A Chatbot for Activities of Daily Living Assessment Standardization

Systems Demo - Clinical Informatics

S35

Zhecheng Sheng, Raymond Finzel, Aditya Gaydhani, Michael Lucke, Sheena Dufresne, Maria Gini, Serguei VS Pakhomov

University of Minnesota, Twin Cities

#AMIA2023

Disclosure

I and my spouse/partner have no relevant relationships with commercial interests to disclose.

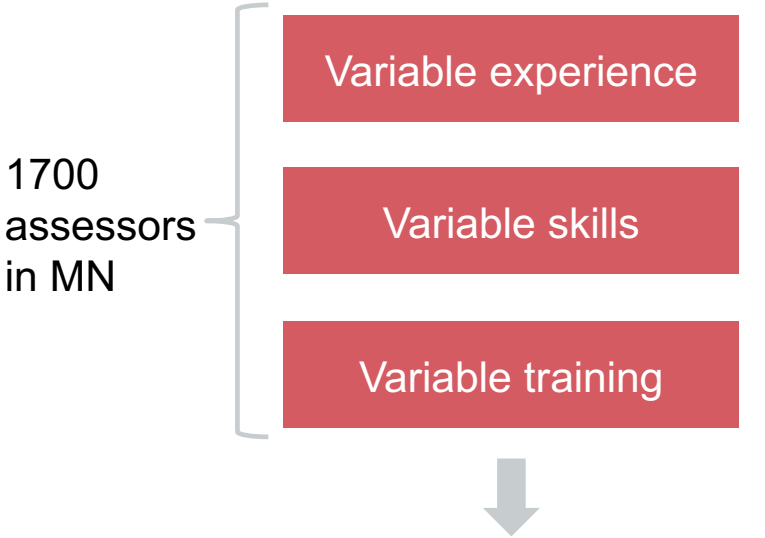
Learning Objectives

After participating in this session the learner should be better able to:

- Understand the challenges associated with assessing individuals for their ability to function in daily lives.
- Explain how an AI-driven chatbot designed to simulate interactions between health assessors and individuals with varying levels of disability can be used to train health assessors.
- Explain the advantages and limitations of integrating pre-trained large language models in a conversational agent designed to impersonate individuals with varying degrees of functioning.

Motivation

The Minnesota Department of Human Services (DHS) conduct state-wide assessments for eligibility for human services (e.g., personal care assistance).



A coaching system would be beneficial to:

- Train new assessors get prepared for real-world scenarios
- Help standardize the assessment results from experienced assessors

Domains of functioning

Broad areas of functioning

- Self-care
- Household Management
- Meals and Meal Prep
- Mobility

Case identification

- Public resources
- Specific needs
- Verbal assessment

ADLs currently covered by the conversational agent

1. dressing 2. grooming 3. bathing 4. toileting 5. incontinence accident management 6. house-keeping light 7. housekeeping heavy 8. laundry 9. finance 10. food consumption 11. meal preparation 12. meal planning 13. mobility 14. transfer 15. mode of transfer 16. positioning 17. mode of positioning 18. fine motor skills

Verbal Interactions with Individuals

80-year-old male

I must get my wife or a caregiver to help me bathe. I can't do it myself.

She helps me wash, dry off, and put on my clothes. She also helps me shave.

Tell me about how bathing goes for you?

What sort of help does your wife give you?

Assessor

Data Collection

Data Sources

The knowledge base relies on the data from 10,000 historical assessments provided by DHS.

- Each assessment is fully anonymized and contains only basic demographics.
- Each assessment contains details about how the individual performs in each of the 18 ADL domains.
- Short assessor notes are also included to describe the assessed individual‘s difficulties, preferences or devices used. These are frequently abbreviated or entered in shorthand notation.

Data Collection

Synthetic Profile

We selected 10 anonymized assessments and creates synthetic profiles based on demographics and enhanced assessor notes

- Independence level is translated into numerical ratings
- Populated with intents and action types based on short note
- Translate short notes into conversational responses

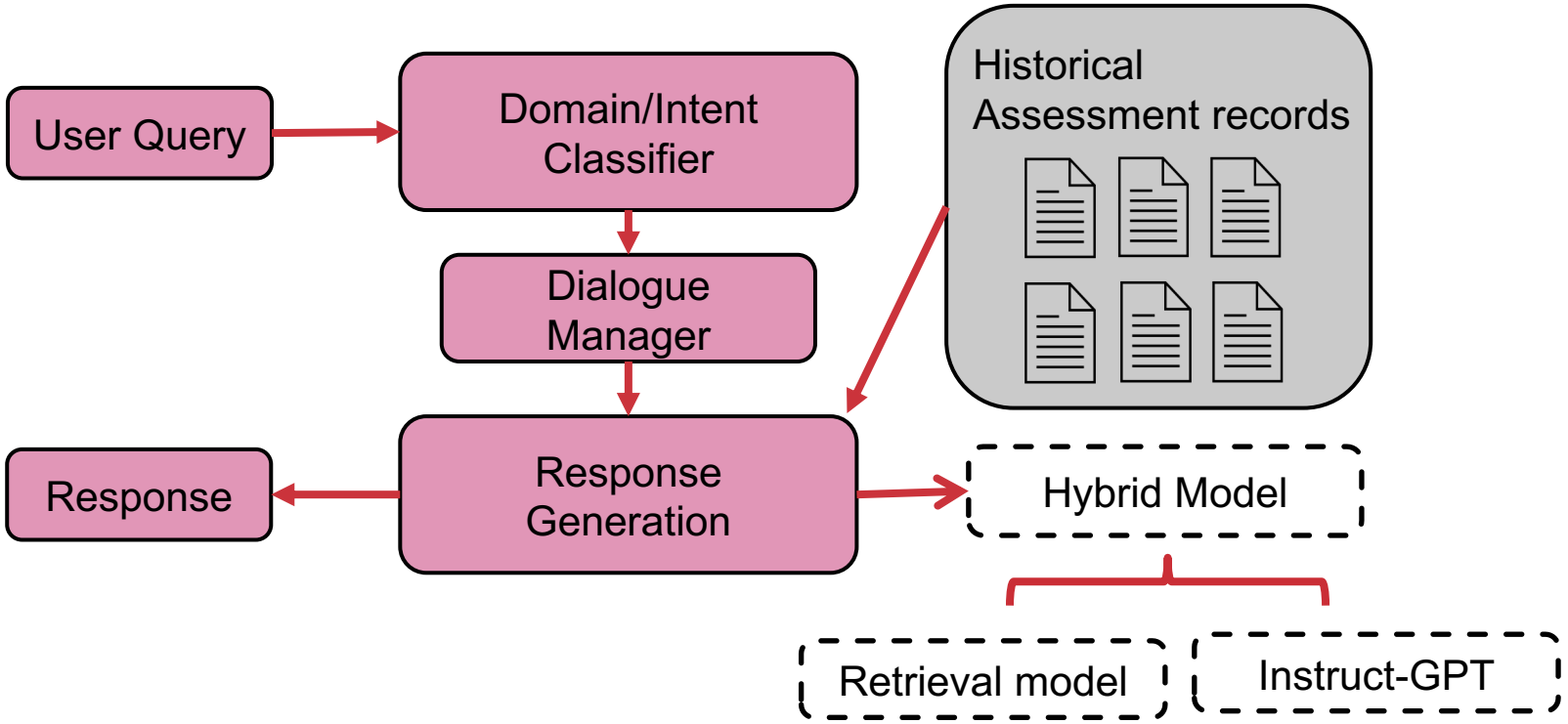
Assessor note:

Prefers shower

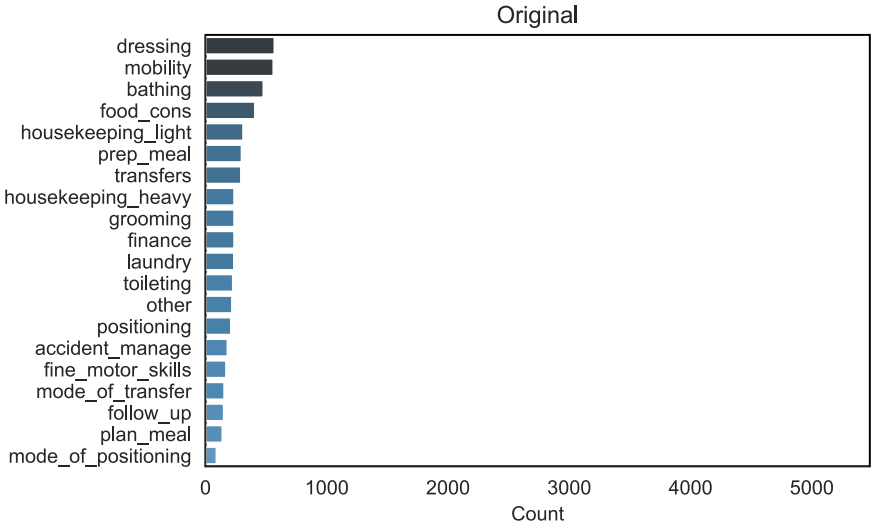
I do not like baths, I prefer to shower. I like taking long showers…

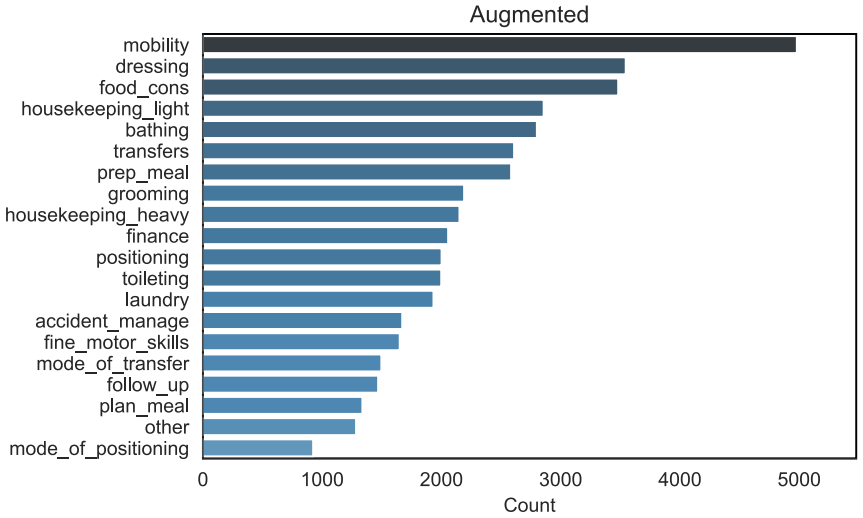
ID	Age	Gender	Avg rating	#utterances
3b1	27	Female	3.41	252
3b108	64	Male	2.73	259
3b77	71	Female	3.23	196
3b84	84	Male	2.57	148
3b86	52	Male	3.53	206
4d18	86	Female	3.58	233
4d23	60	Male	3.78	114
4d26	96	Female	3.54	81
4d29	42	Female	1.74	50
4d4	63	Female	3.07	213

Conversational Agent



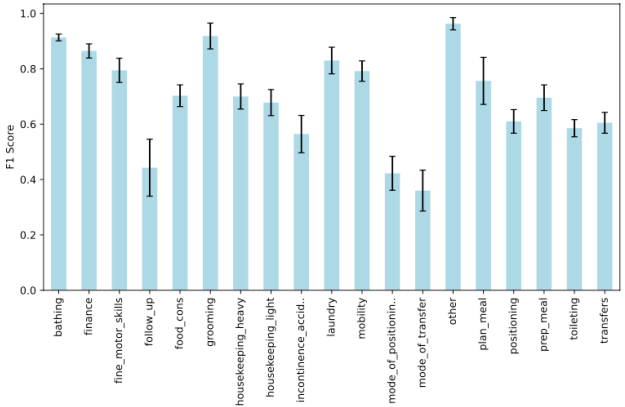
NLU: Query classification



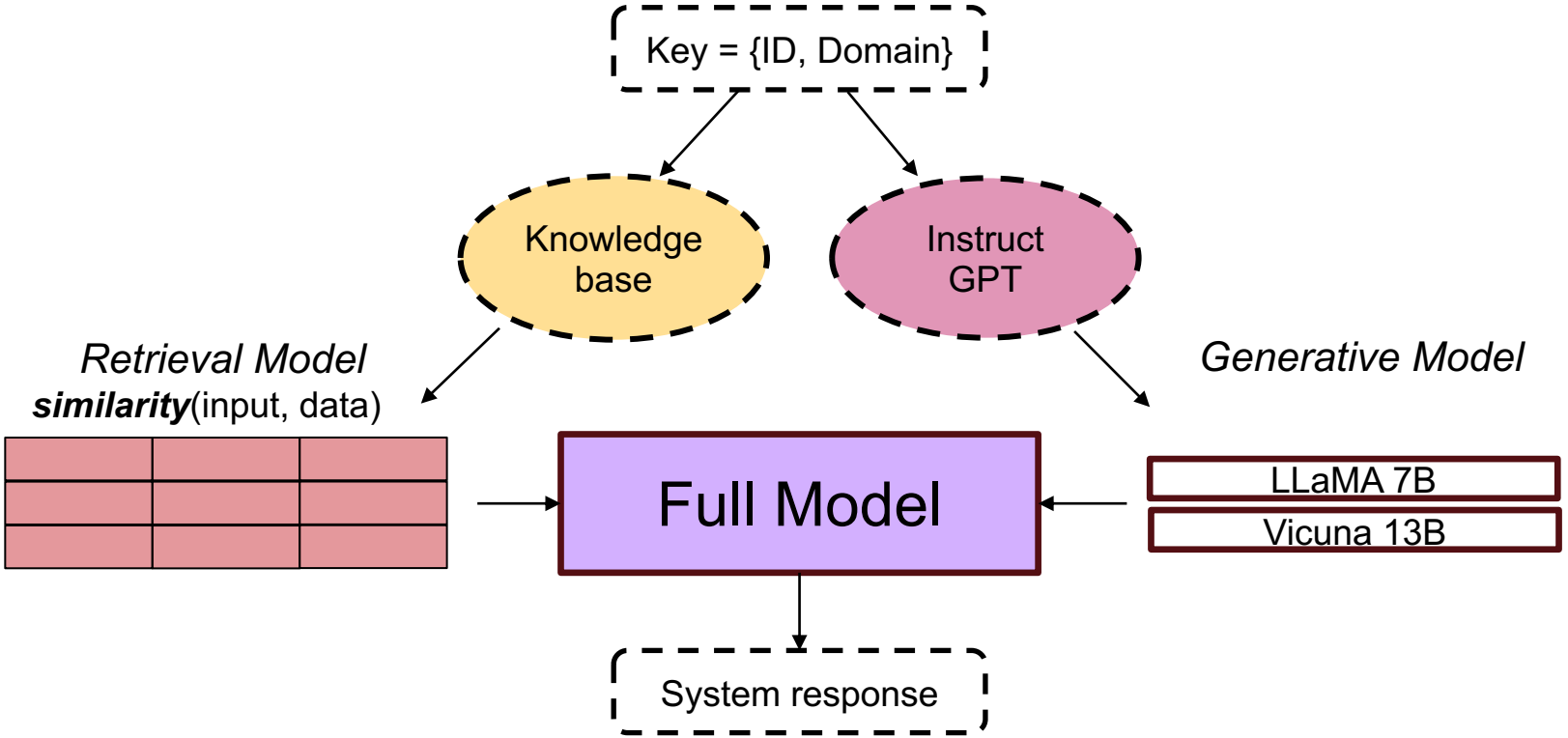


Classification Results

Experiments	Accuracy	F1-weighted	F1-micro	F1-macro
LR + Original	$0.703_{(0.702-0.704)}$	$0.708_{(0.707-0.709)}$	$0.703_{(0.702-0.704)}$	$0.606_{(0.604-0.608)}$
LR + Augmented	$0.696_{(0.694-0.705)}$	$0.702_{(0.700-0.711)}$	$0.696_{(0.694-0.705)}$	$0.615_{(0.613-0.623)}$
BERT$_{base}$ + Original	$0.747_{(0.729-0.760)}$	$0.744_{(0.727-0.756)}$	$0.747_{(0.729-0.760)}$	$0.649_{(0.635-0.670)}$
BERT$_{base}$ + Augmented	$0.726_{(0.720-0.733)}$	$0.729_{(0.723-0.738)}$	$0.726_{(0.720-0.733)}$	$0.639_{(0.630-0.651)}$
RoBERTa$_{base}$ + Original	$0.759_{(0.745-0.767)}$	$0.757_{(0.740-0.766)}$	$0.759_{(0.745-0.767)}$	$0.667_{(0.629-0.698)}$
RoBERTa$_{base}$ + Augmented	$0.727_{(0.720-0.732)}$	$0.731_{(0.725-0.737)}$	$0.727_{(0.720-0.732)}$	$0.641_{(0.633-0.648)}$
DeBERTa$_{v3}$ + Original	$\mathbf{0.762}_{(0.752-0.782)}$	$\mathbf{0.759}_{(0.746-0.781)}$	$\mathbf{0.762}_{(0.752-0.782)}$	$\mathbf{0.683}_{(0.652-0.708)}$
DeBERTa$_{v3}$ + Augmented	$0.732_{(0.728-0.738)}$	$0.736_{(0.732-0.741)}$	$0.732_{(0.728-0.738)}$	$0.646_{(0.643-0.651)}$



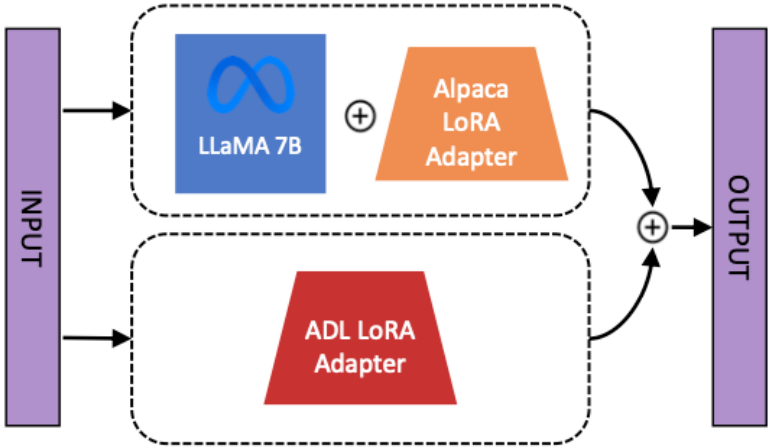
NLG: Response generation



Prompt design & Finetuning

Prompt for all Query:

> Converse briefly about {domain} as if you {plain English functioning} and you are {age} {gender}.



Internal Functional Testing

Survey I: Fixed Questions

Model	Sensibleness	Specificity	Realness	Favorite
Fine-tuned LLaMA 7B	3.67	3.92	1	1
Zero-shot Vicuna 13B	4.50	5.00	0	1
Full module w/ LLaMA 7B	4.92	4.33	5	4

Survey II: Adaptive Questions

Model	Contradict to KB	Contradict to History
Fine-tuned LLaMA 7B	4	1
Zero-shot Vicuna 13B	5	2
Full module w/ LLaMA 7B	1	0

Preliminary Usability Evaluation

- Actual intended users of the system
  - DHS assessors
- Metrics for each system response:
  - Sensibleness, Specificity
- Metrics for overall system usability
  - Ease of use, Understandable, Enjoyable, Realistic, Time, Satisfaction, Recommend to another

Results of preliminary Evaluation

- 299 responses triggered from 42 sessions
  - Average sensibleness: 2.71 out of 4
  - Average specificity: 2.13 out of 4
- 12 surveys about the system are returned
  - Average rating for *Easiness* is 4 out of 5
  - Among other measurements (*Understandable, Enjoyable, Realistic, Time, Satisfaction, Recommend*), the ratings are within 2.6 - 3.6

Limitation and Future Direction

- System performance largely depends on classification accuracy
  - Collect larger amount of utterances with minimum noise for training
  - Design a better prompt template and get rid of the query classification
- Large language model can be better fine tuned for conversation
  - Improved hardware to host biggest foundation models
  - Collect more example conversations for finetuning
- Hybrid model is sub-optimal
  - Adaptive retrieval
- More formal evaluation is in need
  - Develop formal manual/guidelines for using the web interface
  - Evaluate with well-defined metrics

Contributions

- Introduced a novel conversational dataset for ADL assessment
- Demonstrated a chatbot architecture that can be generalized in other health assessment tasks
- Illustrated a use case of large language model in healthcare to improve clinical practice

Acknowledgement

This work is supported by funding from the Minnesota Department of Human Services.

Thanks the people at DSD and MNIT for help with project specifications, gathering of historical data, and expert guidance on domain-specific aspects of the project. We would also like to thank Pamela Miller, Sidney Kiltie, and Elise Moore for help with transforming certified assessor notes to natural language format and Julia Garbuz for helping to develop and conduct the surveys of DHS assessors.

Useful Links

ADL conversational url: https://mndhs.rxinformatics.net/

Prior work on this project:

- https://ceur-ws.org/Vol-2760/paper2.pdf
- https://aclanthology.org/2021.eacl-demos.38.pdf
- https://aclanthology.org/2023.dialdoc-1.8.pdf

Vicuna: https://lmsys.org/blog/2023-03-30-vicuna/

LLaMA: https://ai.meta.com/llama/

Thank you!

Email me at:
sheng136@umn.edu