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

Anuja Tayal[1], Devika Salunke[2], Barbara Di Eugenio[1], Paula G. Allen-Meares[3],
Eulalia P. Abril[4] , Olga Garcia-Bedoya[3], Carolyn Dickens[3] , Andrew D. Boyd[2]
University of Illinois Chicago, IL, USA
[1]Department of Computer Science, [2]Department of Biomedical and Health Information Sciences,
[3] Department of Medicine, [4] Department of Communications

Introduction

- LLM based dialog systems - difficult to evaluate
  - Do not operate within rigid strict boundaries
  - Lack transparency regarding data source
  - Fail to reliably follow user prompts
- Limitations are critical when facilitating medical conversations
- Human Evaluation remains gold standard [1]
- Need for controlled probing evaluations with real stakeholders
- **Goal**: conduct within-group user study to compare 2 versions
  - **HFFood-GPT** - based on GPT-4 [2]
  - **HFFood-NS**- Task Oriented Dialog System (TODS) with neuro-symbolic architecture

User Study

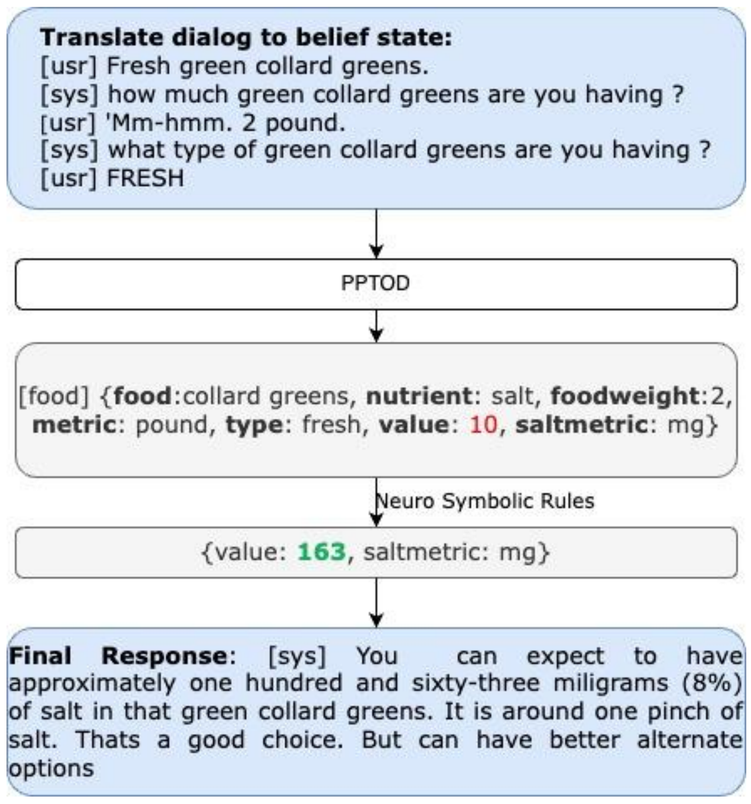
- 20 (13 Males, 7 Females) African American Heart failure hospitalised patients
- Within-subject design - patients asked 5 food questions to both systems
- **Conversational**: Systems can ask clarification questions
- All interactions - oral- consent , system interaction
- **Pre-Survey Questionnaire**- Health Literacy, Digital Health Literacy
- **Post-Survey Questionnaire** - Feedback, preference, helpfulness, likelihood of use

HFFood-GPT - Zero Shot GPT-4 based prompt

- *asks users about the **food type, cooking method, and portion size, one question at a time, to accurately determine the salt content.***
- *calculates the estimate salt content and compares it to the recommended daily intake of 2000mg.*
- ***refrains from giving health advice and suggesting from consulting a professional for dietary guidance.***
- *Answers are **kept under 40 words** ,it only searches the data provided in the JSON*
- ***Users do not know about the data file, so don't discuss it**.*

HFFood-NS

- T5 based - with neuro-symbolic rules [3]
- Ability to add fail-safe
- System responses were template-based

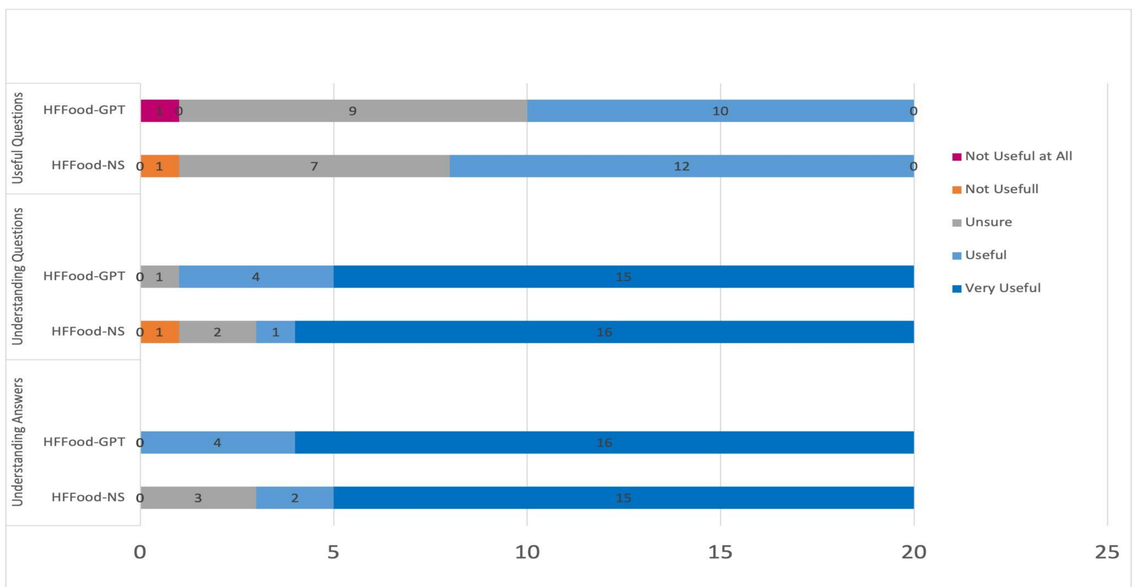


HFFood-NS with Interaction

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%
Salt Value Accuracy	37%	24%
Slot Accuracy	56%	89%

Intrinsic Evaluation: Comparison of 2 systems



Extrinsic Evaluation of 2 systems

- **Preference**- 11 HFFood-NS, 9 - HFFood-GPT.
- 11 participants (55%) preferred informal terms like pinches or dashes.

	HFFood-NS
Missed Slot	27
Wrong food identified	9
System Error	8
Internet	6

Error Analysis of HFFood-NS

	HFFood-NS	HFFood-GPT
Allow Error Analysis	✓	x
Reliable	✓	x
Handles Complex Query	x	✓
Fluent	x	✓
Concise	✓	x
More Constrained	✓	x
Faster to Deploy	x	✓

Pros and Cons of 2 systems

Conclusions

- In-house system is more accurate, completes more tasks, less verbose,
- ChatGPT: makes fewer speech errors, requires fewer clarification questions, handles complex query more effectively
- Greater Control on LLMs needed; Neuro-symbolic methods offers solutions

Acknowledgement

DPI Cycle 1 Seed Funding Program Award, NSF award IIS 2232307

References

[1] M A Walker, D J Litman, C A Kamm, A Abella. 1998. Evaluating spoken dialogue agents with paradise: Two case studies. Computer Speech & Language
[2] GPT-4 - Technical Report
[3] A Tayal, B Di Eugenio, D Salunke, A D. Boyd, C Dickens, E. Abril, O Garcia, P A-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 24, Italia. ELRA and ICCL