A Neuro-Symbolic Approach to Control Task-Oriented Dialog Systems

Anuja Tayal

UIN: 657111812

Department of Computer Science

Ph.D. Thesis Proposal

June 11 2025

Committee:

Prof Barbara Di Eugenio

Prof. Natalie Parde

Prof. Debaleena Chattopadhyay

Prof. Sourav Medya

Prof. Olga Garcia-Bedoya

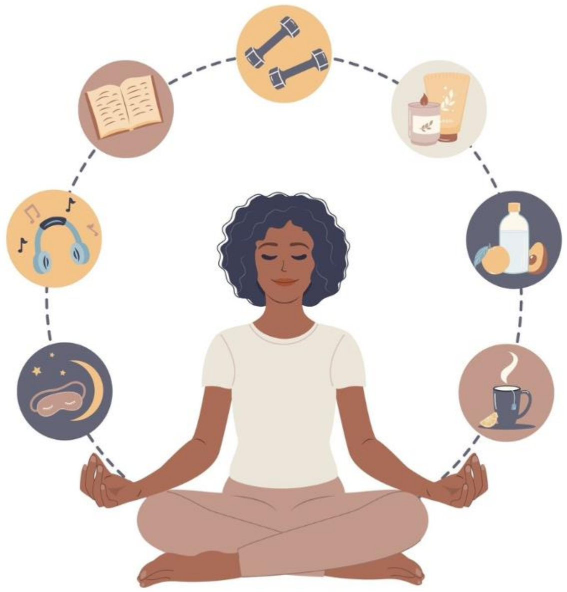
Heart Failure

- Nearly 50% of Americans have a chronic disease
- Chronic illnesses last 1+ year and greatly affect life
- Heart Failure (HF) – chronic illness - heart can't pump enough blood
- Highest readmission: patients 65+
- HF patients must self-care

[1] Lewsey, S. C. and Breathett, K.: Racial and ethnic disparities in heart failure: Current state and future directions. Current Opinion in Cardiology, May 2021.

Self-Care

- Self-care means managing symptoms, treatment, emotions, and lifestyle changes to maintain a satisfactory quality of life for as long as possible. [1]
- Traditional self-care design was medically focused, not patient experience
- Patients often lack knowledge and depend on others



[1] Toukhsati S, Jaarsma T, Babu A, Driscoll A, Hare D. Self-Care Interventions That Reduce Hospital Readmissions in Patients With Heart Failure; Towards the Identification of Change Agents. Clinical Medicine Insights: Cardiology. 2019;13. doi:10.1177/1179546819856855

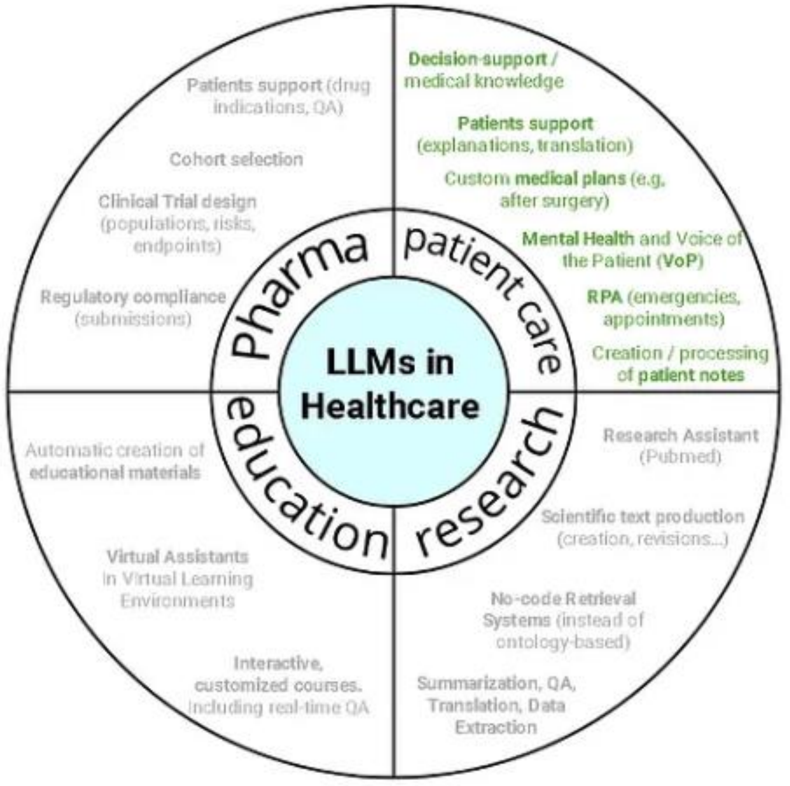
Minorities face worst outcomes

- Minority groups (African American (AA) and Hispanic/Latino (H/L)) face worse outcomes
- Causes:
  - genetics,
  - healthcare gaps
  - socioeconomic factors
  - low health literacy
- Self-care resources often lack cultural relevance for minorities

[1] Nayak, A., Hicks, A., and Morris, A.: Understanding the complexity of heart failure risk and treatment in black patients. Circulation: Heart Failure, 2020.

HealthCare NLP

- NLP offers solutions in healthcare
- NLP for self-care, patient education not much explored
- Addressing racial and cultural disparities
- Challenges
    - Limited datasets due to IRB
    - Evaluation Challenges



LLMs in Healthcare: Patient Care

AIM

- Explore conversational architectures that deliver self-care information to AA HF patients

Table of Content

Motivation

- Understand how PEs convey self-care strategies
- Patients spoke very little
- **Spans multiple domains**: salt intake, exercise, fluid management, medical adherence, understanding of condition

Speaker	Utterance
Patient	Yeah, I don't, I dont't do the frozen meal.
Educator	Okay.
Patient:	I was basically doing the uh, vegetables.
Educator:	Okay.
Patient:	Frozen vegetables,
Educator:	They should be fine.
Patient:	Yeah.
Educator:	But... but, I do want you to start looking at those nutrition labels.
Patient:	Okay.
Educator:	And look for something that says less than 5%
Patient:	Okay.
Educator:	So, the other we always want you to do is, um, of course take all your medicines like you're supposed to.
Patient:	Which I didn't do last night.
Educator:	Okay.

Patient-Educator Session

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. Association for Computational Linguistics, Online, 41–46.

AIM

- Explore conversational architectures that deliver self-care information to AA HF patients
- Not a traditional dialog agent or QA system
  - Supports multi-turn interactions
  - Patient takes initiative
  - The agents asks clarification questions
- **Focus**: Salt Intake and Exercise

Just use an LLM…?

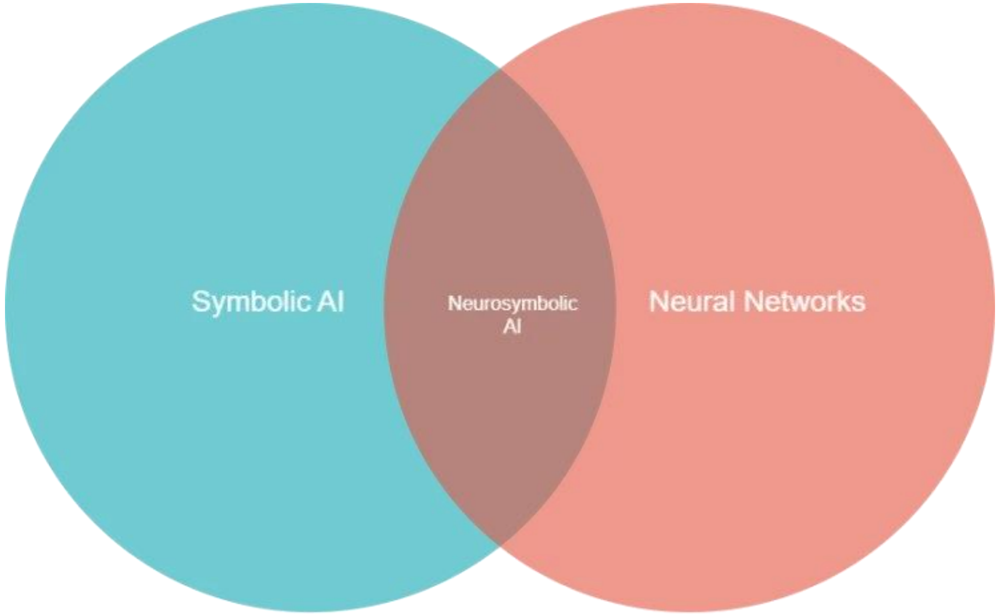
- Offer unique advantages:
  - Contextual understanding
  - Scalability across diverse datasets
  - Shown strong potential in generating synthetic datasets
- Hallucinate
- Difficult to ensure that they stay within the boundaries of specialized domains
- Often have high readability levels- trained on extensive web data, PubMed articles
- Make it more accessible
- Struggle with mathematical reasoning

Need for Control

- Limitations are especially concerning in healthcare dialogue systems.
- Prompting
    - Few-shot prompting, chain of thought prompting
- No guarantee that the models will adhere to all the prompt instructions

Solution: Integrating Neuro-Symbolic Approaches

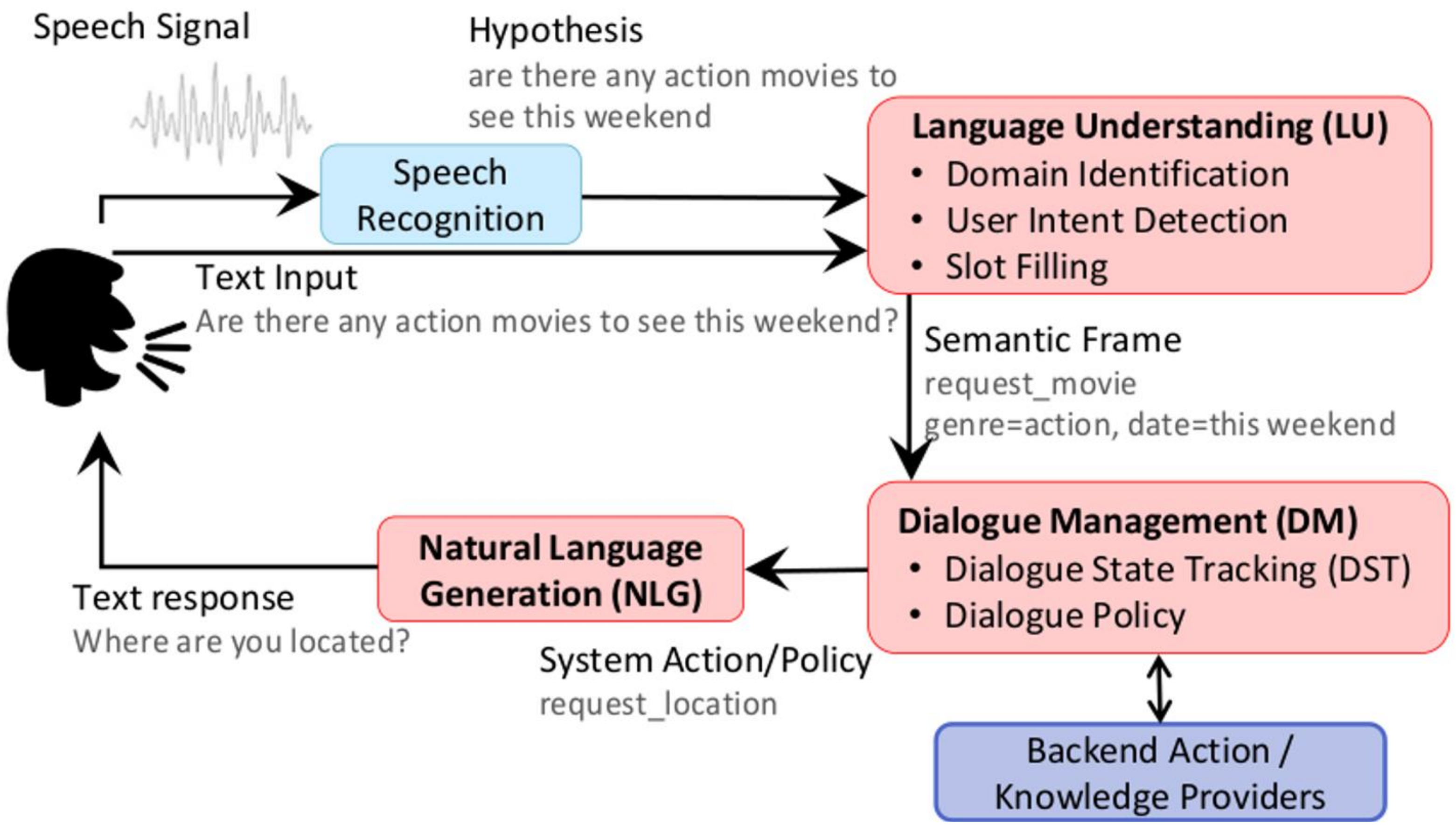
- combining the inference capabilities of symbolic systems with the robustness of neural networks
  - System 1, which is fast, intuitive, and associative (akin to large language models)
  - System 2, which is slower, more deliberate, and logical, represents the symbolic reasoning component.



AIM

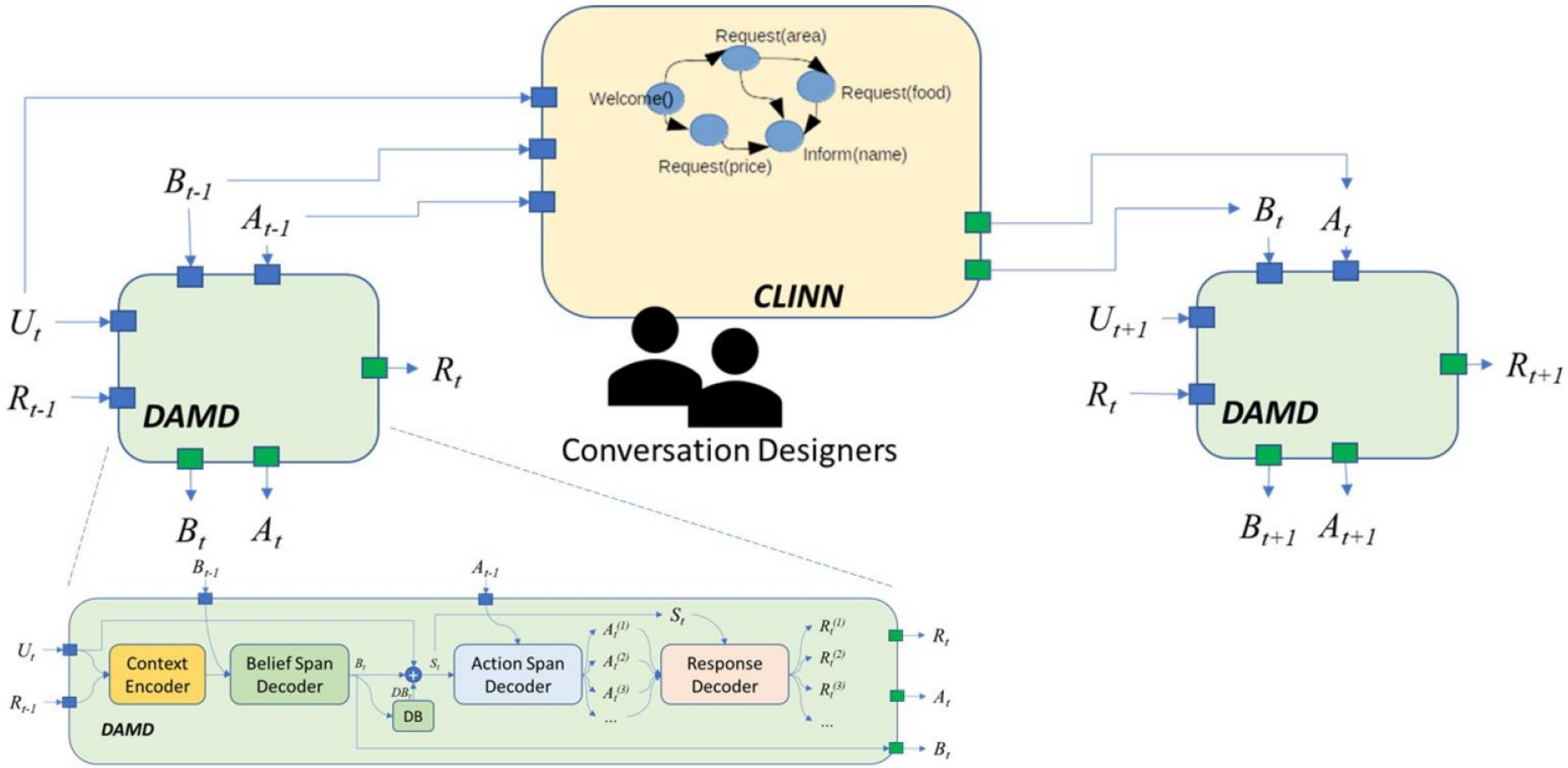
- Explore **conversational architectures** that deliver self-care information to AA HF patients
    - Able to add neuro-symbolic rules
    - Add more explainable
    - Make more controllable

Task Oriented Dialog System (TODS) Pipeline



Task Oriented Dialog System Pipeline

Example



Neuro-Symbolic Architecture

AIM

- Explore conversational architectures that deliver self-care information to AA HF patients
    - Explore a **hybrid task-oriented dialogue model** that combines the strengths of task-oriented systems and language models (LMs/LLMs).
    - The task-oriented dialogue system **incorporates neuro-symbolic rules**, enhancing interpretability by providing clearer insight into the system's understanding and enabling more thorough error analysis

Research Questions

- **Data**: What are the different ways to prompt LLM to generate synthetic conversations in the absence of patient-oriented self-care dialogues, and is prompting enough to control/ personalise the conversations?
- **Methodology**: How can we effectively combine the strengths of task-oriented dialog systems (TODS) and LMs/LLMs to create a hybrid dialog model, and integrate neurosymbolic rules with the LM/LLM-based dialog system to control the dialog system.
- **Evaluation**: How do heart failure AA patients perceive and interact with a neurosymbolic TODS compared to an LLM-based system?

Patient Educator Dialogs

- Patients spoke very little
- Half of patients responses were filler words
- Remaining primarily focused on salt and dietary habits

Speaker	Utterance
Patient	Yeah, I don't, I dont't do the frozen meal.
Educator	Okay.
Patient:	*I was basically doing the uh, vegetables.*
Educator:	Okay.
Patient:	Frozen vegetables,
Educator:	They should be fine.
Patient:	Yeah.
Educator:	But... but, I do want you to start looking at those nutrition labels.
Patient:	Okay.
Educator:	And look for something that says less than 5%
Patient:	Okay.
Educator:	So, the other we always want you to do is, um, of course take all your medicines like you're supposed to.
Patient:	Which I didn't do last night.
Educator:	Okay.

Excerpt of Patient-educator session

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. Association for Computational Linguistics, Online, 41–46.

HFChat – Dialog Flow dialog agent

- Gupta and Salunke collected 16 Q/A pairs with medical professionals.
- First Iteration of Dialog system, was built using DialogFLow
- Evaluation:
  - User study was conducted
  - 14 participants (12 AA, 2H/L)
  - Semi-structured interview followed by interaction by HFChat
  - 35% of participants were unable to name their condition
  - Able to name their conditions : asked questions about comorbidities
  - Some were concise while others prompted HFChat to ask follow-up or cq questions
- Limited to QA format
- Relied on a dataset of 16Q/A pairs

Assess Reading level

- NIH recommendation
  - writing health information at a 6th-7th grade level
- Assessed the reading level of responses generated by ChatGPT

Reading Level	HFChat		ChatGPT		
	Group1	Group2	Group1	Group2	Group3
SMOG	10.07	10.22	12.91	12.71	12.77
FKGL	8.94	8.54	12.45	19.81	10.76
FKRE	58.36	64.23	45.05	45.69	46.99

Assessing Reading Level

Reading level

Measures	HFChat		ChatGPT	
	HCP1	HCP2	HCP1	HCP2
Amount of Information (# topics/answer)	4.5	1.8	5.9	2.6
Accuracy	2.7	1.7	5.6	2.5
% Accuracy- Correct Information/ Total information per answer	62.6	96.6	95	96.6
Relevance- Total No of Relevant answers	2/10	7/10	7/10	3/10

Objective Evaluation

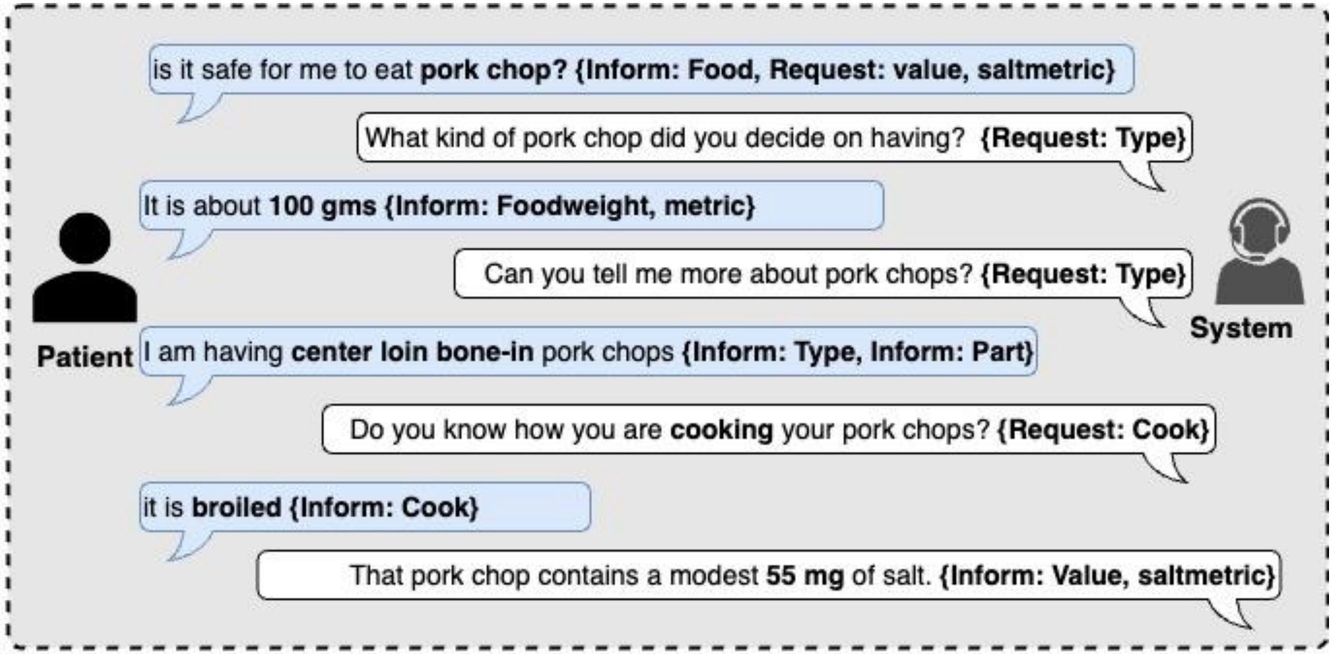
Table of Content

Salt Intake Conversational System

- HF patients must monitor and limit salt intake.
- Salt intake - key topic during HF education sessions
- AAs are more affected by HF
    - Have greater sensitivity to salt
    - Encounter challenges like food deserts
    - Rely more on processed salty foods
    - Only 58% can read the salt content on nutrition labels
    - have low numerical literacy

Salt Intake Conversational System

- **Goal**: Task-Oriented Dialogue System to help HF patients to monitor salt
- Requires numerical ability



Sample Template Conversation

Tayal, A., Di Eugenio, B., Salunke, D., Boyd, A. D., Dickens, C. A., Abril, E. P., Garcia-Bedoya, O., and Allen-Meares, P. G.: 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, eds. D. Demner-Fushman, S. Ananiadou, P. Thompson, and B. Ondov, pages 93–103, Torino, Italia, May 2024. ELRA and ICCL.

Dataset Creation

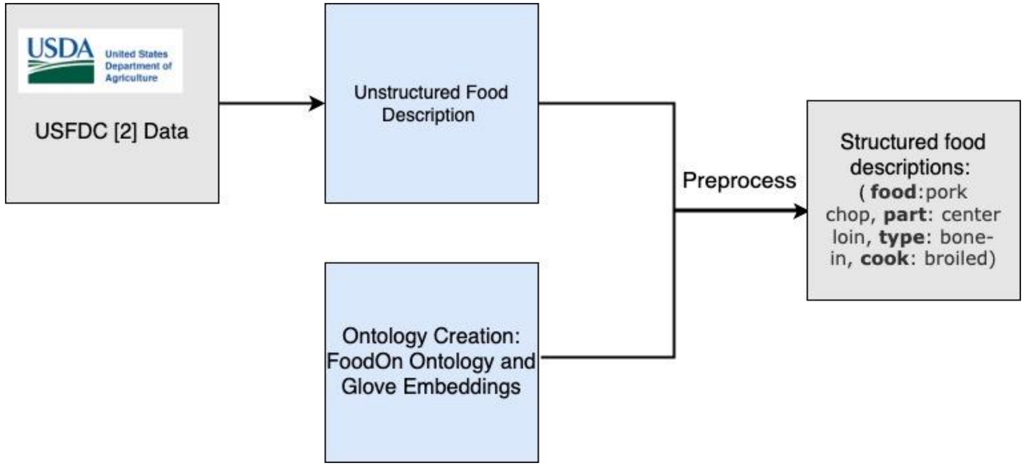
Food Descriptions	Salt Value
Pork, fresh, loin, top loin (chops), boneless, separa-ble lean and fat, raw	48
Pork, fresh, loin, center loin (chops), bone-in, sepa-rable lean and fat, cooked, broiled	55
Pork, fresh, blade, (chops), boneless, separable lean and fat, cooked, broiled	58
Pork, fresh, loin,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

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

- Provides salt content for standard food measurements
- Users: Often don't frame queries in these standard terms
- Unstructured food descriptions
- Lacks clarity on the significance of each component

[1] USFDC. 2022. Us food data central.

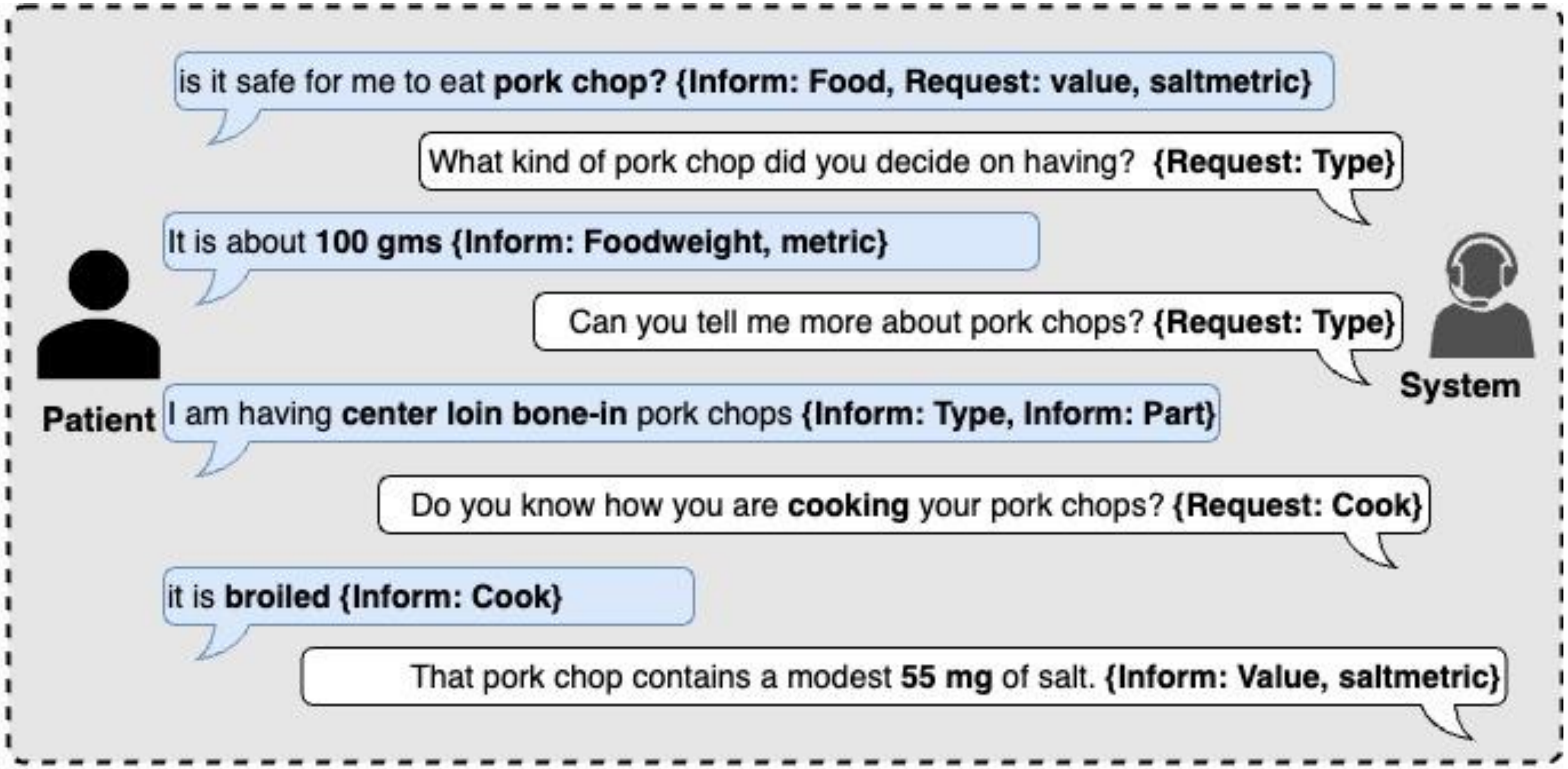
Dataset Creation



Dataset Creation Pipeline

- Ontology Creation
  - FoodOn
  - GloVe
- Lack of conversational dataset
- No annotation

Conversational Dataset Creation



Sample Template Conversation

Conversational Dataset Creation

- Template based conversational dataset (MultiWOZ [1])
- Conversation alternates between user and system
- **Slots**: food, cook, type, animal, part, foodweight, metric
- System's questions and user's responses are randomly selected from template
- Types of turns
    - Matching answers
    - Misaligned answers
    - Changing answers

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

Distributional Characteristics

[1] P Budzianowski, T-H Wen, BoH Tseng, I Casanueva, S Ultes, O Ramadan, M Gašić. 2018. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. 2018 EMNLP

Research Questions

- Is a Language Model sufficient to build a conversational system that requires numerical ability, or is a hybrid system required by integrating neuro-symbolic rules?
- Can we use neuro-symbolic rules externally to control the output of the system?
- **Methodology**: How can we effectively combine the strengths of task-oriented dialog systems (TODS) and LMs/LLMs to create a hybrid dialog model, and integrate neurosymbolic rules with the LM/LLM-based dialog system to control the dialog system.

Training the Dialog System

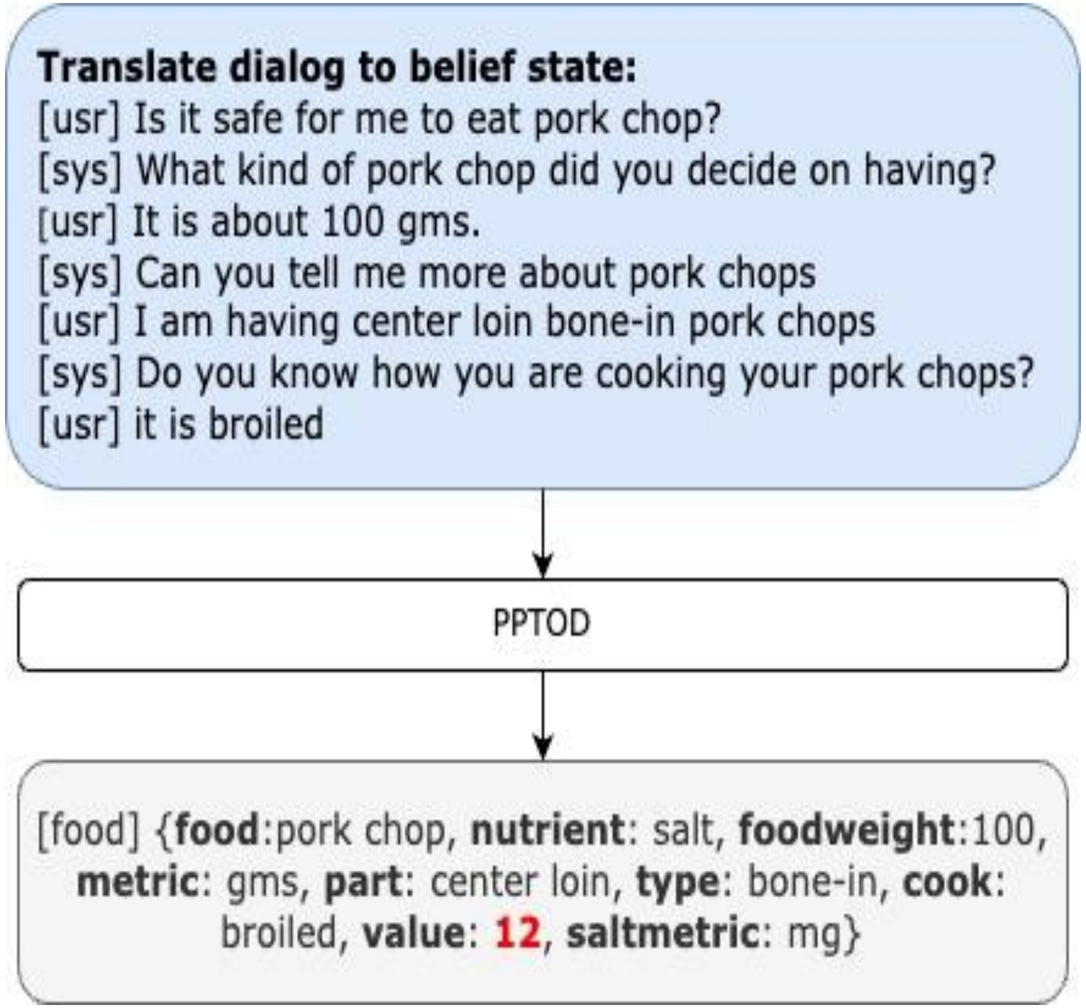
PPTOD (Plug-and-Play Task Oriented Dialog System) [1]

- T5 based - model designed for task-oriented dialogue [6]
- Adept at in-context learning employing customized prompts
- Modular

How do we use PPTOD?

- Trained with maximum likelihood objective
- Few-Shot Training

Performance: correctly identified most slot values but not salt values



Trained PPTOD Model

[1] Y Su, L Shu, E Mansimov, A Gupta, D Cai, Y Lai, Y Zhang. 2022. Multi- task pre-training for plug-and-play task-oriented dialogue system. In Proceedings of the 60th Annual Meeting of the ACL

Results

	Train Size	Epochs	Inform	Success	BLEU
PPTOD	100	8	71.43	0	24.99
PPTOD	300	7	75.00	5.00	34.30
PPTOD	500	9	82.86	2.86	29.81
PPTOD	1000	7	93.50	2.70	29.00

PPTOD Evaluation Metrics

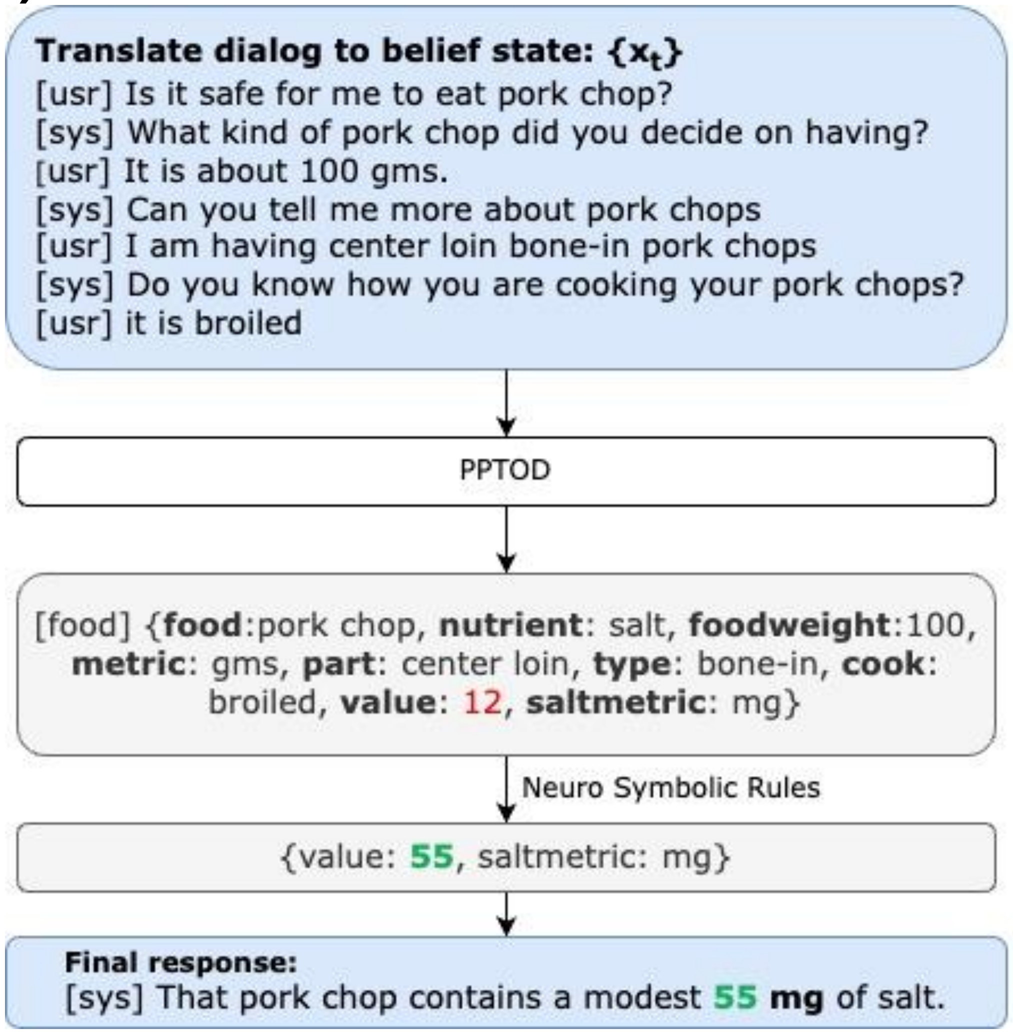
Proposed Model (NS-PPTOD)

NS-PPTOD

- PPTOD+ Neuro-symbolic rules
- Applied external to the system

NeuroSymbolic Rules:

- Retrieval of accurate salt value from the database
- Mathematical calculation of the correct salt value for queried food weights
- Able to respond to queries for non-standard food quantities (bowl, plate)



NS-PPTOD Model

Results

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

Increase in performance when using NS-PPTOD compared to PPTOD

Results

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

Increase in Joint-Accuracy when using NS-PPTOD compared to PPTOD

Addressing the Research Questions

- **Is a Language Model sufficient to build a conversational system that requires numerical ability, or is a hybrid system required by integrating neuro-symbolic rules?**
- Fine-tuning is insufficient for training a dialog system that requires numerical reasoning
- **Can we use neuro-symbolic rules externally to control the output of the system?**
- Incorporating neuro-symbolic rules help
- Able to control the salt value.
- Experimental results show that integrating neuro-symbolic rules led to a 20% improvement compared to a fine-tuned model.

- **Methodology**: How can we effectively combine the strengths of task-oriented dialog systems (TODS) and LMs/LLMs to create a hybrid dialog model, and integrate neurosymbolic rules with the LM/LLM-based dialog system to control the dialog system.

Research Questions

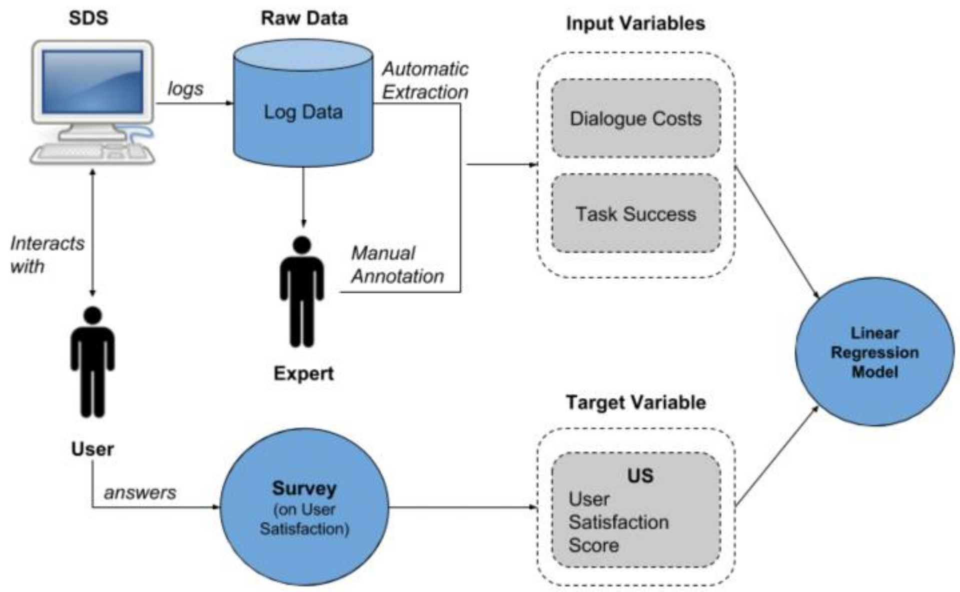
Evaluation: How do heart failure AA patients perceive and interact with a neurosymbolic TODS compared to an LLM-based system?

- With LLMs prevailing, there is a need for a controlled probing evaluations with real stakeholders which can highlight the advantages and disadvantages of a LLM based system with more traditional architecture with neuro-symbolic rules.
- How do LLM-based systems compare to traditional neuro-symbolic rule-based architectures in terms of performance, usability, and stakeholder satisfaction in real-world scenarios?

Was the evaluation comprehensive

- LLM-based dialog systems- difficult to evaluate
  - Lack of transparency about the data source
  - Do not operate within strict boundaries

- Limitations are critical when in patient-centric environment
- Human Evaluations remain gold standard [1]
- Need for controlled, probing evaluations with real stakeholders which can highlight advantages and disadvantages of more traditional architectures and those based on generative AI.



Paradise Framework

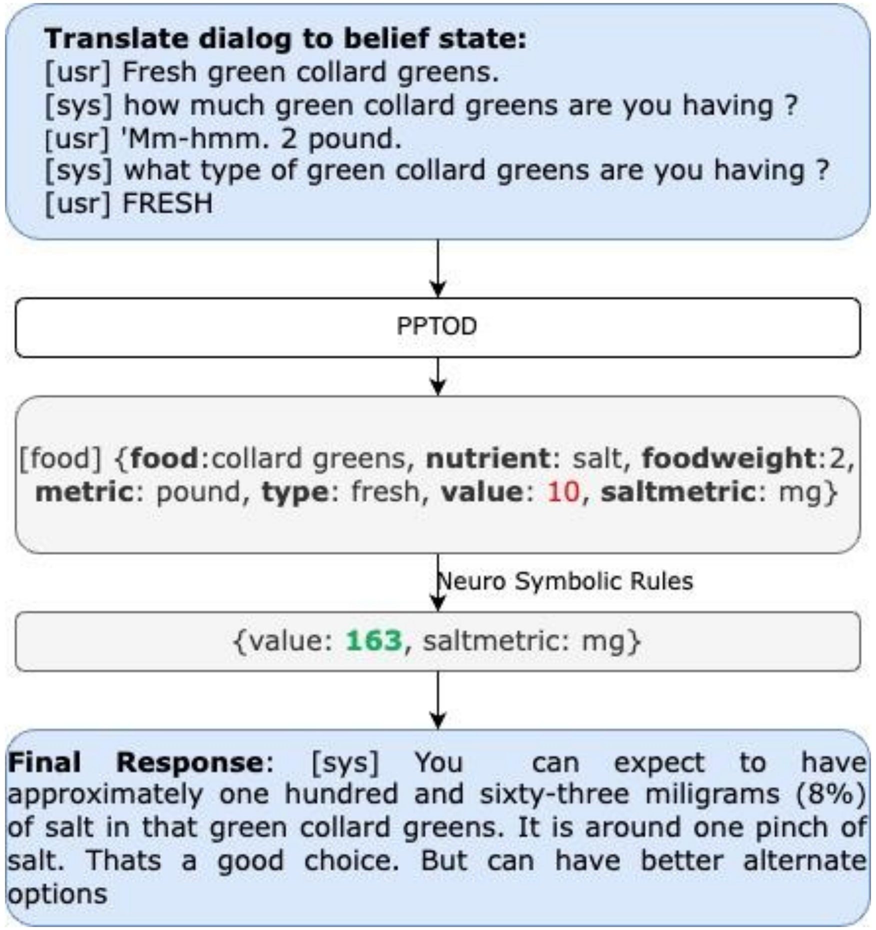
[1] Marilyn A Walker, Diane J Litman, Candace A Kamm, and Alicia Abella. 1998. Evaluating spoken dialogue agents with paradise: Two case studies. Computer Speech & Language, 12(4):317–347.

User Study

- Create 2 versions of dialog systems
- Compares two dialog systems— the NS-PPTOD system with a ChatGPT-based system using a within-subject design.
- **Aim** Evaluate and contrast the effectiveness of both systems offering insights into their respective impact in real-world scenarios.
- Conduct Intrinsic and extrinsic analysis using pre and post survey questionnaire to evaluate the 2 systems with AA patients

HFFood-NS

- Only the DST module of NS-PPTOD was used for reliability
- If a slot remains unfilled, system attempts to query the user about it up to 2 times.
- Template based system responses.
- In addition of giving salt value in mgs, gave the value in percentage of daily value and pinches also (informal)



HFFood-NS Model Interaction with the patient

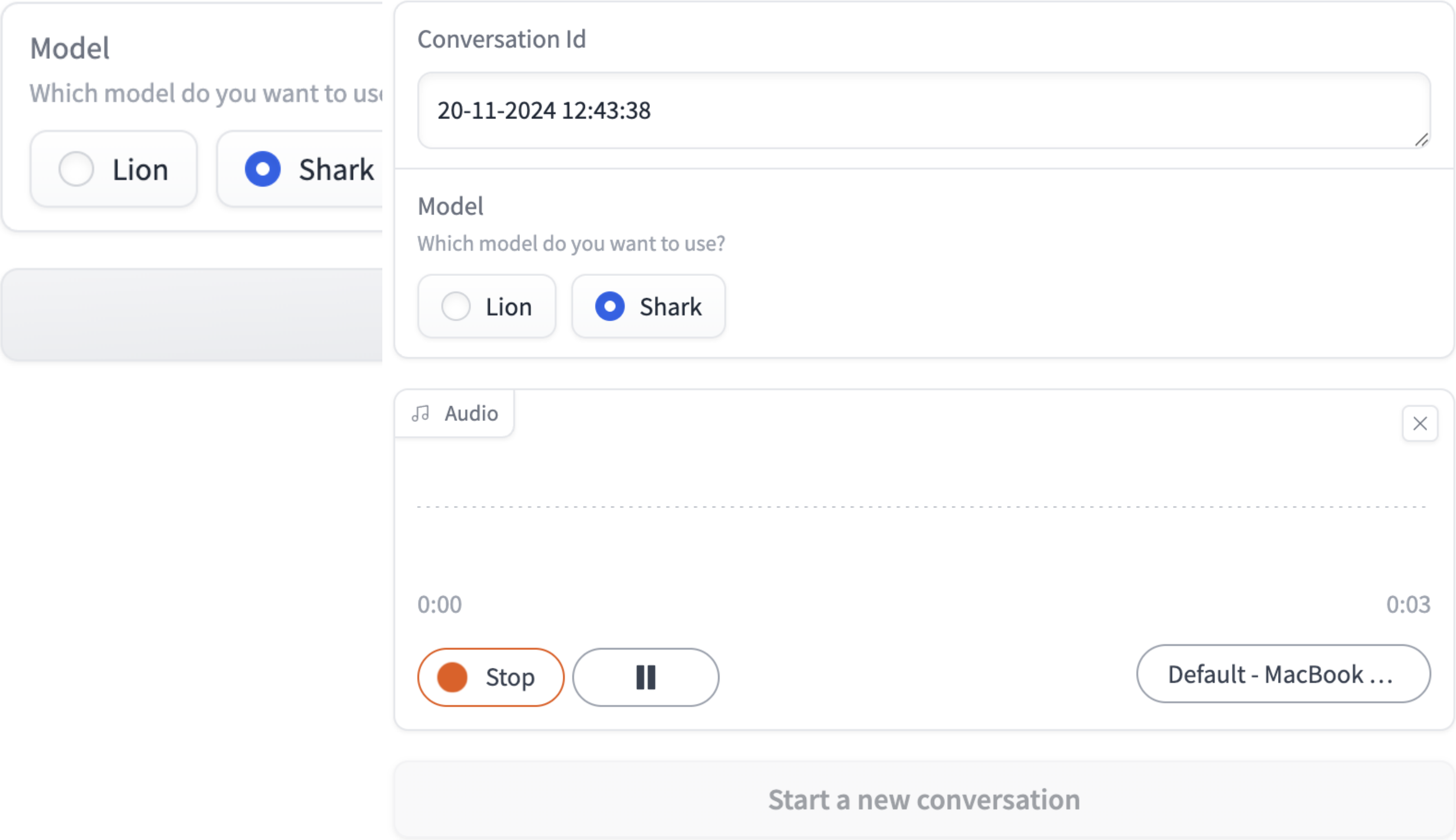
HFFood-GPT

- analyze the salt content in various foods.
- methodically asks users about the **food type, cooking method, and portion size, one question at a time.**
- 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 a**nd suggesting from consulting a professional for dietary guidance.
- 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.**
- **Do not look for information on the web.**

User Study

- Recruitment Criteria
  - 20 AA patients
  - Age – 18-89 yrs of age
  - Heart failure
  - English Speaking
- Procedure
  - Oral Consent
  - Pre-Survey Question (Health and Digital Literacy)
  - Test 2 systems
  - Post-Survey Questionnaire (Feedback on 2 systems, preference, helpfulness, likelihood of use)

UI

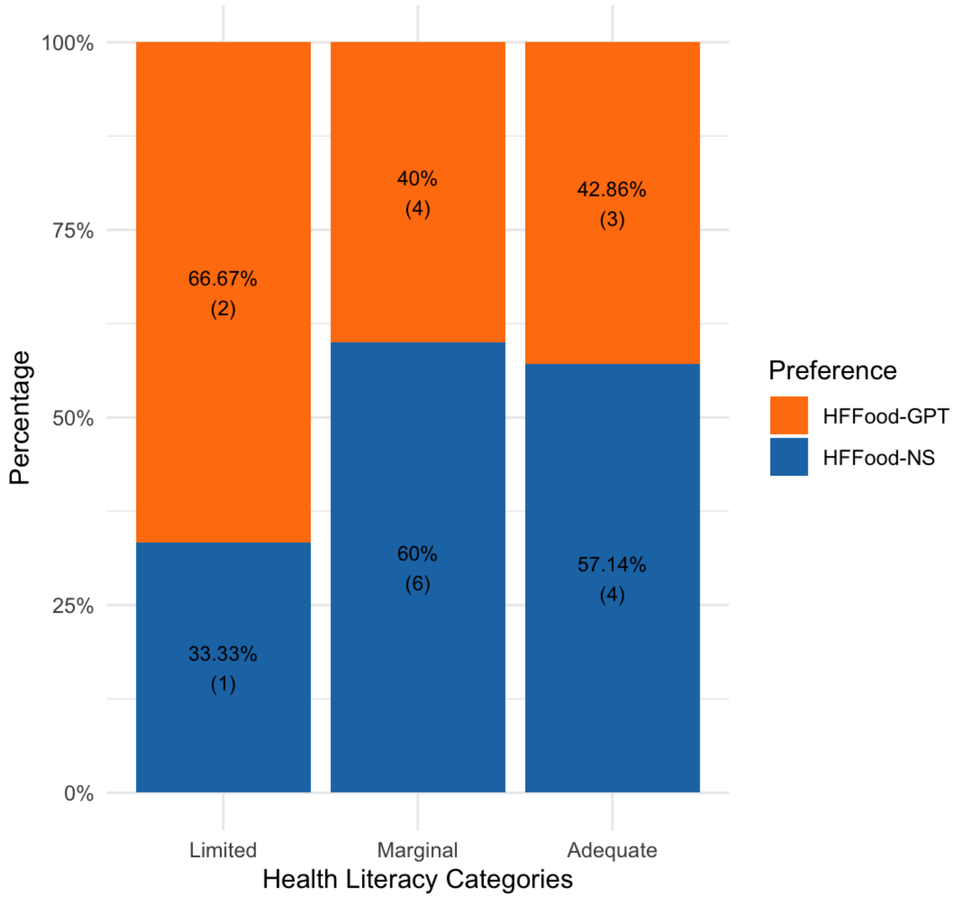


Model
Which model do you want to us
Lion
Shark
Conversation Id
20-11-2024 12:43:38
Model
Which model do you want to use?
Lion
Shark
Audio
0:00
0:03
Stop
Default - MacBook ...
Start a new conversation

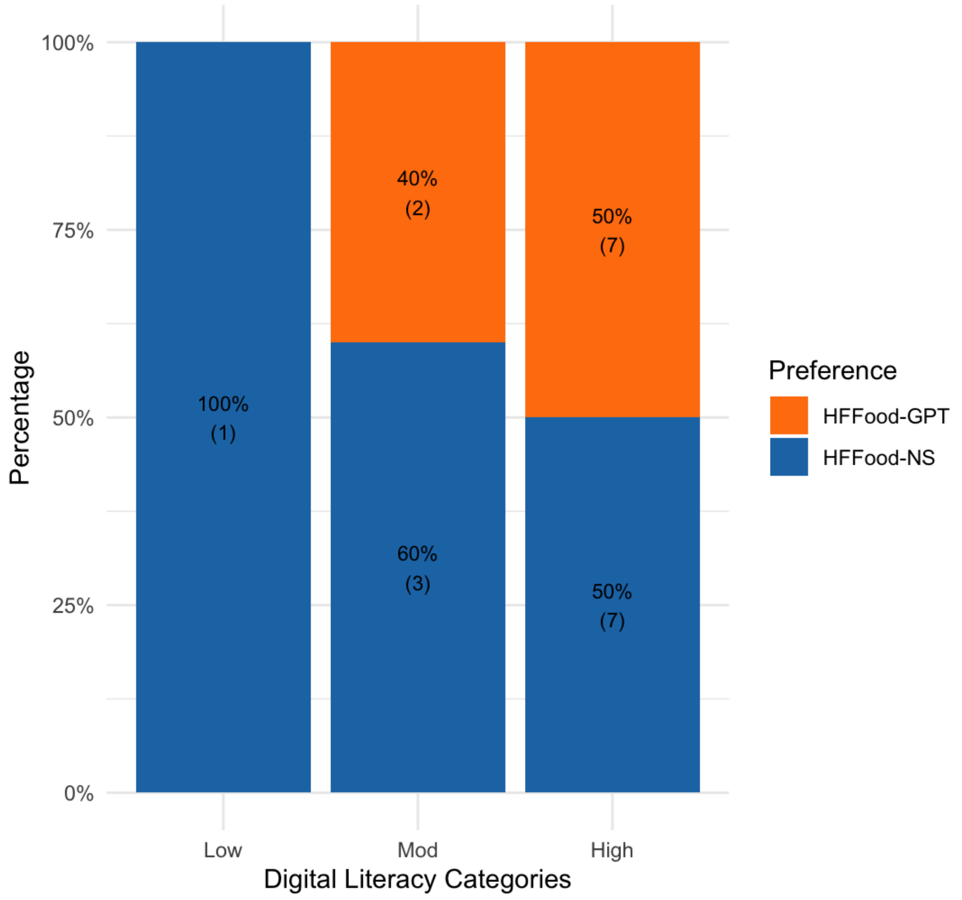
Statistics

- Approached 70-80 patients
- Recruited 23 patients
- 20 patients completed the study (mean 58.75, sd=14.32)
- Took 6 months
- Challenges

Pre-Survey



100%
75%
50%
25%
0%
Percentage
66.67%
(2)
33.33%
(1)
40%
(4)
60%
(6)
42.86%
(3)
57.14%
(4)
Limited
Marginal
Adequate
Health Literacy Categories
Preference
HFFood-GPT
HFFood-NS



100%
75%
50%
25%
0%
Percentage
100%
(1)
40%
(2)
60%
(3)
50%
(7)
50%
(7)
Low
Mod
High
Digital Literacy Categories
Preference
HFFood-GPT
HFFood-NS

Intrinsic Evaluation

- Used USFDC dataset as a baseline to evaluate the accuracy of the 2 systems
- Treated HFFood-GPT as a TODS by categorizing the provided values into slots manually
  - Challenging to evaluate due to its black box nature
  - Evaluated similar to HFFood-NS
- GPT provided values in absolute value, range, categorical values

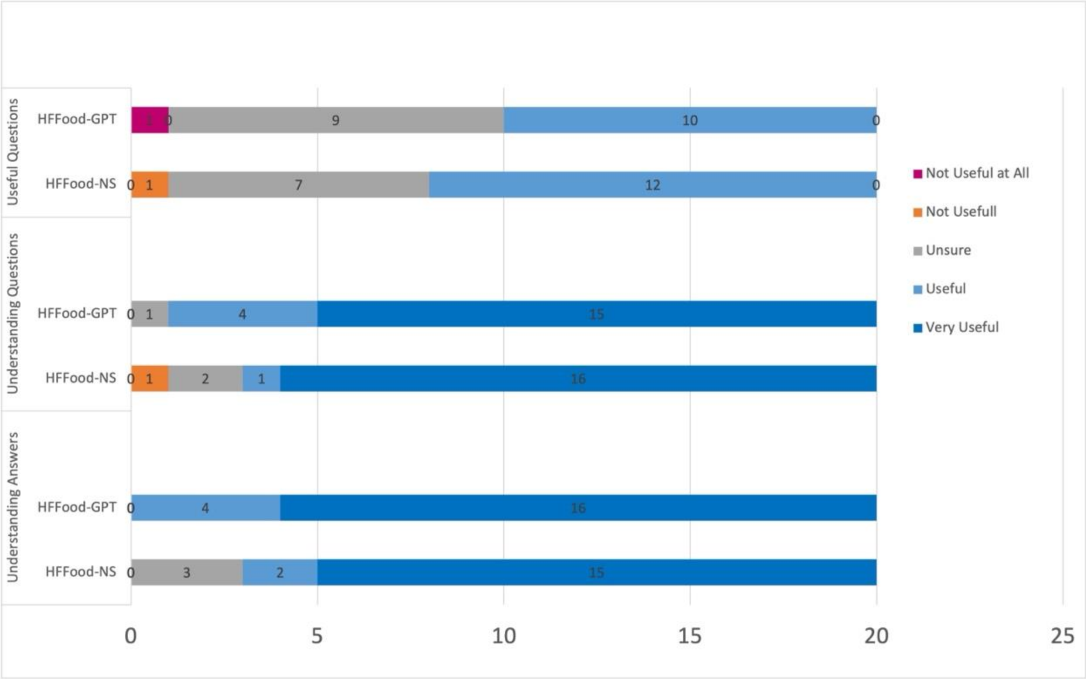
Intrinsic Evaluation

	HFFood-NS	HFFood-GPT
Avg No of Turns	3.6	3
Avg Processing Time	6.7	11.4
Avg System Words	14.5	54.5
Avg Retries	2	1.7
Avg WER	.483	.41
Task Completion	84%	62%
Task Accuracy System	37%	24%
Task Accuracy Net	11%	20%
Slot Accuracy	56%	89%

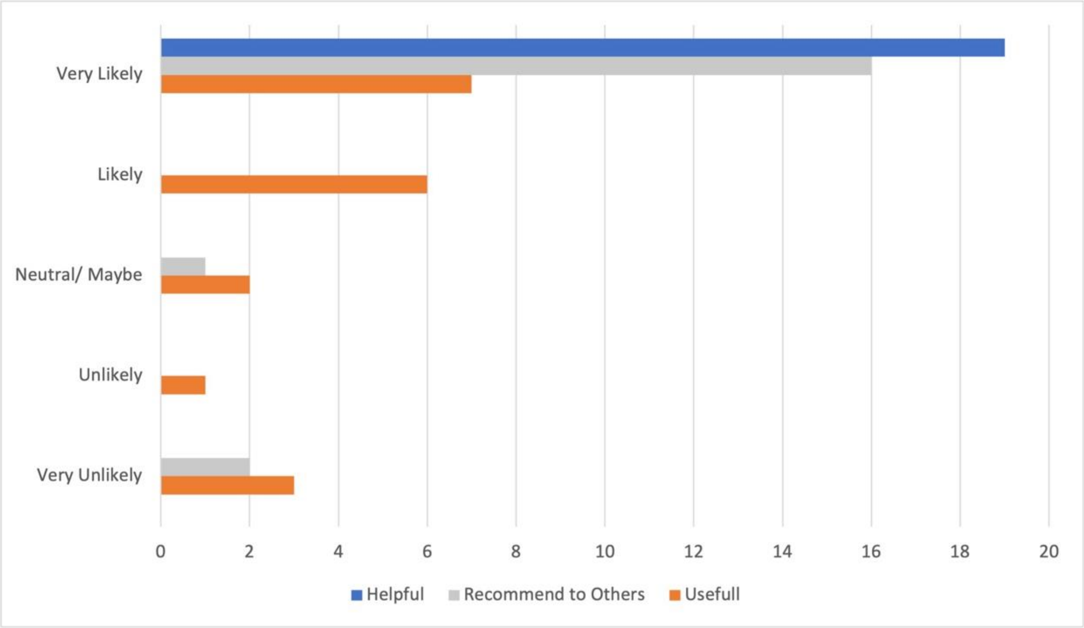
Preference

- **Preference-** 11 HFFood-NS, 9 - HFFood-GPT.
- No recall bias ($r = 0.3$, $p = 0.19$).
- **Reasons for preference**
  - HFFood-NS - precise and to-the-point answers, faster flow, concise responses.
  - HFFood-GPT - ease of understanding questions, clear, detailed explanations, better-formulated questions.
- 11 participants (55%) preferred informal terms like pinches or dashes.

Extrinsic Evaluation



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



Post Survey Analysis: How users perceived the conversational system

Pay Attention to salt

- Except 5 people everyone said they paid attention to salt.
- Most people preferred Informal methods to measure salt.

Measurement Method	Count
Informal (Pinches/ dash)	11
Daily Value	3
mg	4

HFFood-NS

- Ability to conduct thorough and precise error analysis
  - Ability to add fail safe
- Greater Control over the system
  - Remain aligned with its primary goal
  - GPT asked unnecessary questions sometimes- color of the bell pepper
  - Ensured consistency in responses (10%-gms, 10%-categorical, 8%-ranges, 2%-tsps.)

Error Analysis

Error		HFFood-NS samples
Missed Slot		23
No data Fetched	Complex Query	6
	System Error	6
	Food Not in USFDC	3
	Internet Connectivity	3
Metric Mismatch		5
Wrong food identified		8
Slot not in USFDC		4

HFFood-NS

Speaker	Utterance
System	Double-dipped Italian beef with cheese and hot giardiniera peppers.
Patient	what is the serving size of the Italian beef ?
System	about the size of a cell phone.
Patient	What is your Italian beef portion like ?
PE	The size of a cellphone, fat cellphone.
	you will find about nine hundred and fourteen miligrams (46%) of salt in Italian beef. It is around six pinch of salt. You are about to reach half of your daily intake.

HFFood-NS Example

HFFood-GPT

- Handled diverse food descriptions easily
  - Serving size,
  - HFFood-NS – most errors in determining food-weights and metric slots.
- Attempted to gather recipe and estimate based on ingredients.
  - But can be overly specific and ask too many clarification questions
- People identify foods with brand names- hurricane whopper
- Gave user options – 30% of turns
- Did not adhere to prompt
  - Refer db to users- 38.5% of turns
  - 1 question at a time- 7% of turns

HFFood-GPT

Speaker	Utterance
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?

HFFood-GPT Example

HFFood-GPT

Speaker	Utterance
Patient	The salt content of lasagna.
System	It appears there was no direct match for lasagna in the provided data. To give you the best possible estimate, could you tell me if the lasagna you're inquiring about is homemade, from a restaurant, or a prepackaged product?
Patient	From a restaurant. Used to be homemade, now from a restaurant.
PE	I'm working with a certain set of data to provide salt content estimates. Since your lasagna is from a restaurant, I'll calculate the salt content based on typical values for restaurant meals.\n\nBefore proceeding, could you specify the portion size?
	That's more than six forkfuls, let's say 12, I need 12, yeah.

HFFood-GPT Example

Comparison

	HFFood-NS	HFFood-GPT
Task Completion	✅	❎
Accuracy	✅	❎
Slot Accuracy	❎	✅
Fewer Speech Errors	❎	✅
Less Processing Time	✅	❎

Comparison

	HFFood-NS	HFFood-GPT
Error Analysis	✅	❎
Controlled	✅	❎
Reliable	✅	❎
Predictable	✅	❎
Complex Query	❎	✅
Gave Optios	❎	✅
Fluent	❎	✅
Concise	✅	❎
Faster to Deploy	❎	✅

Addressing Research Questions

Evaluation: How do heart failure AA patients perceive and interact with a neurosymbolic TODS compared to an LLM-based system?

- **HFFood-NS** – more accurate, completes more tasks, more controllable
- **HFFood-GPT**- handles complex queries more effectively, diverse responses
- 2 systems complement each other

Research Questions

- **Data**: What are the different ways to prompt LLM to generate synthetic conversations in the absence of patient-oriented self-care dialogues, and is prompting enough to control/ personalise the conversations?
  - How can prompting be leveraged to generate synthetic, patient-oriented dialogues?
  - Is Prompting sufficient to generate appropriate self-care conversations?
  - Do the generated conversations convey empathy towards the patients?
  - Is generating reasoning before conversations more effective than directly generating the conversations?

How can prompting be leveraged to generate synthetic, patient-oriented dialogues?

- ChatGPT was used to generate simulated conversations using 5 distinct prompting strategies
  - According to the domain
  - According to Race
  - AAVE
  - SDOH
  - Reasoning-based Prompts

Generating Conversations According to Domain

You are a healthcare educator focusing on heart failure. Your purpose is to answer and ask follow-up questions related to heart failure.
Simulate {**lines**} round conversation between African American heart failure patient and healthcare educator where the patient asks for recommendations for {**domain**}.
You should **empathetically** communicate medical information in a simple manner. You should focus on guiding the patient towards the importance of {**domain**} in heart failure.
Scripts should be generated in following format: [speaker] [utterance]
(Do not use phrases such as "consult with your healthcare provider", assume that you are the healthcare provider)

Prompt for generating conversations according to domain

Generating Conversations According to Domain

Speaker	Utterance
Patient	Okay, but what kind of foods should I be eating or avoiding?
PE	Great question! In general, you should aim to eat more fruits, vegetables, whole grains, and lean protein. These provide heart-beneficial nutrients. **Limit salt, sugar, unhealthy fats**, and alcohol intake as they can worsen heart conditions.
Patient	Is it bad for me to eat fried chicken or pork ribs with my condition? I have been eating them all my life.
PE	I understand that these foods are part of your cultural and personal food habits. **However, these are typically high in saturated fats which can increase cholesterol levels.** Try to limit it, or prepare them in healthier ways. Maybe bake the chicken instead of frying, and use herbs and spices for flavor instead of excessive salt or fat.
Patient	How about beverages? Can I still drink things like sweet-natured tea or alcohol?
PE	**Sweet tea and alcohol can add extra calories and sugar to your diet, which can lead to weight gain and can strain your heart. Try drinking more water or unsweetened beverages and limit alcohol intake.**

Excerpt of conversation where the patient is advised to limit calories and sugar, along with salt intake, and advised to drin k more water instead of less.

Generating Conversations According to Race

	Caucasian	White	AA
Fluid			
Food			
Exercise			
Self-care	.88	.87	.86

Simulate a 10-round conversation between a Caucasian patient with heart failure and a clinician where the patient asks for fluid intake recommendation.

Cosine Similarity Values Compared to No Race

AAVE

Speaker	Utterance
Patient	Thank you for letting me know. I often enjoy foods like watermelon and okra, which I've heard have high water content. Do I need to count those in the 1.5 to 2 liters?
PE	Yes, you're correct. Foods like watermelon and okra do contribute to your fluid intake. While you do need to consider them, the 1.5 to 2-liter guideline typically includes both liquids and foods with high water content
Patient	I enjoy herbal teas, but I've also heard about the effects of caffeine. Should I stick to caffeine-free options?
PE	Herbal teas are a good choice, especially if you want to avoid caffeine. Some herbal teas can have health benefits too. Just be sure to read the labels and choose options that are free of caffeine and low in added sugars.

Generating Conversations According to SDOH Features

You are a healthcare educator focusing on heart failure

Speaker	Utterance
Patient	I want to exercise, but I'm not sure what's safe for my heart. Any suggestions?
PE	Absolutely! Walking is a great option. It's low-impact and helps strengthen your heart.
Patient	But, my neighborhood is not safe for walking, what can I do?
PE	I understand. **What about a stationary bike or a treadmill at home?**
Patient	I can't afford to buy exercise equipment. Is there anything else I can do?
PE	Sure! You can do chair exercises or household chores which can also help to keep you active.

Reasoning-based Prompt

Premise: Socio-economic factors, cultural influences, and healthcare access impact heart failure management in African-American patients, leading to disparities in disease outcomes.
Reasoning: Considering the patient's age, gender, living in an unsafe neighborhood, and being below the poverty line, it is crucial to address the barriers she may face in accessing safe and affordable exercise options. Lack of resources, fear of safety, and limited access to healthcare facilities may hinder her ability to engage in physical activity for heart failure management.
Solution: Encourage the patient to start with **simple, low-cost exercises at home,** such as **walking in a safe area, using household items as weights, or following online workout videos**. Emphasize the importance of consistency and starting slow to avoid injury.
Anticipate Barrier: The patient may struggle to find a safe and affordable environment to exercise, leading to inconsistency in her physical activity routine.
Solve Barrier: Provide resources for community centers, local parks, or discounted gym memberships that offer safe and affordable exercise options. Encourage the patient to enlist the support of family members or friends to exercise together for added safety.
Educate: Educate the patient on the importance of regular physical activity in managing heart failure, the benefits of exercise in improving heart function and overall health, and the impact of socio-economic factors on health disparities. Emphasize the need to prioritize her health and well-being despite the challenges she may face.

Reasoning-based Prompts

Speaker
Patient
PE
Patient
PE
Patient
PE

You are a healthcare educator focusing on heart failure.

You are a healthcare educator focusing on heart failure. Your purpose is to answer heart failure patient questions. You should empathetically communicate medical information in a simple manner.
Patient Description:
gender: {gender}
socio-economic condition: {socio_economic}
neighborhood: {neighborhood}
age: {age}
{Reasoning}
Given the patient description and reasoning, simulate lines round conversation between African American heart failure patient and healthcare educator where the patient asks for recommendations for domain
Scripts should be generated in the following format: [speaker] [utterance] between the patient and the healthcare educator. Each educator's turn should not be longer than 20 words and should use simple English.

Research Questions

- **Data**: What are the different ways to prompt LLM to generate synthetic conversations in the absence of patient-oriented self-care dialogues, and is prompting enough to control/ personalise the conversations?
  - How can prompting be leveraged to generate synthetic, patient-oriented dialogues?
  - Is Prompting sufficient to generate appropriate self-care conversations?
  - Do the generated conversations convey empathy towards the patients?
  - Is generating reasoning before conversations more effective than directly generating the conversations?

Evaluation

- Qualitative Evaluation
  - Distributed a questionnaire to 10 NLP PhD students sp
- Quantitative Evaluation

	Round Adherence Rate	Format Adherence Rate
Domain	0	1.0
AAVE	0.04	0
SDOH	.64	0
Reasoning	0.35	0.23

Is Prompting sufficient to generate appropriate self-care conversations?

- Do the generated conversations adhere to the convention of human conversation?
- Were the generated conversations appropriate?

Do the generated conversations adhere to the convention of human conversation?

	Follow-up ratio	Ratio of Words
Domain	0.01	3.4
AAVE	0.85	2.68
SDOH	.003	1.83
Reasoning	0.02	1.8

Were the generated conversations appropriate?

- At least 20% of the conversations: Evaluators identified instances where the HE should have provided a different answer.
- *This includes water, juice, and other beverages – only positive examples*
- *When asked about whether they should be concerned about drinking too much – HE provided information about the dangers of drinking water too quickly*

Do the generated conversations convey empathy towards the patients?

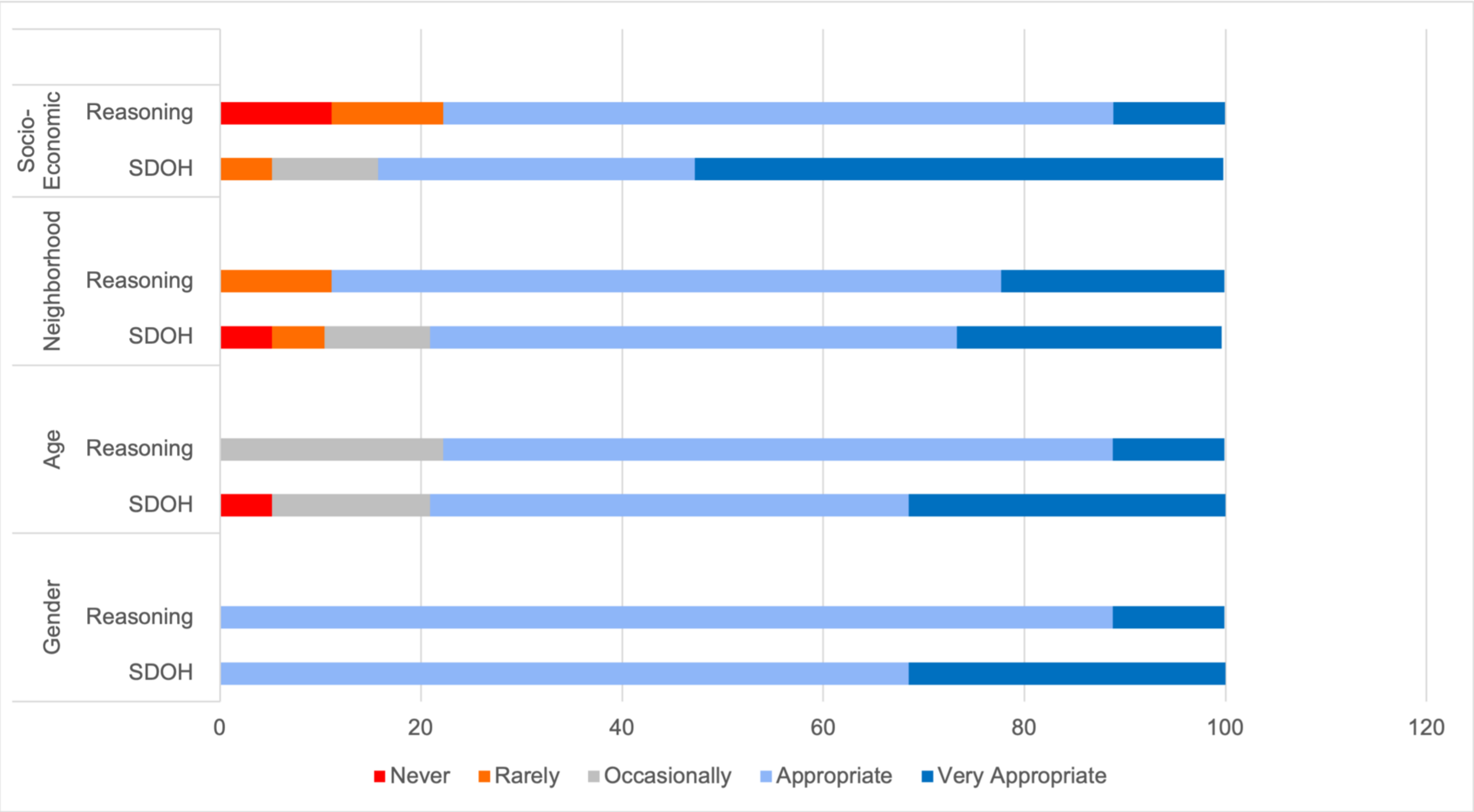
- Level of perceived empathy was relatively unchanged:
  - 25% of the communication receiving a Likert scale rating of 3 or lower.

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

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

- Tone-deaf, lacked empathy
- Robotic

Personalise Conversations based on SDOH Features



Addressing the Research Questions

- **Data**: What are the different ways to prompt LLM to generate synthetic conversations in the absence of patient-oriented self-care dialogues, and is prompting enough to control/ personalise the conversations?
  - Prompting alone is insufficient to control or personalize conversations.
  - Struggled to follow even basic instructions
    - such as adhering to a set number of dialogue rounds, limiting word count, or asking appropriate follow-up questions.
  - While it can incorporate SDOH features and improve dialogue quality through reasoning prior to generation, it remains unsuitable for direct deployment in patient-centric settings due to the lack of controllability.

Recap

- Neuro-symbolic rules were applied externally to control salt values,
  - Improved accuracy but kept reasoning separate from the model.
- An alternative is fine-tuning the model with symbolic rules
- In TODS, dialog acts represent user intent; training with them embeds symbolic reasoning more naturally.
- HFChat-NS used template-based responses—more control, less flexibility.
- ChatGPT outputs were more natural and diverse.
- Just prompting is insufficient

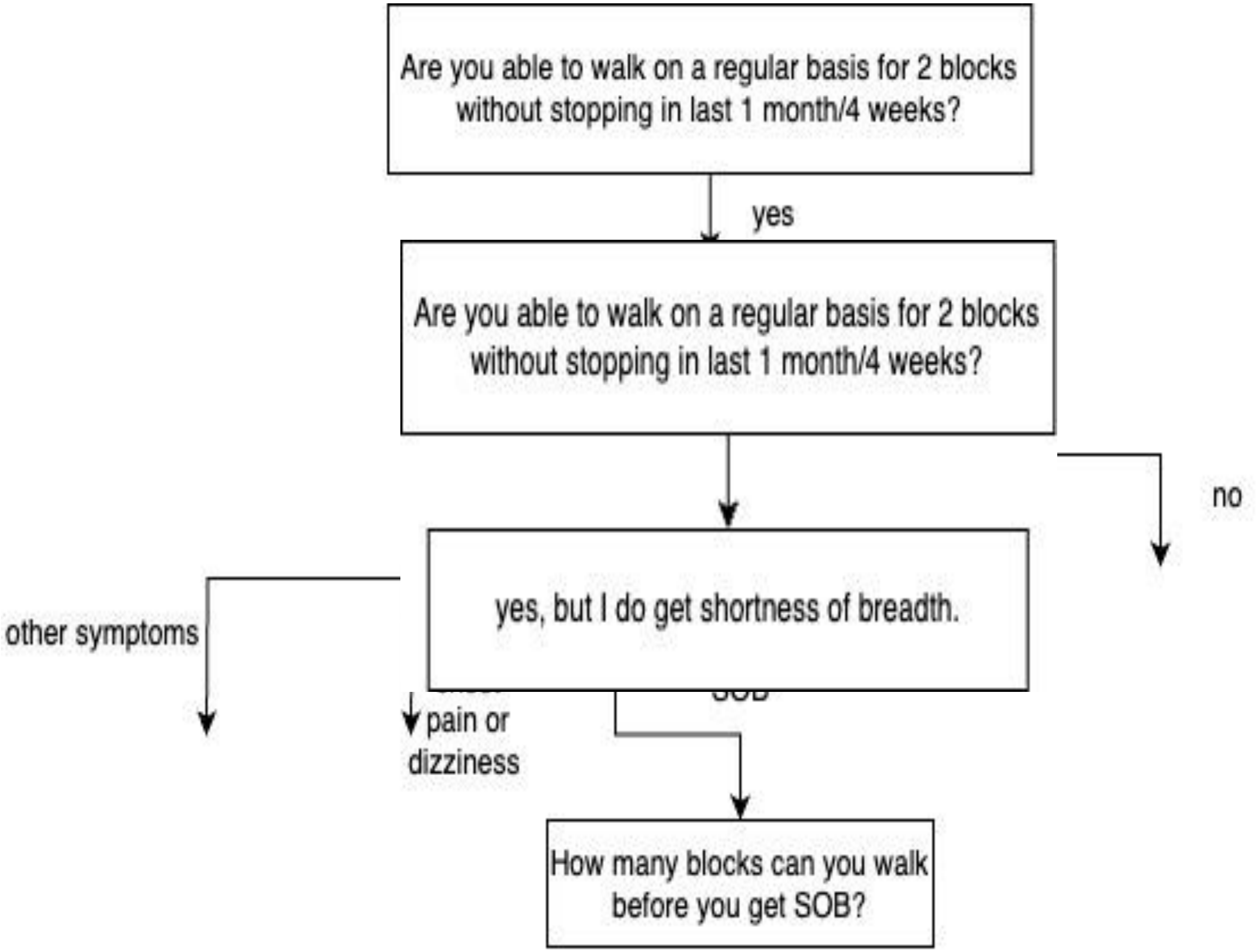
Exercise based dialog system

- Recommend at least 150 minutes of moderate intensity exercise per week
- Challenges
  - no pre-existing ontology
  - Routine activity that demands constant motivation
  - Needs to be personalized
  - Individual differences in fitness levels further heighten the complexity
  - Should be actionable- not only be persuasive

Data

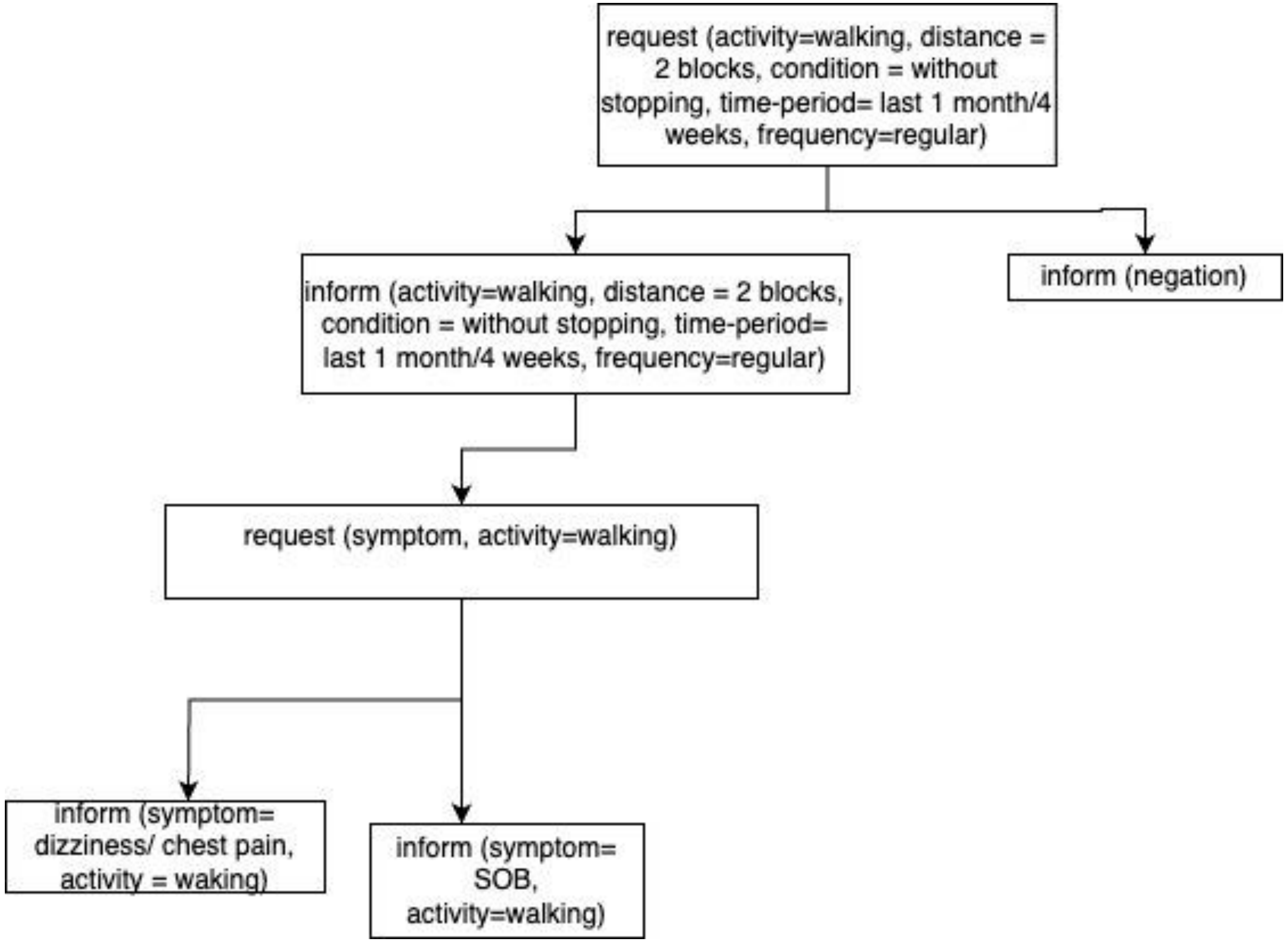
- Consulted HealthCare Professionals to gain insights into how the conversation should take place.
- Created dialog paths to model patient-educator interactions and generate synthetic data.

Data

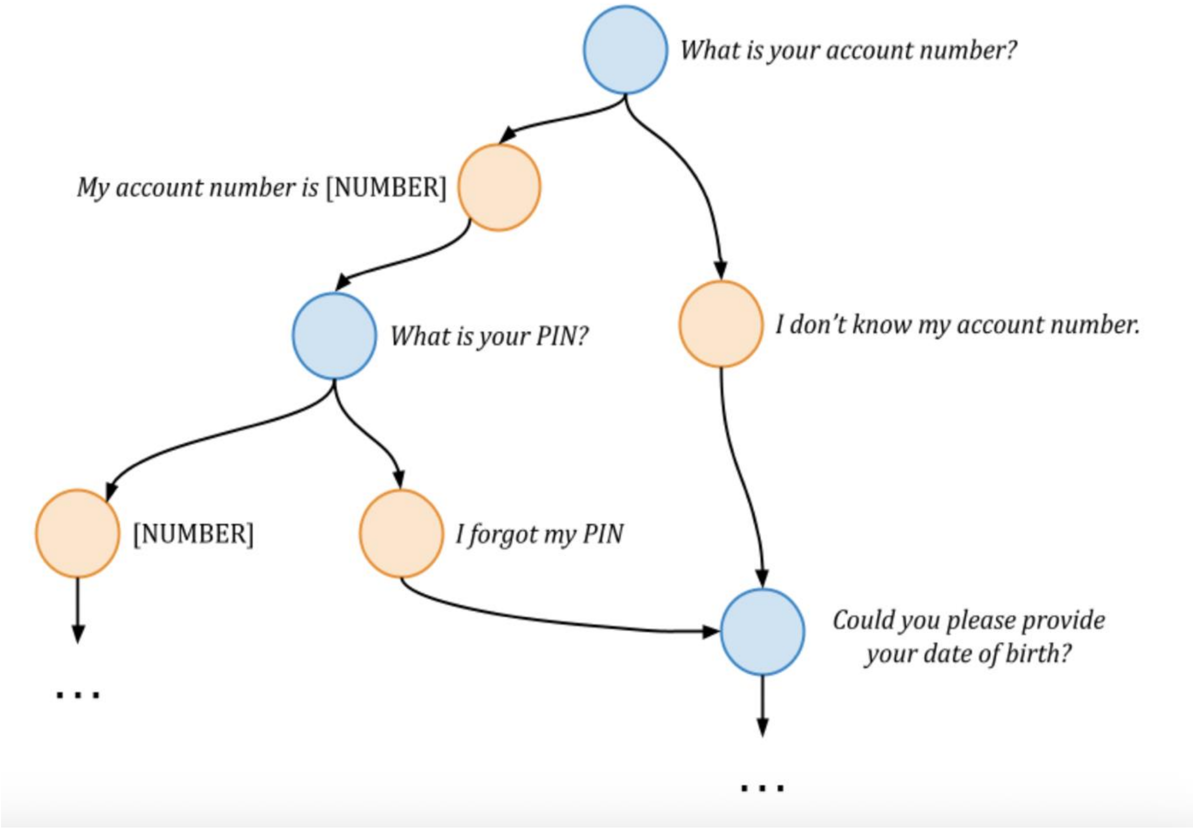


A Sample Conversation Path

Schema Graphs



Schema Graphs



Mehri, S. and Eskenazi, M.: Schema-guided paradigm for zero-shot dialog. In Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue, eds. H. Li, G.-A. Levow, Z. Yu, C. Gupta, B. Sisman, S. Cai, D. Vandyke, N. Dethlefs, Y. Wu, and J. J. Li, pages 499–508, Singapore and Online, July 2021. Association for Computational Linguistics.

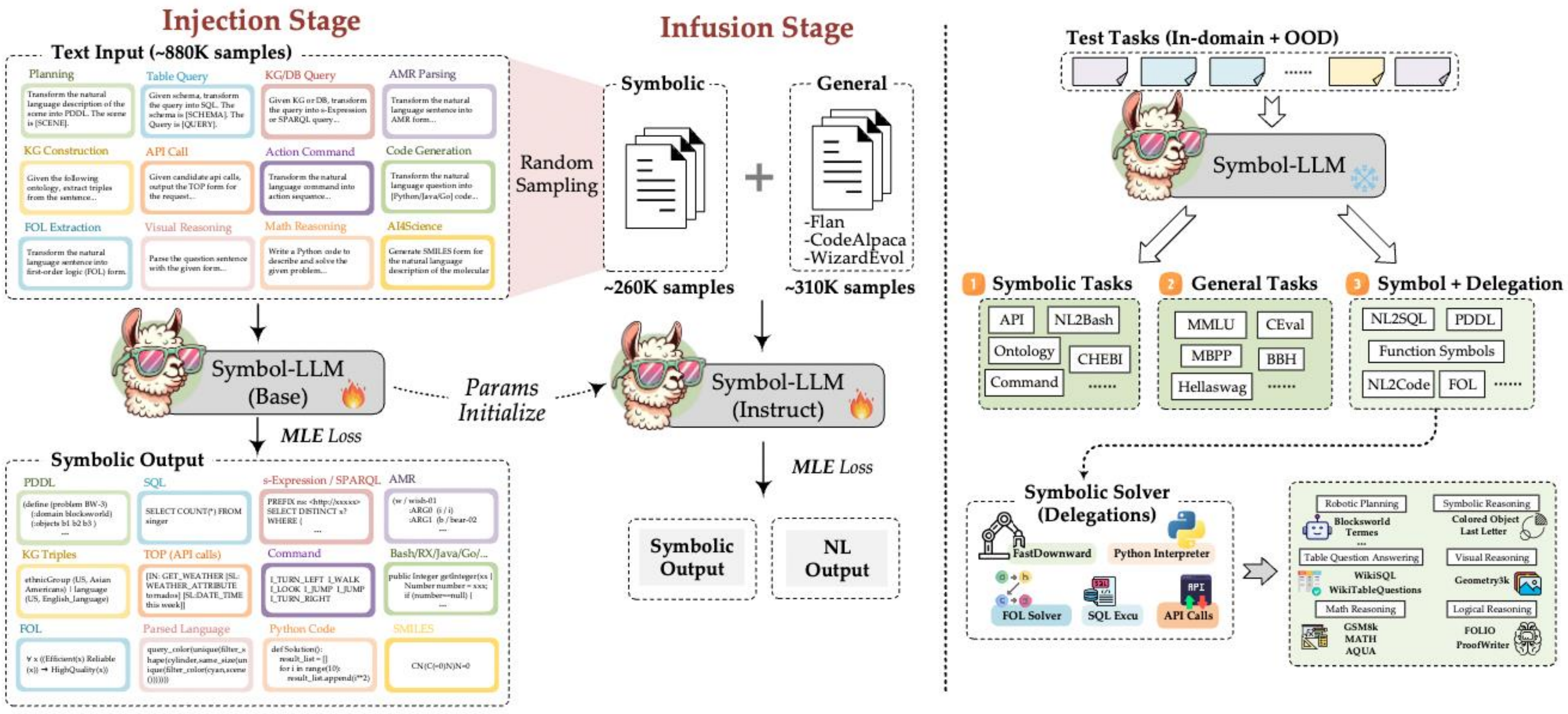
Proposed Work

- *How can we effectively combine the strengths of TODS and LMs/LLMs to create a hybrid dialog model? Specifically:*
    - *How can we integrate dialog acts to control the flow of the conversation?*
    - *How can different persuasion strategies be integrated to enhance the generation of patient education responses?*
- *How do users/patients/older adults perceive such a system?*

Hybrid Dialog System

- Not Generate Responses Directly
- Dialog Management
  - Schema-Graphs will provide more control to the flow of the conversation
  - Predict next dialog act
  - Train T5, Symbol-LLM model on schema
- Response Generation
  - Examine patient-educator conversations for persuasive communication strategies
  - Use strategy to generate response

Symbol-LLM



(a) Tuning Phase (b) Test Phase

Xu, F., Wu, Z., Sun, Q., Ren, S., Yuan, F., Yuan, S., Lin, Q., Qiao, Y., and Liu, J.: Symbol-LLM: Towards foundational symbol-centric interface for large language models. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), eds. L.-W. Ku, A. Martins, and V. Srikumar, pages 13091–13116, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

Persuasion?

Speaker	Utterance
PE	*You have to ask. Um, exercise, regularly. You know, it sounds with this one to two miles you're walking on a daily basis, we're going to get you back up to that.*
Patient	Okay.
PE	*That's a great way to keep that going. There's no reason to stop, once we get you feeling better. Um, it used to be back in the day, maybe 20 years ago, people would say, "Well, you know, I've got to take it easy." That's not the case with heart failure. We want you to get up where you can do it. We don't want you to push yourself. . .*
Patient	Right
PE	*If you're short of breath, but. . . and then, we want you to check your weight every day. Do you own a scale?*

Patient-educator conversation excerpt illustrating the use of persuasive communication

Persuasion Strategies

provide_insulin_information	Provide information in response to a question on insulin or diabetes.
ask_concerns	Ask about concerns related to insulin
propose	Suggest trying insulin
personal_related_inquiry	Ask about some personal information related to the context
task_related_inquiry	Ask about desire to try insulin for better diabetes control
logical_appeal	Provide logical reasoning to trying insulin
emotion_appeal	Emotionally appeal to why they should try insulin
credibility_appeal	Use research studies to convince why they should try insulin
ask_about_consequence	Ask about the result of the described action or situation
ask_about_antecedent	Ask about the reason or cause of the described state or event.
ask_for_confirmation	Confirm the agreement to try insulin
suggest_a_solution	Provide a specific solution to a problem in a form of a question
suggest_a_reason	Suggest a specific reason or cause of the event or state described by the speaker in a form of a question

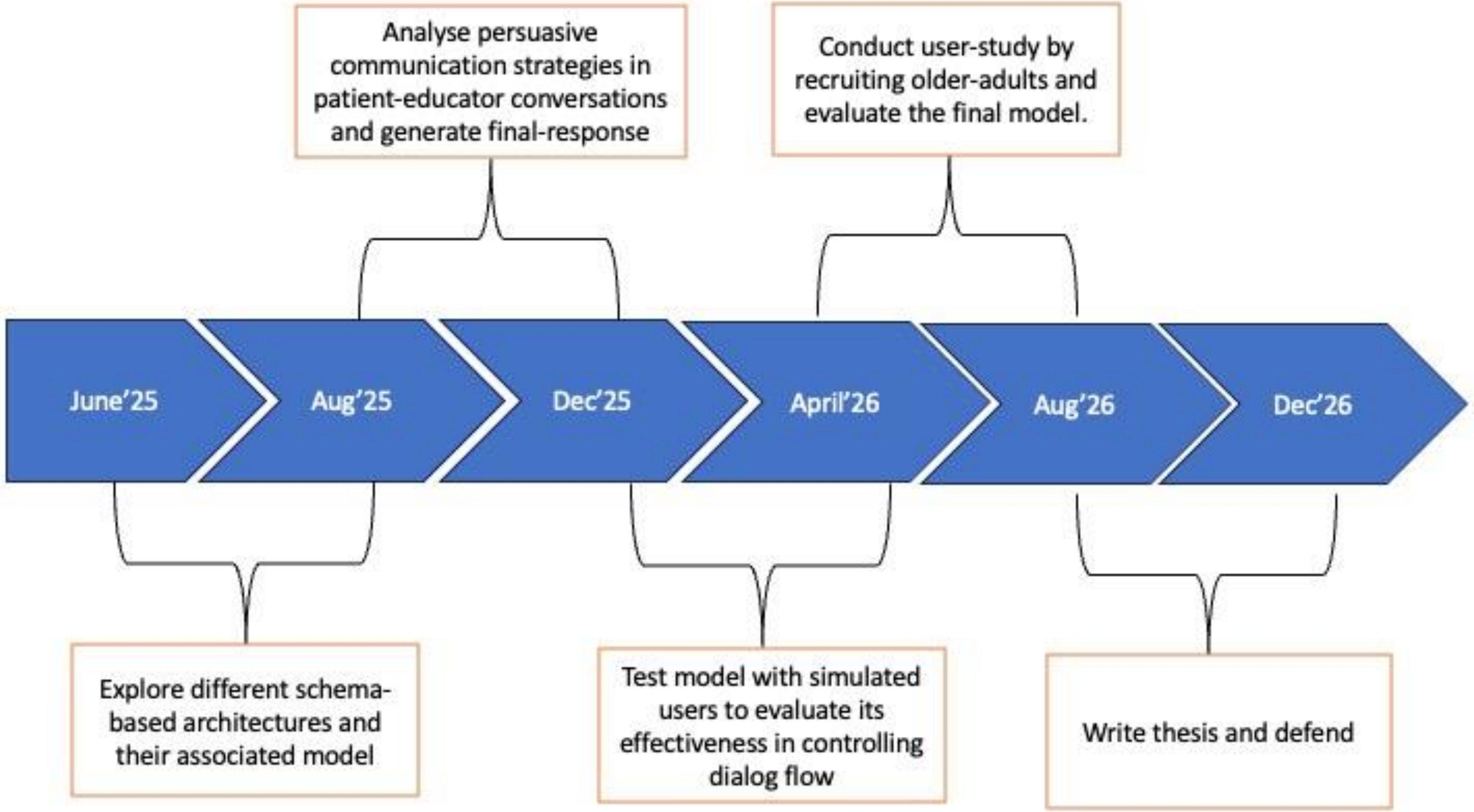
Proposed Work

- Our core hypothesis is that an exercise dialog system can be effective for patients when it is actionable and can adapt to both communication strategies and reading level.

Evaluation

- **Automated Metrics:** Joint Goal Accuracy, inform and success rate.
- Test on simulated users
- User-study with older adults.

Timeline


Analyse persuasive communication strategies in patient-educator conversations and generate final-response
Conduct user-study by recruiting older-adults and evaluate the final model.
June'25
Aug'25
Dec'25
April'26
Aug'26
Dec'26
Explore different schema-based architectures and their associated model
Test model with simulated users to evaluate its effectiveness in controlling dialog flow
Write thesis and defend

Thank You