were external to the system. Task Accuracy Net, on the other hand, includes all cases and reflects the overall accuracy, incorporating both system behavior and speech errors.

	HFFood-NS	HFFood-GPT
Avg No of turns	3.6	3.0
Avg Processing Time	6.7	11.4
Avg Words by System	14.5	54.5
Avg Retries	2.0	1.7
Avg WER	0.48	0.41
Task Completion	84%	62%
Task Accuracy System	37%	24%
Task Accuracy Net	11%	20%

Table 1: **Intrinsic Evaluation**

HFFood-NS achieved higher task completion accuracy than HFFood-GPT, largely due to its ability to incorporate a fail-safe mechanism, which will be explained in Section 6. While HFFood-NS achieved higher task accuracy at the system level, its task accuracy net was lower than that of HFFood-GPT. This could be due to two reasons. First, HFFood-NS recorded a higher word error rate (WER) (Morris et al., 2004), despite both systems using the same text-to-speech and speech-to-text components (Table 1). Second, HFFood-GPT’s higher task accuracy net was influenced by the fact that 50% of its correct responses were range-based values. Section 6 offers additional insights into these results. It is important to note that the neuro-symbolic system (Tayal et al., 2024) provides significant value; without it, both the task accuracy system and the task accuracy net would have been 0%.

	Correct	Incorrect*	Speech Error	P-SE
Food	87	14	5	1
Cook	39	21	11	2
Type	44	31	17	2
Foodweight	32	58	18	0
Metric	28	62	20	0

Table 2: **HFFood-NS Slot Accuracy**

*Incorrect includes Speech Error and P-SE.

HFFood-GPT outperformed HFFood-NS in slot accuracy (excluding salt value). This may be attributed to the more flexible and unpredictable nature of GPT-based interactions—it often understood user intent correctly but struggled to fetch the appropriate values from the database. Detailed slot accuracy analyses for both systems are presented in Table 2 and Table 3. Among incorrect slots, some are Speech Errors or Partial Speech Errors (P-SE), which likely stemmed from the noisy hospital environment, patients’ accents, or the overall conditions in a hospital setting. We define a partial speech error as an instance where the user’s utterance is only partially produced or distorted, rather than being completely substituted or omitted (speech error).

	Correct	Incorrect*	Speech Error	P-SE
Food	94	7	3	4
Cook	46	5	5	0
Type	58	15	9	6
FoodWeight	62	5	5	0
Metric	62	5	5	0

Table 3: **HFFood-GPT Slot Accuracy**

*Incorrect includes Speech Error and P-SE

5.2. Extrinsic Evaluation: User Perceptions

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 it. In 10 cases where HFFood-NS was tried first, 7 patients favored it, while in 10 cases where HFFood-GPT was tried first, 6 patients favored it. But, the preference for the first system was not statistically significant, as indicated by a Pearson correlation ($r = 0.3$, $p = 0.19$).

Figure 2 shows how participants rated the two systems in the post-survey questionnaire - please note that in the following we are referring to questions asked by the system, and system’s answers. The three dimensions are usefulness and understandability of questions, and understandability of

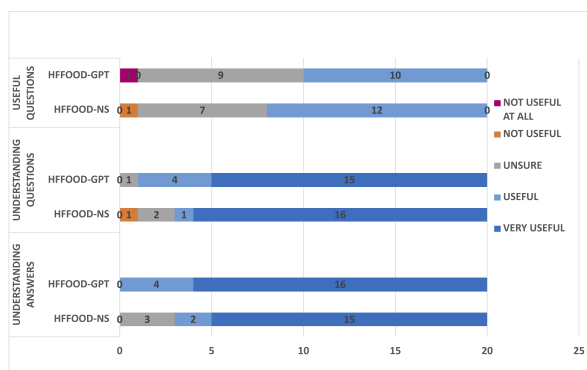


Figure 2: Post-survey: extrinsic evaluation of the two systems on the usefulness and understandability of questions and answers.

answers: at least 50% of subjects rate all these aspects as 4 or 5 (1-5 Likert scale).

A comparison of preferences based on health and digital health literacy, as assessed through the pre-survey questionnaire, is presented in Figure 3. For example, it appears low digital health literacy patients prefer HFFood-NS. However, the sample size is too small to draw definitive conclusions, and the findings are based on self-reported measures.

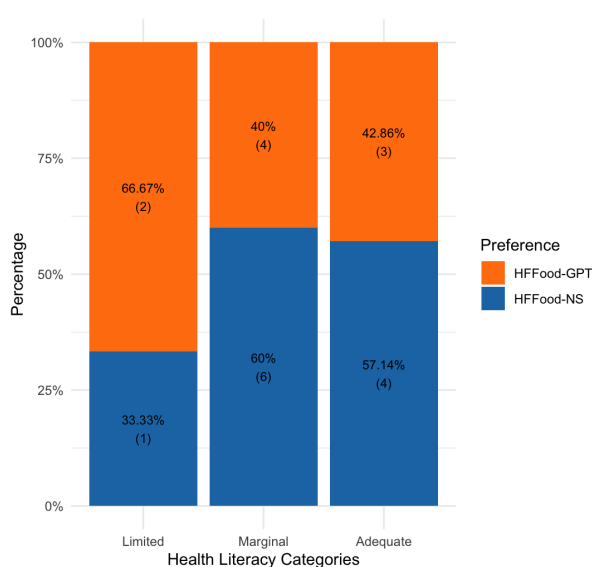
When we asked the patients about the reasons for their preferences, those who favored HFFood-NS highlighted its precise and to-the-point answers, faster flow, and concise responses. On the other hand, the patients who preferred HFFood-GPT appreciated the ease of understanding its questions, clear and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

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

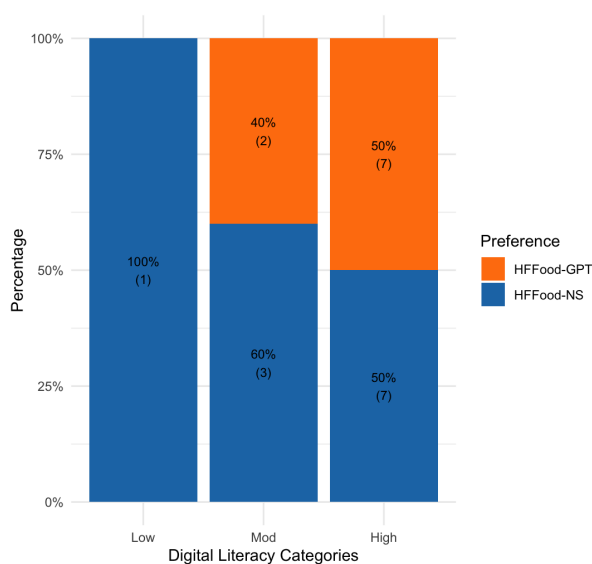
Would patients use the DA in their daily life?

Figure 4 highlights that all the patients found the systems helpful, with most indicating they would recommend them to others. The majority found both systems useful.

Qualitative measures of salt content. In the post-survey, 55% of the patients (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.



(a) Health Literacy Vs Preference



(b) Digital Health Literacy Vs Preference

Figure 3: User preferences by health and digital health literacy levels (numbers in brackets indicate patient counts)

6. Discussion: comparison between the two systems

One significant advantage of employing a neuro-symbolic system combined with a traditional TODS is the ability to conduct **more thorough and precise error analysis**. This capability is crucial, particularly in patient-centric systems. By utilizing this approach, one can accurately pinpoint the areas where the system fell short.

Table 4 provides a detailed breakdown of the error analysis performed on HFFood-NS, which we describe in detail:

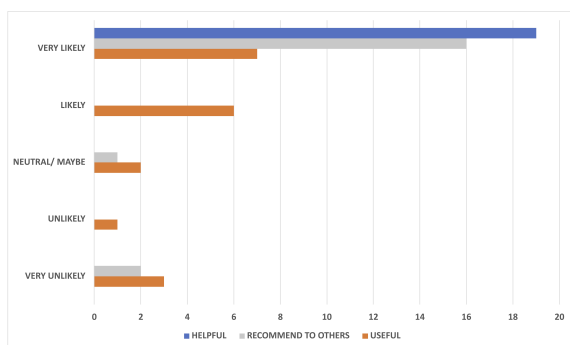


Figure 4: **Post Survey:** How patients perceived the DAs

	HFFood-NS
Missed Slot	23
No Data Fetched	18
Metric Mismatch	5
Wrong Food Identified	8
Slot not in USFDC	4

Table 4: **HFFood-NS: Error analysis**

1. **Missed Slot:** Inability of the system to identify the value of a particular slot.
2. **No Data Fetched:** Inability of the system to retrieve data from the database due to the following reasons:
 - **Complex Query:** HFFood-NS could only calculate salt content for a single food item, while the patients often provided more descriptive inputs (e.g., "*What is the sodium content in two grilled hot links with three ounces of barbecue sauce?*"). Such complex queries exceeded the system's capability but were handled effectively by HFFood-GPT. Of five complex queries, HFFood-NS correctly identified the food slot in three cases but failed to retrieve data.
 - **System Error:** In 6 cases, the system failed to retrieve values due to internal issues, such as exception handling failures or other internal faults.
 - **Food Not in USFDC** Although the USFDC is comprehensive, it lacks certain food items—such as local brands like Jays Potato Chips and the Hurricane Whoper—which occurred in three cases.
 - **Internet Connectivity Issue:** Constant internet problems at the hospital impacted several studies, with at least three confirmed cases.
3. **Metric Mismatch:** In five instances, incorrect calculations were caused by inconsistencies

in weight units (e.g., confusing milligrams with grams).

4. **Wrong Food Identified:** In 8 instances, HFFood-NS identified the food slot incorrectly. These errors occurred due to complex food descriptions or speech errors.
5. **Slot not in USFDC:** The patients were imaginative, leading to incomplete slots (excluding the *food* slot, already addressed under the *Food Not in USFDC* type) for items such as *charred lollipop lamb chops*, which were not present in the USFDC database.

This analysis highlights the nuanced nature of errors encountered. By systematically addressing these issues, one can iteratively develop a more robust and reliable dialogue system.

Another key benefit of the neuro-symbolic approach is the ability to add a fail-safe and make it more **controlled**. For instance, when the system fails to infer the food weight and metric, it assumes 100g, thereby increasing the task completion rate.

Moreover, the **enhanced control** offered by the neuro-symbolic system ensures the system is aligned with the primary goal—a consistency that HFFood-GPT often struggled with. For example, in one instance, HFFood-GPT asked a clarification question about the color of a bell pepper— an irrelevant detail when determining the salt amount. The rule-based structure of HFFood-NS helps avoid such detours, maintaining goal-oriented, streamlined interactions.

In addition to improved control, the neuro-symbolic system ensured consistency in responses by providing values in milligrams (mg), which made the system both **more patient-friendly and reliable**. In contrast, HFFood-GPT showed some inconsistency in its responses, with 10% of values given in grams (g), 10% as categorical values, 8% as ranges instead of precise values, and 2% in teaspoons (tsp) and the rest in mg. This inconsistency can pose challenges for patients, particularly those with limited numerical literacy (Dickson and Riegel, 2009), who may not easily recognize that 1 gram equals 1000 milligrams, as illustrated in Figure 5, which shows an excerpt of an HFFood-GPT conversation with the patient. Maintaining consistency in responses is a crucial step to improving accessibility and ensuring users can easily interpret the information provided.

On the other hand, HFFood-GPT was **fluent** and 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*", rather than using standard units like ounces, grams, or

Speaker	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?
Patient	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.
Patient	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?

Figure 5: HFFood-GPT "lecturing" a patient about weight metrics

pounds. Two patients used unique measurements, such as *spoonfuls* and *forkfuls*, while two of them relied on size comparisons, like describing a hot dog as "the size of a cellphone" or quantifying food weight as "3 inches of bacon". HFFood-GPT demonstrated flexibility in **handling these complex queries**, 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 the patients.

	HFFood-NS	HFFood-GPT
Error Analysis	✓	✗
Controlled	✓	✗
Reliable	✓	✗
Predictable	✓	✗
Complex queries	✗	✓
Gave Options	✗	✓
Fluent	✗	✓
Concise	✓	✗
Easier and faster to develop	✗	✓

Table 5: Pros and Cons of HFFood-NS and HFFood-GPT: usability, reliability, and flexibility.

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, because HFFood-GPT's behavior was sometimes unpredictable, it occasionally asked for food weights in grams or ounces. For instance, in Figure 5, when the system requested a weight in grams, the user—unaware of the difference—responded in milligrams. While it was positive that HFFood-GPT attempted to clarify the distinction, its subsequent responses became overly pedantic. Patients may not always know precise measurements and are more likely to express quantities in terms familiar to them.

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 a choice. For instance, when one patient mentioned *chicken breast*, the system correctly identified the slot despite a speech error, likely because chicken breast is the most common form of chicken consumption.

When building HFFood-GPT, we set specific rules to constrain its behavior. We limited responses to 40 words, but HFFood-GPT was **verbose**, with an average of 54.5 words per response (Table 1). 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 mention 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.

Table 5 compares the two systems across design features (controlled, error analysis), and reliability. HFFood-NS, while more controlled, allows for detailed and precise error analysis, making it more reliable and predictable. It made more speech errors, but is more concise and faster. In contrast, HFFood-GPT is a closed box—less reliable and predictable—but delivers more fluent responses, handles complex queries, provides users with options; however, it tends to be more verbose.

7. Conclusions and Future Work

We conducted a user study with 20 hospitalized African American patients with heart failure, focusing on real patients—rather than typical study participants like Mechanical Turk workers or students. Using a within-subject design, we compared a NeuroSymbolic System (HFFood-NS) and a GPT-based dialog system (HFFood-GPT). Evaluating the two dialog systems in a patient-centric environment provided meaningful insights. While neither system emerged as a clear winner, the study highlighted key differences: 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. These findings highlight that the two systems complement each other, suggesting the potential of a hybrid approach. This direction is particularly promising for healthcare dialogue systems, where more control is needed.

Moving forward, we aim to develop hybrid conversational systems that leverage the strengths of both systems by decoupling dialog management and response generation. The dialog management will integrate dialog acts to control the conversation flow while utilizing LLMs to enhance response generation (He et al., 2018; Zhang et al., 2025).

8. Limitations and Ethics Statement

We presented our findings within the context of a specific, patient-centric use case—a medical dialogue system designed to help African American patients with heart failure understand the salt content of foods. This use case is uniquely complex: the user study was conducted in a real-world hospital setting with an African American population that has historically experienced lower levels of health literacy and numeracy.

Given the sensitive nature of working with patient populations, ethical considerations were central to the study’s design and deployment. Participation

was voluntary and conducted in accordance with institutional ethical guidelines, with appropriate consent obtained from participants. Importantly, the study did not affect patient treatment in any way; the system was used solely for research purposes and did not influence clinical decision-making or medical care.

While these findings are preliminary and should be validated across other domains, we anticipate similar trends emerging: the two systems complement each other, suggesting the potential of a hybrid approach.

9. Acknowledgments

For financial support, we acknowledge a DPI Cycle 1 funding award and NSF grant IIS 2232307.

10. Bibliographical 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.
- Forough Arabshahi, Jennifer Lee, Mikayla Gawarecki, Kathryn Mazaitis, Amos Azaria, and Tom Mitchell. 2021. Conversational neuro-symbolic commonsense reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 4902–4911.
- Satanjeev Banerjee and Alon Lavie. 2005. [ME-TOR: An automatic metric for MT evaluation with improved correlation with human judgments](#). In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.
- 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. Multi-WOZ - 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.
- DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damaí Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yudian Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. [Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning](#).
- Jan Deriu, Alvaro Rodrigo, Arantxa Otegi, Guillermo Echevoyen, Sophie Rosset, Eneko Agirre, and Mark Cieliebak. 2021. Survey on evaluation methods for dialogue systems. *Artificial Intelligence Review*, 54:755–810.
- 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.
- He He, Derek Chen, Anusha Balakrishnan, and Percy Liang. 2018. Decoupling strategy and generation in negotiation dialogues. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2333–2343.
- Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. [Survey of hallucination in natural language generation](#). *ACM Comput. Surv.*, 55(12).

- 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.
- Chin-Yew Lin. 2004. [ROUGE: A package for automatic evaluation of summaries](#). In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- 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.
- Rodrigo Nogueira, Zhiying Jiang, and Jimmy Lin. 2021. Investigating the limitations of transformers with simple arithmetic tasks. *arXiv preprint arXiv:2102.13019*.
- 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 Altschmidt, 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, Madeleine 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, Na-

- talie 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, Lillian 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](#).
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. [Generative agents: Interactive simulacra of human behavior](#). In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, New York, NY, USA. Association for Computing Machinery.
- Jing Qian, Hong Wang, Zekun Li, Shiyang Li, and Xifeng Yan. 2023. [Limitations of language models in arithmetic and symbolic induction](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9285–9298, Toronto, Canada. Association for Computational Linguistics.
- 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.
- 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.
- Oscar J. Romero, Antian Wang, John Zimmerman, Aaron Steinfeld, and Anthony Tomasic. 2021. A task-oriented dialogue architecture via transformer neural language models and symbolic injection. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 438–444, Singapore and Online. Association for Computational Linguistics.
- Chandan Singh, Jeevana Priya Inala, Michel Galley, Rich Caruana, and Jianfeng Gao. 2024. Rethinking interpretability in the era of large language models. *arXiv preprint arXiv:2402.01761*.
- Karen Sparck Jones and Julia R Galliers. 1995. *Evaluating natural language processing systems: An analysis and review*. Springer Science & Business Media.
- 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.
- Anuja Tayal, Devika Salunke, Barbara Di Eugenio, Paula Allen-Meares, Eulalia Puig Abril, Olga Garcia-Bedoya, Carolyn Dickens, and Andrew Boyd. 2025. [Towards conversational assistants for health applications: using ChatGPT to generate conversations about heart failure](#). In *Proceedings of the 26th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 527–537, Avignon, France. Association for Computational Linguistics.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. [Llama: Open and efficient foundation language models](#).
- 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.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.
- 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.
- Joseph Weizenbaum. 1966. [Eliza—a computer program for the study of natural language communication between man and machine](#). *Commun. ACM*, 9(1):36–45.
- Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik R Narasimhan. 2023. [Tree of thoughts: Deliberate problem solving with large language models](#). In *Thirty-seventh Conference on Neural Information Processing Systems*.
- Yi-Ting Yeh, Maxine Eskenazi, and Shikib Mehri. 2021. [A comprehensive assessment of dialog evaluation metrics](#). In *The First Workshop on Evaluations and Assessments of Neural Conversation Systems*, pages 15–33, Online. Association for Computational Linguistics.
- Taedong Yun, Eric Yang, Mustafa Safdari, Jong Ha Lee, Vaishnavi Vinod Kumar, S. Sara Mahdavi, Jonathan Amar, Derek Peyton, Reut Aharoni, Andreas Michaelides, Logan Schneider, Isaac Galatzer-Levy, Yugang Jia, John Canny, Arthur Gretton, and Maja Matarić. 2025. [Sleepless nights, sugary days: Creating synthetic users with health conditions for realistic coaching agent interactions](#).
- Zhen Zeng, Jianzong Wang, Ning Cheng, Tian Xia, and Jing Xiao. 2020. [AlignTTS: Efficient feed-forward text-to-speech system without explicit alignment](#). In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6714–6718.
- Chao Zhang, Xin Shi, Xueqiao Zhang, Yifan Zhu, Yi Yang, and Yawei Luo. 2025. [DecoupledESC: Enhancing emotional support generation via strategy-response decoupled preference optimization](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 22189–22215, Suzhou, China. Association for Computational Linguistics.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.
- Hongjian Zhou, Fenglin Liu, Boyang Gu, Xinyu Zou, Jinfa Huang, Jinge Wu, Yiru Li, Sam S. Chen, Peilin Zhou, Junling Liu, Yining Hua, Chengfeng Mao, Chenyu You, Xian Wu, Yefeng Zheng, Lei Clifton, Zheng Li, Jiebo Luo, and David A. Clifton. 2024. [A survey of large language models in medicine: Progress, application, and challenge](#).
- Zhenpeng Zhou, Ahmad Beirami, Paul Crook, Pararth Shah, Rajen Subba, and Alborz Geramifard. 2020. Resource constrained dialog policy learning via differentiable inductive logic programming. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6775–6787, Barcelona, Spain (Online). International Committee on Computational Linguistics.

A. HFFood-GPT prompt

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

*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.***

B. 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?

C. 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?