UNIVERSITY OF MINNESOTA

DEPARTMENT OF HUMAN SERVICES

61 ACL 2023 DialDoc

A Dialogue System for Assessing Activities of Daily Living: Improving Consistency with Grounded Knowledge

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

Cognitive AI Lab @

University of Minnesota, Twin Cities

Activities of Daily Living (ADL)

- Measure of functioning

  Cognitively

  Perceptually

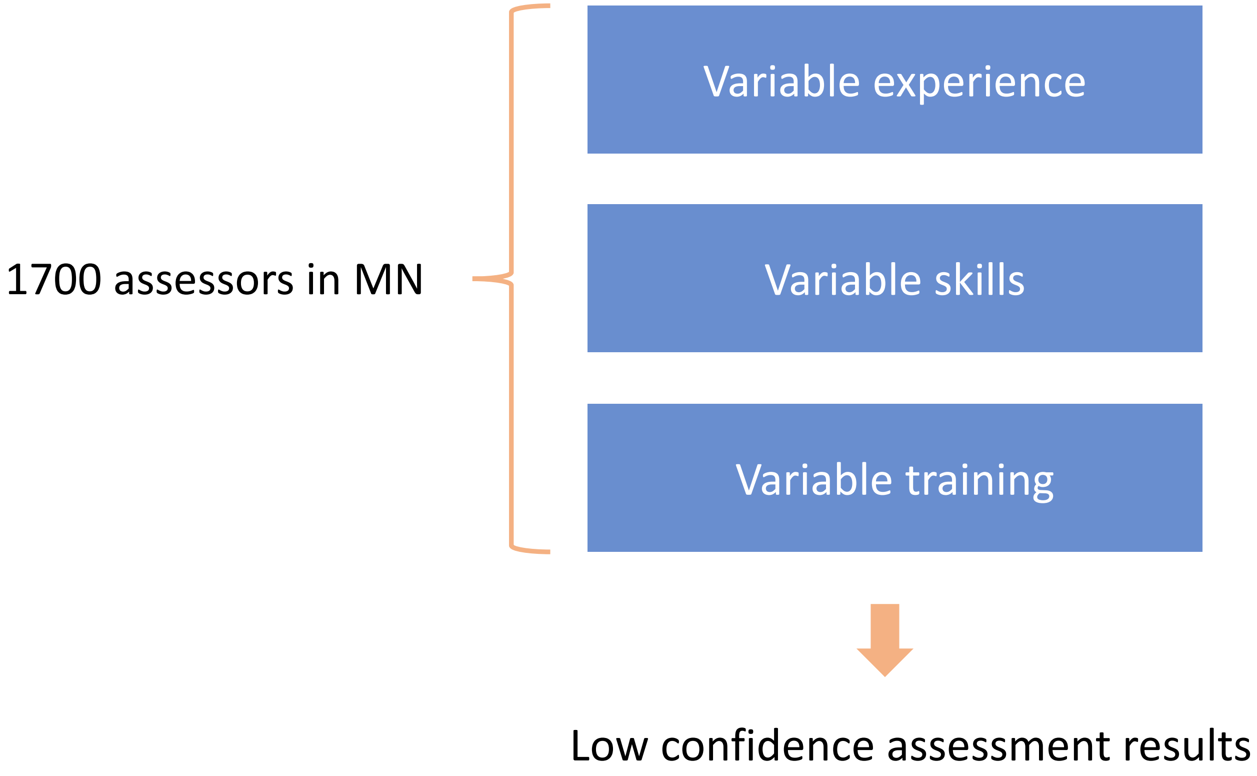
  Physically

- Case identification

  Require support

  Significant public resources

  Verbal assessment



1700 assessors in MN

Variable experience

Variable skills

Variable training

Low confidence assessment results

Interact with Participants

50-year-old, male

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

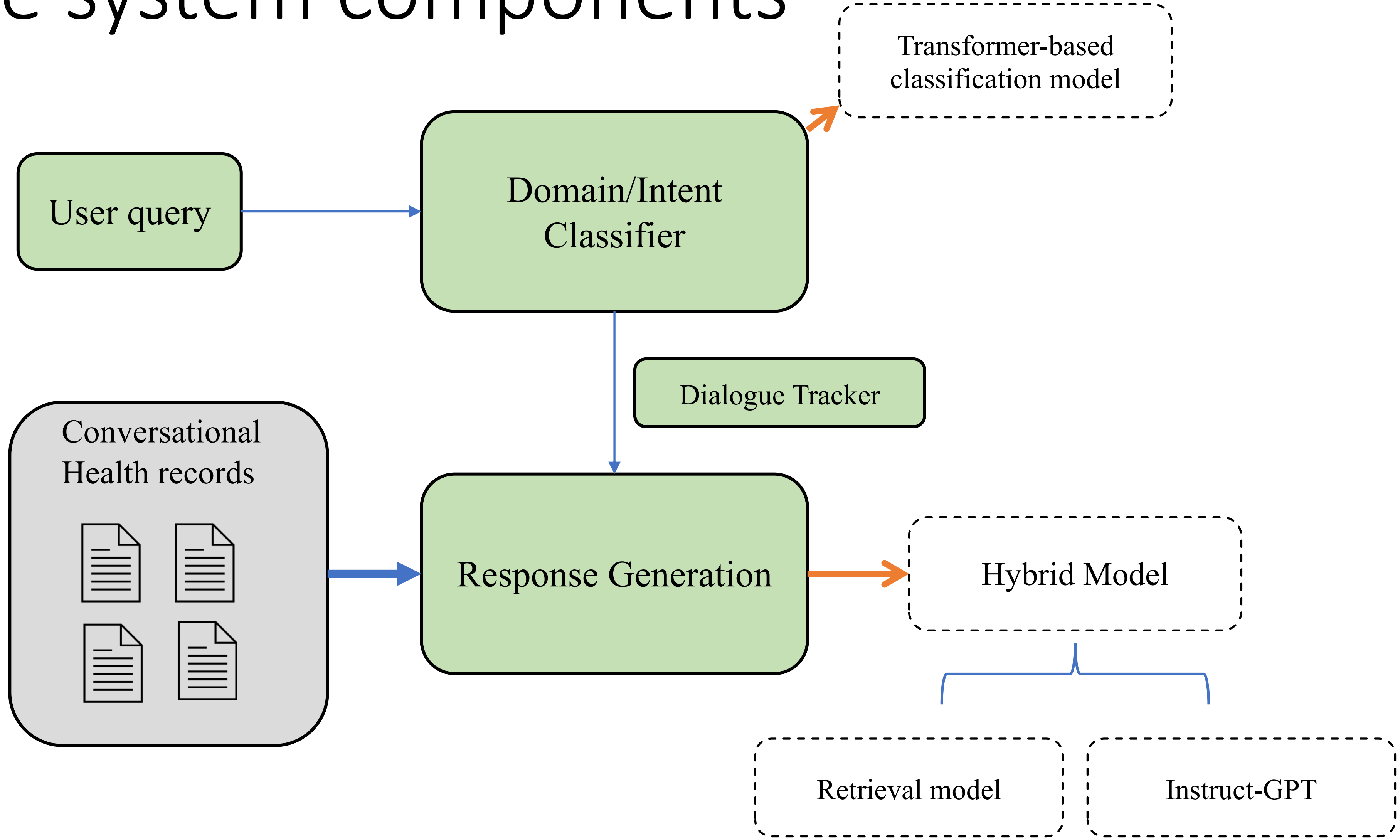
She helps me wash, dry off, and put on my clothes. She also helps me shave.

Assessor

Tell me about how bathing goes for you?

What sort of help does your wife give you?

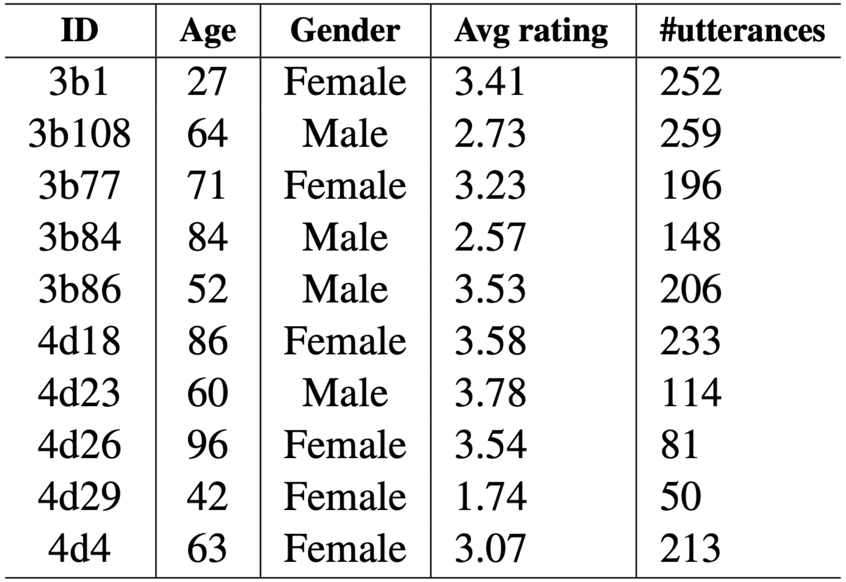
Dialogue system components



Data

- Fully anonymized personal records
- Collected from experienced certified assessors
    - 10000 historical assessments
    - Includes details about individual's ability for ADL
- Includes short notes during the interview
- Create synthetic profiles based on demographics and translated notes

Synthetic Profiles



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

Avg rating across domains: 1 is independent, 5 is completely dependent

Exp.

Prefer shower ➡ *I do not like baths, I prefer to shower.*

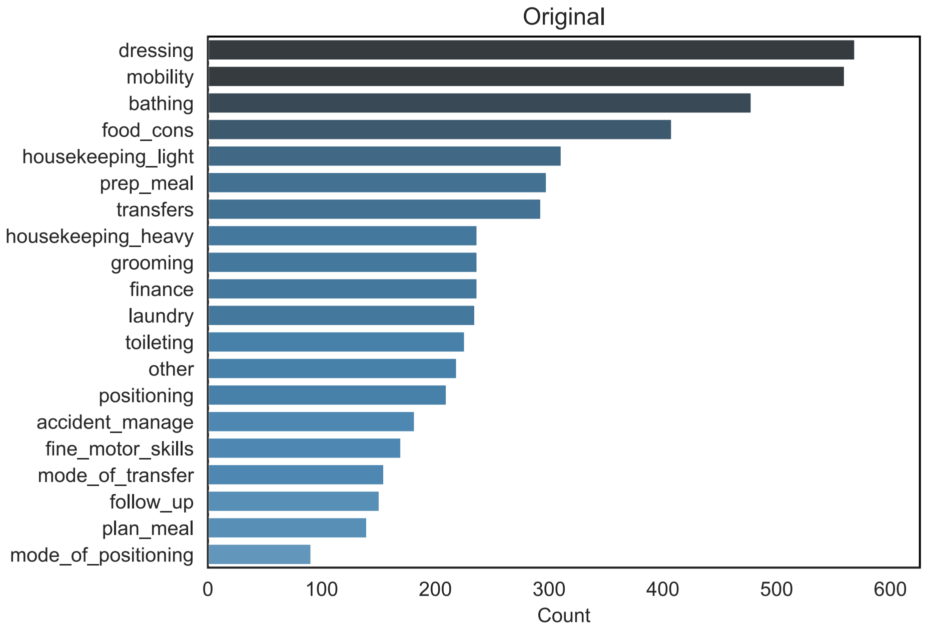
Factual Consistency with Knowledge

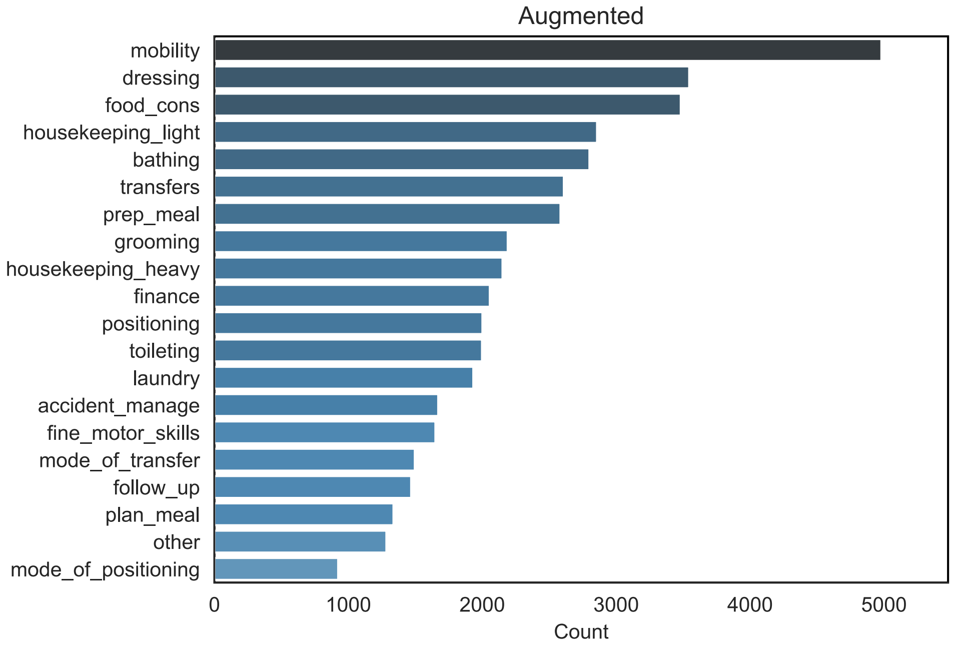
- The ability of a model to generate responses that are accurate and consistent with the information present in a verified knowledge base.
- Ungrounded language model can always generate hallucinations
- Factual consistency is important for tasks requiring accurate information. (Q&A, dialogue systems, chatbot, etc.)

Query classification

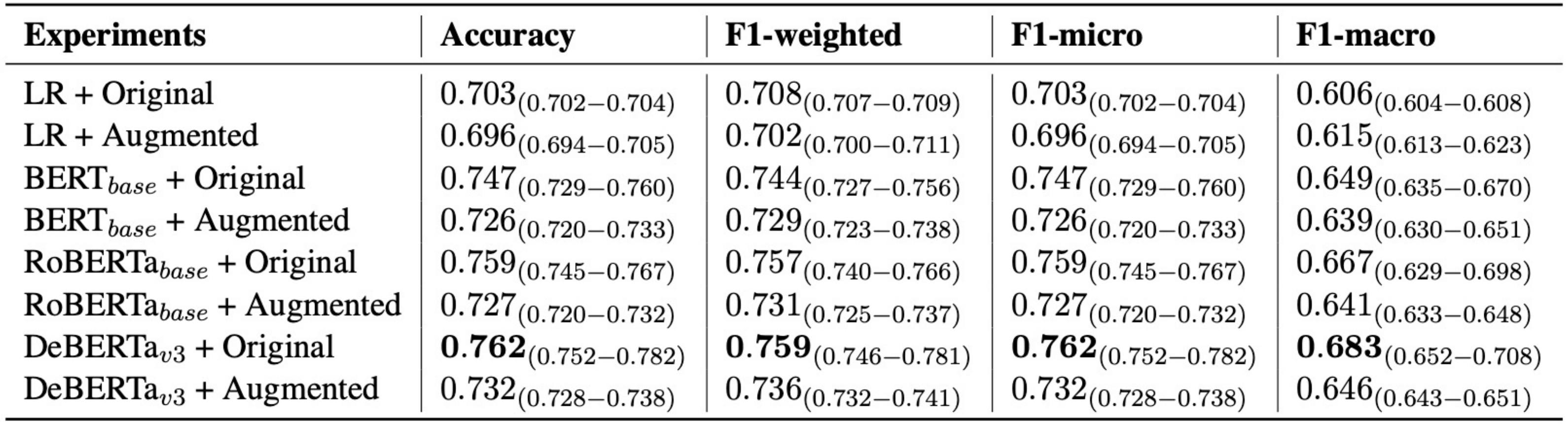
ADL to consider

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 prepa- ration 12. meal planing 13. mobility 14. transfer 15. mode of transfer 16. positioning 17. mode of positioning 18. fine motor skills



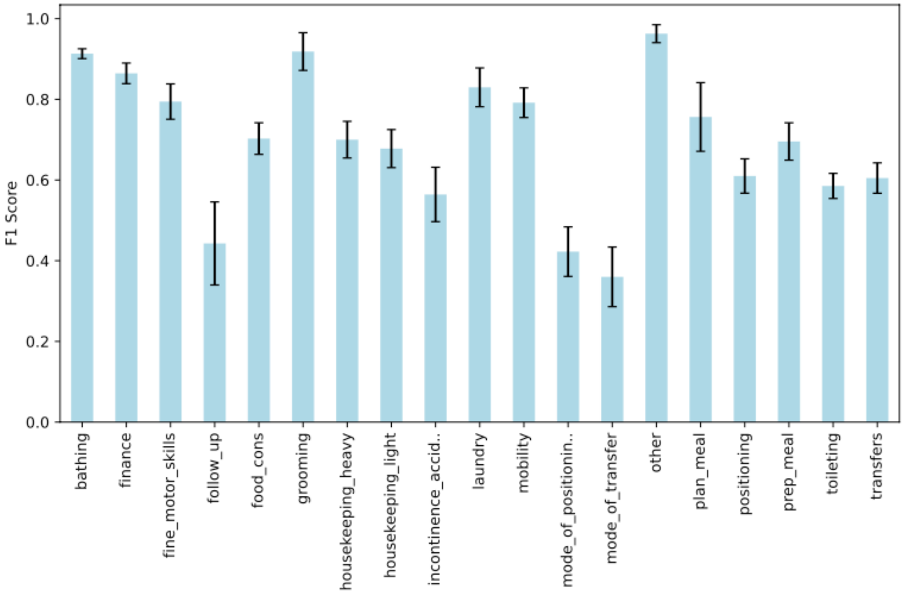


Classification Results

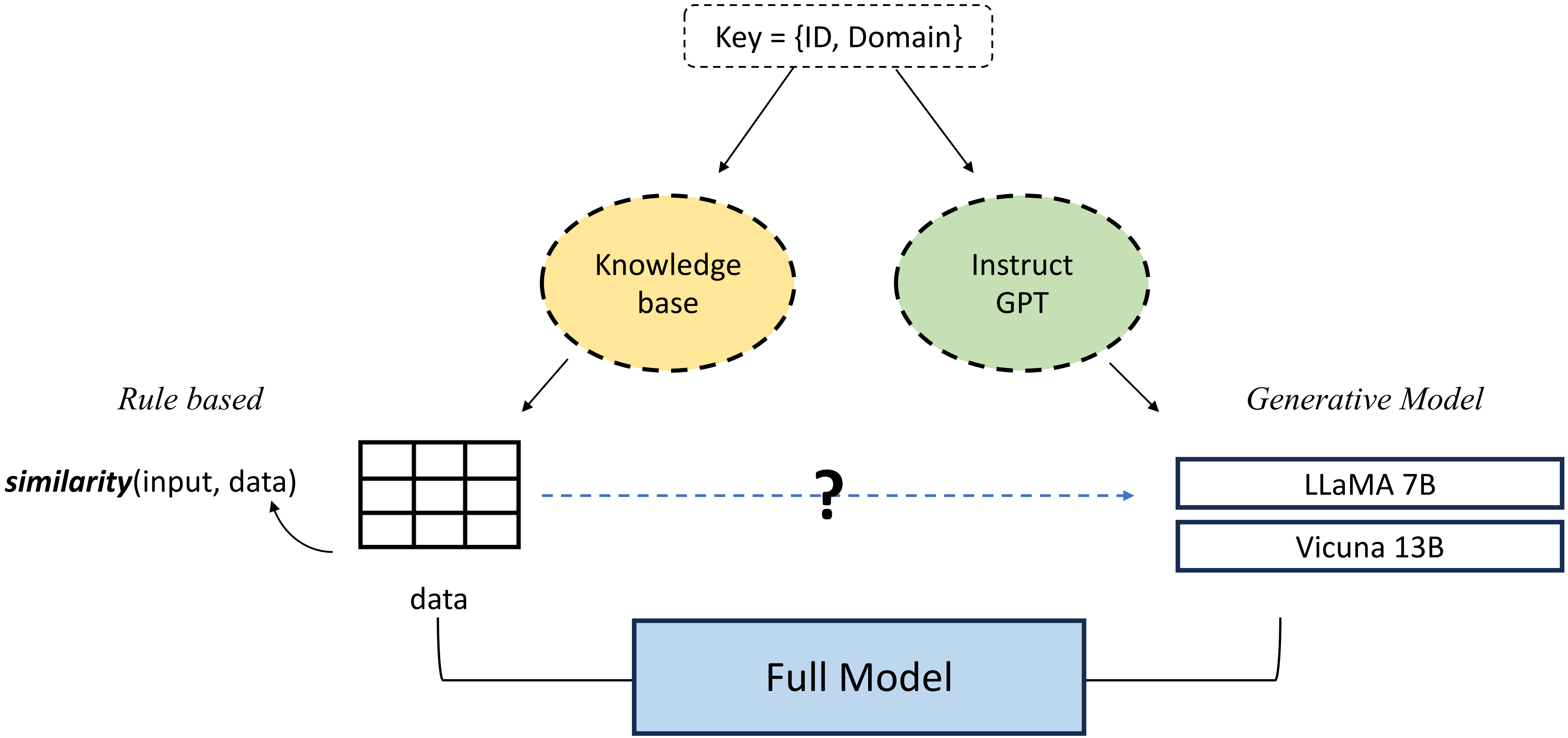


Experiments	**Accuracy**	**F1-weighted**	**F1-micro**	**F1-macro**
LR + Original	0.703$_{(0.702-0.704)}$	0.708$_{(0.707-0.709)}$	0.703$_{(0.702-0.704)}$	0.606$_{(0.604-0.608)}$
LR + Augmented	0.696$_{(0.694-0.705)}$	0.702$_{(0.700-0.711)}$	0.696$_{(0.694-0.705)}$	0.615$_{(0.613-0.623)}$
BERT$_{base}$ + 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	**0.762**$_{(0.752-0.782)}$	**0.759**$_{(0.746-0.781)}$	**0.762**$_{(0.752-0.782)}$	**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$_{(0.643-0.651)}$

Class-wise F1 score across all experiments



Response generation



Prompt design & Finetuning

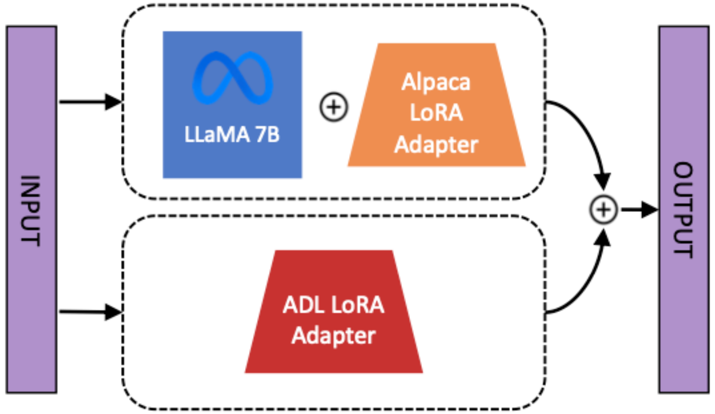
General Prompt:

Write your next response in the following conversation about {domain} as if you {plain English functioning} and you are {age} {gender}.

Follow-up Prompt:

Provide more details to this statement about {domain} as if you {plain English functioning} and you are {age} {gender}.

Finetune via Low-Rank Adaptation



Survey Results

Survey I: **Fixed question**

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 with LLaMA 7B	4.92	4.33	5	4

Survey II: **Adaptive question**

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

Conclusion & Limitation

- Introduce a novel conversational dataset for ADL assessment
- Preliminary evaluation shows combining knowledge base with generative model can improve factual consistency
- Accuracy of Domain/Intent Classification is essential to guarantee the quality of response

Limitation

- Formal framework needs to be designed to enable large-scale of human evaluation and quantitatively comparison is needed for different system iterations.
- More data is needed for minor domains for classifier training
- Current hybrid mode for NLG is sub-optimal