A Neuro-Symbolic Approach to Monitoring Salt Content in Food

Anuja Tayal[1], Barbara Di Eugenio[1], Devika Salunke[2], Andrew D. Boyd[2], Carolyn A Dickens[3], Eulalia P Abril[4], Olga Garcia-Bedoya[3], Paula G Allen-Meares[3]
[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, bdieugen, dsalun2, boyda, cdickens, eulalia, ogarciab, pameares }@uic.edu

Abstract

We propose a dialogue system that enables heart failure patients to inquire about salt content in foods and help them monitor and reduce salt intake. Addressing the lack of specific datasets for food-based salt content inquiries, we develop a template-based conversational dataset. The dataset is structured to ask clarification questions to identify food items and their salt content. Our findings indicate that while fine-tuning transformer-based models on the dataset yields limited performance, the integration of Neuro-Symbolic Rules significantly enhances the system's performance. Our experiments show that by integrating neuro-symbolic rules, our system achieves an improvement in joint goal accuracy of over 20% across different data sizes compared to naively fine-tuning transformer-based models.

Keywords: Dialogue Systems, Neuro-Symbolic AI, Heart Failure

1. Introduction

The excessive consumption of salt poses significant public health risks, contributing to diseases such as high blood pressure and heart failure (He et al., 2020). Reducing salt intake has been shown to mitigate these health issues. In 2017, excessive sodium intake was associated with around three million deaths and a significant loss of healthy life years (Roth et al., 2018). Research, including clinical trials and population studies, supports the reduction of salt intake as a means to manage and prevent these conditions. Despite the clear benefits of sodium reduction, public understanding and action are lacking; only 58% of individuals can accurately read sodium content on nutrition labels, and merely 44% can classify food products as high or low in sodium based on standard labeling (Dickson and Riegel, 2009). This gap in knowledge and practice underscores the challenge of addressing dietary sodium intake, with only a handful of countries implementing effective public health interventions.

Therefore, we aim to develop a dialogue system that enables patients to inquire about the salt content in various foods. This system especially aims to support heart failure patients, who must meticulously monitor and reduce their salt intake. More specifically, African American individuals who are more prone to heart failure (Nayak et al., 2020), have a higher sensitivity to salt and face challenges like food deserts and higher consumption of junk foods. This necessitates a specialized dietary management approach to help them monitor and reduce their salt intake effectively. Furthermore, in (Gupta et al., 2020), the authors show that African American patients with heart failure often focus on discussions related to 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. Having said that, the dialog system can be used by anyone who wants to inquire about the salt content in food.

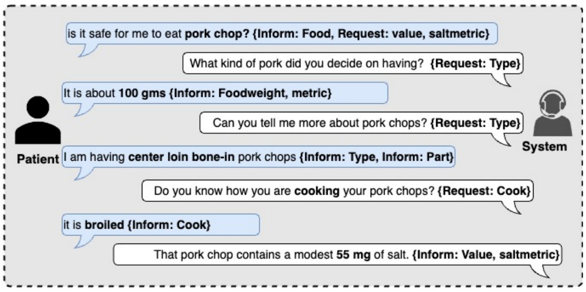

Figure 1: Sample Template Conversation which begins with the user asking about the salt content in food. The system asks clarification questions to determine the precise food item and its salt value.

Creating a dialog agent specialized in food-related conversations and nutrient information is challenging. This is primarily due to the lack of a conversational dataset specifically designed for this domain. Moreover, annotating the dataset is very costly and resource-intensive. To address this issue, we create a template-based conversational dataset (an example is shown in Figure 1) to identify various food items and their salt content. Our approach involves utilizing the USFDC (U.S. Food Data Central) (USFDC, 2022) dataset, which provides detailed food descriptions along

Proceedings of CL4Health 2024 @LREC-COLING-2024, pages 93–103

with their nutrient values. To enhance the system's ability to recognize the different components detailed in the food descriptions, we developed a food ontology. This ontology is constructed using the FoodOn (Dooley et al., 2018) framework and GloVe (Pennington et al., 2014) embeddings, facilitating the identification of various attributes related to the food's cooking and preparation methods. Leveraging this ontology, we create a structured food conversational dataset (Figure 1). As it is a template-based dataset, annotating it is easy and effective. We model the dataset after the state-of-the-art task-oriented dialog dataset MultiWOZ (Budzianowski et al., 2018).

To train the dialog system, we use the end-to-end dialog system PPTOD (Plug-and-Play Task-Oriented Dialogue System) Model (Su et al., 2022). PPTOD extends the T5 framework, especially designed for task-oriented dialogue (TOD) applications. To accurately provide salt content information, our dialog system will address vague user queries, (such as "What is the salt content in pork chops?"), by asking clarification questions. This approach ensures the model understands the specific preparation and consumption context of the food item, allowing us to determine the precise salt content based on how the food is prepared (For ex - beef can be consumed raw, cooked, or as part of a dish like a burger). The dataset along with the trained model is publicly available.[1]

Despite the remarkable proficiency of large pre-trained language models (PLMs) like GPT-3 and T5 (Brown et al., 2020; Raffel et al., 2020) in complex arithmetic reasoning tasks, they occasionally make calculation errors, especially as the mathematical operations in equations become more complex (Wei et al., 2022). Our observations align with these findings, notably that even after fine-tuning, the PPTOD model struggled to compute the salt values for various food items. Moreover, the USFDC provides salt content for standard food measurements, and users may not frame their queries in these standard terms. To rectify this, we propose **NS-PPTOD**, where we integrate PPTOD model with neuro-symbolic rules. These rules are designed to harness the strengths of PLMs while compensating for their weaknesses, enabling the system to retrieve and accurately calculate the salt content from the database. This ensures the system's adaptability in responding to queries about salt content in both standard and non-standard food quantities, thereby offering accurate salt content information and enhancing the system's overall accessibility and effectiveness.

Our experiments show that just fine-tuning a transformer model to predict salt content isn't enough. The integration of neuro-symbolic rules significantly enhances the system's performance, evidenced by a 20% improvement in joint goal accuracy across different dataset sizes. This proves that combining pre-trained language models with neuro-symbolic rules is essential for better accuracy.

In summary, our contributions are

- We propose to develop a food conversation dataset that includes clarifying questions to infer the correct food item and its salt content.
- We finetuned PPTOD on our food conversation dataset using a few-shot approach.
- We propose NS-PPTOD which integrates PPTOD with Neuro-Symbolic rules to infer correct salt values across different food weights.
- We show a 20% increase in joint accuracy compared to the finetuned PPTOD.

2. Related Work

- **HealthCare Dialog Systems** Task-oriented dialogue systems have seen a significant rise in the healthcare sector, where they play a vital role in enhancing various aspects of healthcare. These systems are developed for a wide array of diseases including heart failure (Moulik, 2019; Gupta et al., 2020), cancer (Belfin et al., 2019), mental disorders (Ali et al., 2020), public anxiety (Wang et al., 2020) etc. Their applications extend to several areas, including disease diagnosis (Wei et al., 2018), patient education (Cai et al., 2023; Gupta et al., 2020), and health coaching (Zhou et al., 2022) among others. A comprehensive survey of NLP literature conducted 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.

  (Gupta et al., 2020; Salunke et al., 2023) discuss the development of a dialog agent for self-care needs of heart failure patients, drawing upon insights from educational sessions. The work in (Kearns et al., 2020) explores the Wizard of Oz (WOZ) technique to craft a persona-based health counseling dialog dataset. Additionally, recent advancements have seen the application of Large Language Models (LLMs) in responding to patient inquiries (Chowdhury et al., 2023), though the importance of safety is emphasized. Addressing the limitations in the medical knowledge of LLMs, the study in (Li et al., 2023) undertakes the task of enhancing and fine-tuning the LLaMa model with

[1] https://github.com/anujatayal/NS-Monitoring-Salt-Content-in-Food

a dataset of approximately 100,000 patient-doctor dialogues.

- **Pretrained Language Models (PLMs)** With the advancement in pre-trained language models (PLMs), different systems based on PLMs have been proposed including dialog systems (Lei et al., 2018; Peng et al., 2021). PLMs excel in various tasks, approaching human-like performance. Yet, they struggle in mathematical reasoning, as noted in (Wei et al., 2022). (Qian et al., 2023) shows the limitations of LLMs with complex or lengthy numerical operations. For instance, GPT-3 (Brown et al., 2020) performs well in simple two-digit additions but falters with longer numbers. Similarly, even a fine-tuned T5 model struggles with the accurate addition or subtraction of lengthy numbers (Nogueira et al., 2021), and the challenge escalates with numbers not covered in their training data.

- **NeuroSymbolic AI** Integrating neuro-symbolic approaches offers a solution by combining the inference capabilities of symbolic systems with the robustness of neural networks, creating a composite AI framework adept at reasoning, learning, and cognitive modeling (Garcez and Lamb, 2023). This blend addresses the inherent weaknesses of each system, promising enhanced performance and robustness.

  To address the generalization issues in neural networks, particularly in task-oriented dialogue systems, various neuro-symbolic methodologies have been investigated. (Mehri and Eskenazi, 2021) proposes schema graphs to generalize across various unseen domains and tasks. In (Romero et al., 2021), the authors fine-tuned GPT-2 to generate the text and symbolic representations. DILOG (Zhou et al., 2020) employed inductive reasoning to formulate logical rules, enabling dialog policy training with minimal data to facilitate zero-shot domain transfer. (Arabshahi et al., 2021) used a neuro-symbolic approach to extract multi-hop reasoning and integrate commonsense in a dialog system. These strategies underscore the potential of neuro-symbolic integration to significantly improve the adaptability and efficacy of language models in complex and dynamic tasks.

- **Representing Food in Dialog Systems** Addressing the intricacies of food representation, FoodKG (Haussmann et al., 2019; Chen et al., 2021) explored knowledge graphs to represent food. FoodKG (Haussmann et al., 2019) integrates information from diverse recipe collections and the US Food Data Central (USFDC, 2022), primarily focusing on template-based queries related to ingredients and recipes. However, this framework has limitations, notably in identifying only the primary item in food descriptions and missing key details like the type of food, cooking methods, or quantity, which are included in our methodology. (Fu et al., 2022) explored the role of recommending food to improve mental health. RecipeQA (Yagcioglu et al., 2018) explored multimodal question answering within the context of recipes, while CookDial (Jiang et al., 2022) provides a platform for users to navigate and query cooking recipes more effectively.

3. Dataset Creation

Given the absence of a specialized dataset for conversational inquiries about salt content and the challenges in dataset collection and annotation, in this section, we show in detail how we created the dataset. Dataset creation involves developing a template-based conversational framework to accurately identify food items and their salt content. First, we used the USFDC dataset (USFDC, 2022), and created an ontology using FoodOn (Dooley et al., 2018) and Glove (Pennington et al., 2014) to describe the different components in the food description. By doing so, we were able to distinguish between different slot values. Using the ontology, we created a template-based conversational dataset that mimics human conversation, alternating between the user and the system turns. We define an average of 7 slots while creating the dataset namely- food, cook, type, animal, part, foodweight, metric.

Data Source To construct the dataset, we leverage the extensive food descriptions and nutritional data from the USFDC database (USFDC, 2022). It is renowned for its broad representation of diverse food items and is publicly available. The dataset was created with careful consideration of cultural differences, sourcing its data from the U.S. Department of Agriculture (USDA).

Each food description in the USFDC database consist of unstructured, comma-separated text detailing ingredients, cooking methods, and cutting styles. This format lacks clarity on the significance of each component as demonstrated in Table 6 of Appendix B. To address this, we concatenated these segments using underscores, transforming each into a distinct entity to enhance data clarity and interpretation.

Furthermore, we faced the difficulty of distinguishing whether a food item is a primary ingredient or as part of a larger dish. Items like lettuce and cheese, for example, can be both independent food items

key	Questions
food	What is the {nutrient} content in {food} ? How much {nutrient} in {food}? What is the {nutrient} content in {cook} {food}? How much {nutrient} in {foodWeight} {metric} of {food}? Can my partner with heart issues eat {food}? Is {food} okay for heart patients?

Table 1: Sample Template Questions that user asks to begin the conversation

and components in recipes like pizzas or burgers. To overcome this challenge, we developed a food ontology. This ontology aids in categorizing each food item more accurately, thus improving the overall understanding of the dataset.

Ontology Construction To develop the food ontology, an initial framework is established using FoodON (Dooley et al., 2018), focusing on key relations of *food*, *cook*, *animal* and *part*. These relations were chosen based on their significant impact on altering the salt content in various foods. Moving forward, these specific relations will be utilized to aid in creating and annotating the conversational dataset and pose clarifying questions to users to infer the salt amount in food. To address the limitations in the comprehensiveness of this initial ontology, pre-trained GloVe vectors (Pennington et al., 2014) are utilized to identify words similar to those in the ontology, thus expanding its scope. However, this method inadvertently introduces some items unrelated to food, necessitating manual preprocessing to eliminate irrelevant elements and maintain a focus on food context.

Further refinement of the ontology was needed to incorporate the items that do not associate with existing relations. A new relation, *type*, was created to integrate these components (For example type of cuisine, meat, other food ingredients etc). Drawing inspiration from FoodKG (Haussmann et al., 2019), the first item in each food description is categorized under the *food* relation. This enhanced ontology becomes a valuable tool for mapping each component of the comma-separated food descriptions to the relevant keys. In instances where multiple segments pertain to the *type* relation, their values are concatenated to ensure consistency and clarity.

3.1. Conversational Dataset Creation

Using the ontology and the food descriptions, we aim to develop a template-based conversational dataset that mimics human conversation. The conversation initiates with a user query about the salt content in a specific food item and alternates between the user and the system. The system poses clarification questions, drawing from ontology relations such as the type of food, cooking method, and portion size, which are crucial determinants of salt content.

Leveraging the task-oriented dialogue framework, each turn t is annotated with a belief state B_t, encompassing a list of slot-value pairs and action states *inform* and *request*. Figure 1 illustrates a sample conversation highlighting the belief state and action state for each turn t. The figure demonstrates that within a single turn, it is possible to fill multiple slot values (the 3rd turn of the user). The dataset's format and annotations draw inspiration from the advanced task-oriented dialog dataset, MultiWOZ (Budzianowski et al., 2018). Employing MultiWOZ as a benchmark not only validates the dataset but also enhances its replicability for crafting conversational datasets for other nutrients. Using template-based approach streamlines the annotation process, ensuring uniformity and efficiency, and reducing the cost and time needed for external annotators.

	Dialogue Statistics
# Dialogues	87,425
# Total turns	525,392
Avg turns per dialogue	6
Avg slots	7

Table 2: Dialogue Statistics of the template based conversation data

We consider 3 types of turns in a conversation. We have tried to encapsulate the range of dynamics that can occur in dialogues, ensuring the system is robust enough to handle the fluidity of human conversation.

- **Matching Answers**- This type involves turns where the user's reply is directly pertinent to the system's question regarding a particular slot value. An example is when the system queries about the cooking method, and the user responds specifically about the food's cooking method (the 4th turn of the user in Figure 1).

- **Random Answers**- Occasionally, a user's reply may not correspond to the query posed by the system. For instance, if the system asks about the *food type* and the user responds with information about the *weight* of the item, as demonstrated in the user's second turn in Figure 1, the system needs to adapt. In such situations, the system should recognize and not repeat a question about the weight, since that information has already been provided. Instead, it should proceed to ask another question, possibly continuing to seek details about the *food type*. In .45% of conversations, a turn consists of a random answer.

- **Changing Answers**- People are very indecisive and often tend to change their responses. This category captures the scenarios where users might revise their previous responses. For example, if a user initially mentions that the cooking method is "pan-fried" but later changes it to "boiled," the system needs to update its understanding to reflect this new information, shifting its belief state from "pan-fried" to "boiled." In .45% of conversations, one user turn involves changing the answer.

The dataset is created by generating random conversations, in which, at each turn, the system's questions and the user's responses are randomly selected from the templates. The conversation starts with the user asking about the salt content in food. This initial question is informed by a user study of HFChat (Salunke et al., 2023), where participants frequently asked 3 categories of questions: 1) how much salt in {food} 2) Can I eat {food}? and 3) what kind of {food} can I eat?. These question types, along with similar ones, constitute the initial question, as exemplified in Table 1. The system's objective is to engage in the dialogue by asking clarification questions to ascertain the values of different slots (cook, type, weight etc). To keep the dialogue dynamic and realistic, questions related to these slots are presented randomly. The number of questions is limited; for instance, if the food in question is eaten raw, queries about cooking methods are omitted. Users might not be aware of all system-initiated questions, in which case default values for each slot are assumed. This led to the creation of approximately 87k template-based conversations, each comprising 3-4 exchanges between the user and the system. The statistics of this extensive dataset are detailed in Table 2. As the dataset size is very large, around 87k conversations, a few-shot method is used to train the model.

4. Methodology

Once the conversational dataset was created, we built the NS-PPTOD model by fine-tuning PPTOD on the dataset using few-shot and integrating neuro-symbolic rules.

4.1. Plug and Play Task Oriented Dialog System (PPTOD)

Leveraging the T5's model success, we adopted the PPTOD model for developing the task-oriented dialogue system. PPTOD extends the T5 framework, especially designed for task-oriented dialogue (TOD) applications, and pre-trained on a diverse range of dialog datasets spanning eleven different domains. PPTOD has integrated different TOD modules — Dialogue State Tracking (DST), Natural Language Understanding (NLU), Dialogue Policy (POL) and Natural Language Generation (NLG) —into a single end-to-end architecture using a pipeline approach.

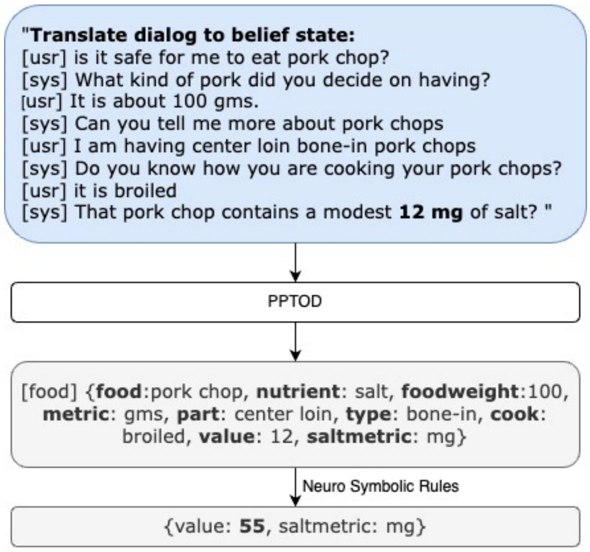

Figure 2: NS-PPTOD Model with Example: This example illustrates that PPTOD alone is not able to identify correct salt value for the food item

The PPTOD model is adept at in-context learning, employing customized prompts for each dialogue component, thus enhancing the relevance of model inputs to specific dialogue scenarios. Each training sample in PPTOD is represented as (x_t, y, z), where $t \in$ NLU, DST, POL, NLG specifies the type of Task-Oriented Dialogue (TOD) task to which the sample belongs. The term x_t refers to the task-specific prompt, formatted as *translate dialogue to A*:, where A represents different aspects such as *user intent*, *belief state*, *dialogue act*, and *system response* corresponding to the NLU, DST, POL, and NLG tasks respectively. The input dialogue context, a concatenation of all preceding dialogue utterances, encompassing both the system's and the user's contributions is represented by y whereas z indicates the target output text. This is exemplified in Figure 2. PPTOD is trained with a maximum likelihood objective and the loss function as shown in Equation 1 where θ is the model parameters.

$$L_\theta = -\sum_{i=1}^{|z|} \log P_\theta(z_i | z_{<i}; x_t, y) \tag{1}$$

PPTOD also has an associated tokenizer, which supports a few-shot learning framework, enabling the system to identify new food-related terms not seen in training. We have fine-tuned it on only 1000 food-related dialogs using the same learning objective as PPTOD.

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

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

We chose PPTOD model for its adaptability, and its ability to support modular decomposition which in turn facilitates the incorporation of specific rules. To further enhance the model, we incorporated neuro-symbolic techniques into the DST framework, aiming to boost the system's interpretative capabilities and its adaptability to intricate dialogue contexts. This integration seeks to fortify the dialogue system, ensuring it remains robust and flexible in managing diverse conversations.

In summary, we chose PPTOD because:

- PPTOD is a state-of-the-art (SOTA) model designed specifically for task-oriented dialogue and is based on the T5 model.
- PPTOD consists of a tokenizer making it possible to use few-shot approach to fine-tune it.
- PPTOD can be modularised to implement neuro-symbolic AI.

4.2. NS-PPTOD

NS-PPTOD is an integration of PPTOD with neuro-symbolic rules. The conversational dataset we created as described in section 3.1 is used to finetune PPTOD. Given the extensive size of the dataset of 87k template-based conversations, a few-shot learning approach is used. Instead of utilizing the entire dataset, limited subsets of samples are selected randomly to finetune PPTOD. As we use a few-shot approach, T5-small is used as the base model to train PPTOD. PPTOD model is finetuned for 10 epochs, employing a batch size of 16 and varying the total dataset size across 100, 300, 500, and 1000 samples. Within the dataset, 10% of the data was allocated as a development set, and another 10% served as the test set.

During the experiments - to be discussed in Section 5, we observed that the model correctly identified slot values but struggled to determine the correct salt values. Additionally, the dataset primarily comprised salt values for standard food weights, like 100 grams, 3 ounces, 1 packet, etc., and lacked data for non-standard food weights that users might inquire about. PPTOD model is fine-tuned so that it accurately learns other slot values, even if it also learns some incorrect salt values. Subsequently, we employ a neuro-symbolic approach that involves two key methodologies to correct the salt values:

- **Retrieval of the accurate salt value from the database:** This step is crucial for standard food weights where exact values are available and can be directly obtained. (as shown in Figure 2)
- **Mathematical calculation of the correct salt value for varying food weights:** This method is particularly beneficial for non-standard food weights, enabling the model to compute salt values based on weight. (In Figure 2, if the user requests the salt value for a different food weight, instead of the standard 100 grams of pork chops.)

Specifically, upon determining the slot values, the system queries the database for the salt content. If the database contains the salt value, it is then retrieved. In cases where the salt value is not available in the database, it is calculated mathematically, based on the weight of the food. We demonstrate this in Figure 2. When the dialog context and prompt, labeled *Translate dialog to belief state*, are processed through PPTOD's DST model to determine belief states, the model successfully infers all slot values except for the salt value (12). The correct salt value (81) is then retrieved from the database by applying Neuro-Symbolic rules.

5. Evaluation

The evaluation of NS-PPTOD encompasses two task-oriented tasks. The first task involves end-to-end dialog modeling, assessed using metrics such as inform rate, success rate, and BLEU score (Papineni et al., 2002). The second task, the Dialog State Tracking (DST) module of PPTOD is evaluated through joint-accuracy. This evaluation was conducted over sample sizes of 100, 300, 500, and 1000.

End-to-End Dialog Modeling After seven epochs of training, the PPTOD model demonstrated a high inform accuracy rate of 93.5% across

1000 samples, as detailed in Table 3. Inform rate reflects the model's adeptness in identifying slot values and the target goal slot. However, its success rate in accurately determining correct salt values was notably low, standing at just 2.7%, a point further highlighted in Table 3. This pattern of low success rate was consistent across other training sizes of 100, 300, and 500 samples. The limited 2% success rate is attributed to the PPTOD model's tendency to predict values at random.

Implementing the NS-PPTOD model resulted in a substantial enhancement, achieving success rate of 71.7%. This improvement was not just limited to the training size of 1000 samples but was also observed consistently across the smaller training sizes. The integration of neuro-symbolic rules with PPTOD evidently plays a crucial role in enhancing the model's capability to accurately predict and determine the correct salt values.

DST Module The evaluation also included the Dialog State Tracking (DST) module of PPTOD, trained for recognizing different belief states such as food, cook, type, weight, and value. Joint accuracy of 58.53% was achieved when PPTOD was used. This performance notably increased to 85.2% for 1000 samples with the addition of neuro-symbolic rules. Similar improvements in joint accuracy were observed for other training sizes. Table 4 displays the enhanced joint-accuracy achieved by using NS-PPTOD compared to PPTOD.

Analysis The improvement in both success rate and joint accuracy can be attributed to a key difference in approach. PPTOD, on its own, tends to memorize values instead of effectively retrieving them from the database, a critical process for accurately determining salt content. However, the application of neuro-symbolic rules in conjunction with PPTOD enhances its capability, enabling it to effectively retrieve values from the database.

5.1. Comparison with ChatGPT

With the advent of ChatGPT, questions have arisen about the necessity of systems such as our NS-PPTOD. To address this,we conducted a comparative analysis between NS-PPTOD and ChatGPT, to highlight the distinct capabilities and applications of NS-PPTOD that are not achievable by ChatGPT.

In this comparison, we specifically focused on their responses to queries about the salt content in foods. Appendix A illustrates ChatGPT's response to the prompt *What is the salt amount in a pork chop?*. ChatGPT's responses are generally comprehensive, explaining variations in salt quantity due to different cooking methods and weights, and often include an average value. NS-PPTOD, conversely, poses targeted clarification questions to precisely identify both the food item and its salt content.

There is also a marked difference in the readability of responses from these two systems. Readability assessments, SMOG (Mc Laughlin, 1969), Flesch-Kincaid Grade Level (FKGL) (Kincaid et al., 1975) , and Flesch-Kincaid Reading Ease (FKRE) (Kincaid et al., 1975), which evaluates text comprehensibility in educational and informational contexts, were used for this comparison. The SMOG and FKGL estimates the education level needed to understand a text, while the FKRE measures the ease of comprehension.

As indicated in Table 5, understanding ChatGPT's content requires an 11th-grade reading level, whereas NS-PPTOD's content is accessible to those at a 2nd-grade reading level. NS-PPTOD's responses align with recommendations for content to be understandable at a 5th-grade reading level or lower (Badarudeen and Sabharwal, 2010). This approach is particularly suitable for the intended audience of our system, the patients served by the hospital at our institution who are mostly of lower socio-economic status (and mostly minorities); this ensures the information is accessible and understandable to a wider audience.

	SMOG	FKGL	FKRE
ChatGPT	11.8	11.1	51.28
NS-PPTOD	0	2.5	95.17

Table 5: Readability Comparison of NS-PPTOD with ChatGPT: A higher FKRE score denotes better readability. Lower scores in SMOG and FKGL indicate better text readability.

6. Conclusion and Future Work

This work introduces NS-PPTOD, a conversational dialogue system specifically designed to help patients, particularly African-American individuals suffering from heart failure, in accurately identifying the salt content in various foods. Managing salt intake is crucial for these patients, and NS-PPTOD offers a specialized solution to aid in their dietary management. The study reveals that simply applying fine-tuning techniques to a transformer model for predicting salt content is inadequate for achieving high accuracy. However, the integration of neuro-symbolic rules into the system resulted in a significant performance boost, enhancing accuracy by 20% across different dataset sizes. This enhancement underscores the necessity of combining pre-trained language models with neuro-symbolic rules to achieve superior accuracy.

Looking ahead, there are plans to conduct an in-depth study comparing the performance of NS-PPTOD with a system based on ChatGPT. This study will involve African American patients with heart failure interacting with both systems in a within-subject setting. Utilizing pre and post-interaction surveys, the aim is to conduct both qualitative and quantitative analyses. These analyses will evaluate and contrast the effectiveness of each system offering insights into their respective impact in real-world scenarios.

7. Acknowledgement

For financial support, we thank the DPI Cycle 1 Seed Funding Program Award and NSF award IIS 2232307

8. Bibliographical References

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.

Sameer Badarudeen and Sanjeev Sabharwal. 2010. Assessing readability of patient education materials: current role in orthopaedics. *Clinical orthopaedics and related research*, 468(10):2572–2580.

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.

Yu Chen, Ananya Subburathinam, Ching-Hua Chen, and Mohammed J. Zaki. 2021. Personalized food recommendation as constrained question answering over a large-scale food knowledge graph. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, WSDM '21, page 544–552, New York, NY, USA. Association for Computing Machinery.

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.

D.M. Dooley, E.J. Griffiths, and G.S Gosal. 2018. Foodon: a harmonized food ontology to increase global food traceability, quality control and data integration. *npj Science of Food*, 2.

Chengcheng Fu, Zhisheng Huang, Frank van Harmelen, Tingting He, and Xingpeng Jiang. 2022. Food recommendation for mental health by using knowledge graph approach. In *Health Information Science: 11th International Conference, HIS 2022, Virtual Event, October 28–30, 2022, Proceedings*, page 231–242, Berlin, Heidelberg. Springer-Verlag.

Artur d'Avila Garcez and Luis C Lamb. 2023. Neurosymbolic ai: The 3 rd wave. *Artificial Intelligence Review*, pages 1–20.

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.

Steven Haussmann, Oshani Seneviratne, Yu Chen, Yarden Ne'eman, James Codella, Ching-Hua Chen, Deborah L. McGuinness, and Mohammed J. Zaki. 2019. Foodkg: A semantics-driven knowledge graph for food recommendation. In *The Semantic Web – ISWC 2019*, pages 146–162, Cham. Springer International Publishing.

Feng J. He, Monique Tan, Yuan Ma, and Graham A. MacGregor. 2020. Salt reduction to prevent hypertension and cardiovascular disease. *Journal of the American College of Cardiology*, 75(6):632–647.

Yiwei Jiang, Klim Zaporojets, Johannes Deleu, Thomas Demeester, and Chris Develder. 2022. Cookdial: a dataset for task-oriented dialogs grounded in procedural documents. *Applied Intelligence*, pages 1–19.

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.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel.

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.

G Harry Mc Laughlin. 1969. Smog grading-a new readability formula. *Journal of reading*, 12(8):639–646.

Shikib Mehri and Maxine Eskenazi. 2021. Schema-guided paradigm for zero-shot dialog. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 499–508, Singapore and Online. Association for Computational Linguistics.

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.

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.

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.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

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.

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.

Gregory A Roth, Degu Abate, Kalkidan Hassen Abate, Solomon M Abay, Cristiana Abbafati, Nooshin Abbasi, Hedayat Abbastabar, Foad Abd-Allah, Jemal Abdela, Ahmed Abdelalim, et al. 2018. Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980–2017: a systematic analysis for the global burden of disease study 2017. *The Lancet*, 392(10159):1736–1788.

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.

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.

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.

Semih Yagcioglu, Aykut Erdem, Erkut Erdem, and Nazli Ikizler-Cinbis. 2018. RecipeQA: A challenge dataset for multimodal comprehension of cooking recipes. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1358–1368, Brussels, Belgium. 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.

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. Appendix: Sample ChatGPT Response

ChatGPT response to the prompt **How much salt in pork chops?**

The salt content in pork can vary significantly based on how it is prepared. Fresh, unprocessed pork naturally contains a small amount of sodium, roughly 60 to 70 milligrams of sodium per 3-ounce serving. However, processed pork products like bacon, ham, sausages, and deli meats can have a much higher sodium content due to added salt during processing for flavor and preservation. For *instance, a slice of bacon can have between 150 to 450 milligrams of sodium, and a serving of ham can contain over 1,000 milligrams.*

For specific sodium content, it's best to check the nutrition labels on packaged pork products. For fresh pork, nutrition databases or guidelines from health organizations can provide general estimates. Remember, if you're watching your sodium intake, opting for fresh, unprocessed pork and limiting processed pork products can help manage your sodium consumption.

B. Appendix: Sample Food Descriptions

Food Descriptions	Salt Value
Pork, fresh, loin, top loin (chops), boneless, separable lean and fat, raw	48
Pork, fresh, loin, center loin (chops), bone-in, separable lean and fat, cooked, broiled	55
Pork, fresh, blade, (chops), boneless, separable lean and fat, cooked, broiled	58
Pork, fresh, loin, sirloin (chops or roasts), boneless, separable lean only, raw	63
Pork, fresh, loin, blade (chops), bone-in, separable lean only, cooked, broiled	76

Table 6: Unstructured Food Descriptions along with their salt value for 100gms of food