Framework to Evaluate LLM's ability to Reason about Physical Disabilities from Medical Perspective

Atharva S. Kashyap (katharva@umich.edu), Patrícia Alves-Oliveira (robopati@umich.edu)

ROBOTICS UNIVERSITY OF MICHIGAN

ICRA

Robot Studio ATLANTA 2025

1 MOTIVATION

"People with Disabilities are left out of public spaces"

-Yasmin Keats, Director, Open Style Lab

- Physically Assistive Robots assist humans through physical interaction to perform Activities of Daily Living. There is an increase in adoption of assistive robots prompting the **need for personalization**.
- Personalization is the ability to adapt to user's specific needs & preferences.

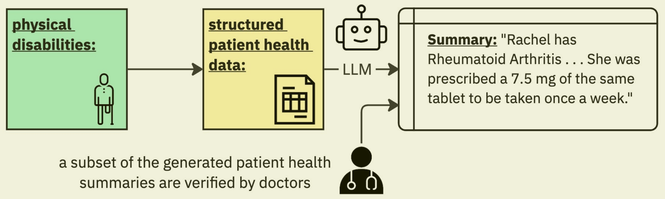
- LLMs seem to have the ability to deliver personalized solutions to users across various tasks and contexts [1].

Can LLMs reason about physical disabilities?

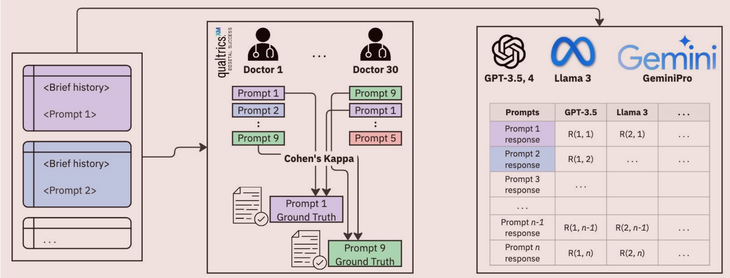
RELATED WORKS

- **Disability Models.** Disabilities are understood through several models (e.g., medical, social, religious) [2]. Medical model frames disability as a "condition" needing "treatment". We investigate whether LLMs can reason about physical disabilities through this lens to personalize assistive robots in healthcare.
- **LLM Evaluation Techniques.** There are several manual and automated evaluation techniques. We expect to adopt LLM-as-a-judge (DeepEval [3]) as an evaluation technique. This method employs prompt engineering to develop evaluation criteria.

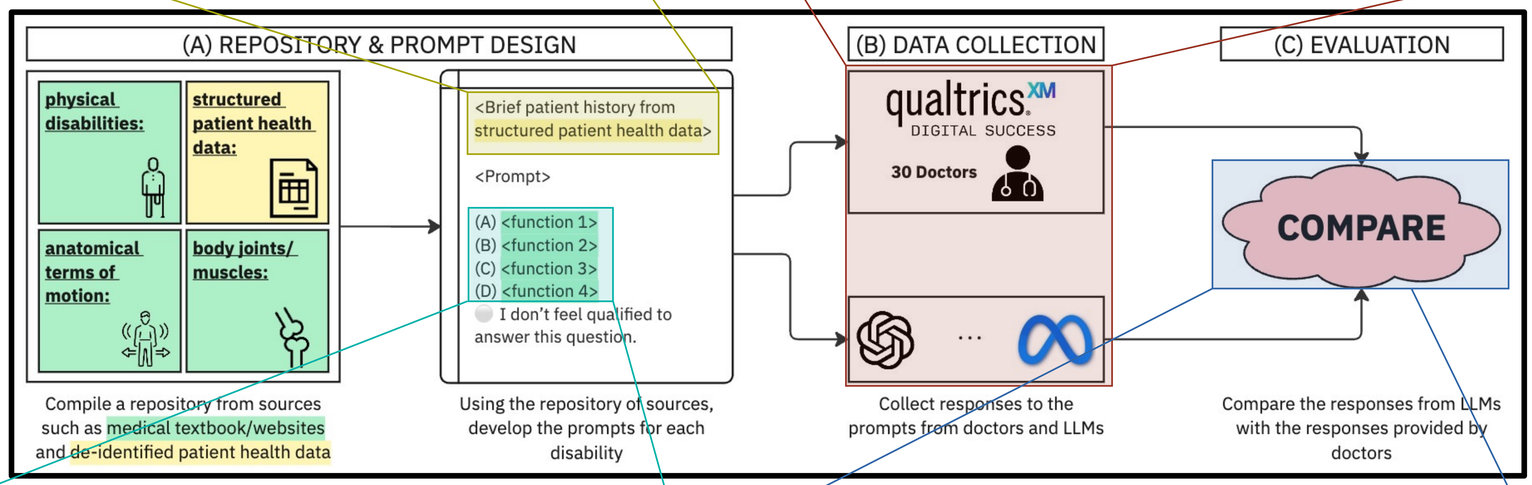
Obtainment of Medical History. To create realistic prompts, we used de-identified structured PHI from DataDirect [5] and employed Phi-3 Mini to generate natural language patient histories, a subset of which was reviewed by physicians for accuracy.



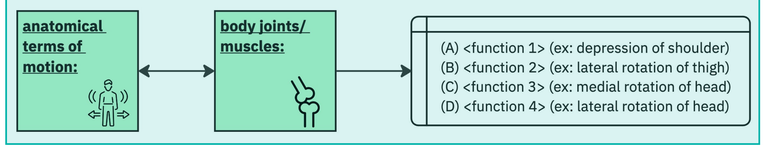
Data Collection. We expect to collect data for each prompt from various LLMs and at least 30 doctors and use Cohen's Kappa to develop a ground truth agreement across the responses from different doctors.



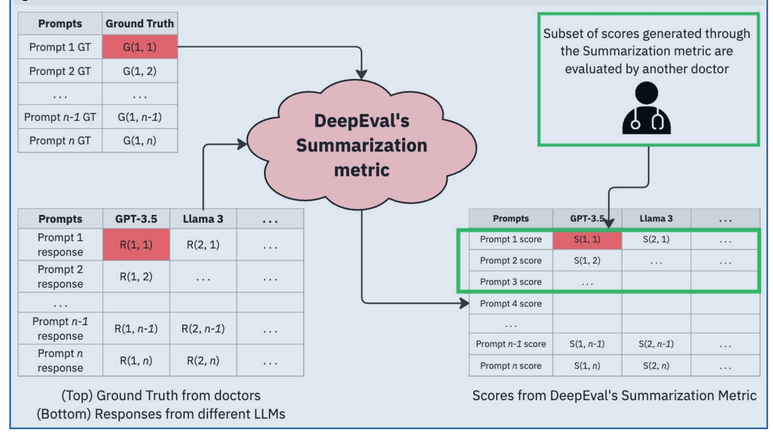
2



Anatomical Terms of Motion. A list of physical functions and relevant body joints/muscles were compiled and incorporated into the prompt choices.



Evaluation. We expect to use DeepEval [3]'s summarization metric to compare LLM response to the ground truth. Is the LLM response a good summary of the ground truth?



(Top) Ground Truth from doctors
(Bottom) Responses from different LLMs

Scores from DeepEval's Summarization Metric

3 LET'S DISCUSS!

- **Limitation.** Medical model of disability is not people-first and does not view disabilities as a lived experience, but rather as something rooted in facts and diagnoses [4]. We may want to consider other models, such as social model or affirmative model.
- **Limitation.** In this work, we define personalization as adapting to a user's needs and preferences. But, in Assistive Robotics, personalization extends beyond functional adaptation to include fostering a sense of warmth and personal connection during human-robot interactions.
- **Limitation.** Language, on its own, has limited visual capabilities. Incorporating VLMs can be another avenue to proceed.
- **Deployment.** What steps are needed to make this framework deployable and testable on assistive robots?
- **Translation to Action.** How can LLM insights be effectively converted into robot behaviors?
- **Evaluation.** Our current evaluation plan involves gathering expert annotations (e.g., doctors) to serve as ground truth for assessing LLM outputs. Are there alternative evaluation strategies that could reduce dependence on human annotators?
- **Privacy.** What safeguards can protect sensitive health data if this framework was deployed in robots?

References:
[1] A. Kabir, S. Shah, A. Haddad, and D. M. S. Raper, "Introducing our custom GPT: An example of the potential impact of personalized GPT builders on scientific writing," World Neurosurgery, vol. 193, p.461–468, Jan. 2025.
[2] M. Retief and R. Letsosa, "Models of disability: A brief overview", HTS Teologiese Studies/Theological Studies, vol. 74, no. 11, Oct. 2018. [Online]. Available: https://www.ajol.info/index.php/hts/article/view/177914
[3] C. AI, "Deepeval: The open-source llm evaluation frame-work", 2025, accessed: 2025-01-15. [Online]. Available: https://github.com/confident-ai/deepeval
[4] S. Brisenden, "Independent living and the medical model of disability", Disability, Handicap & Society, vol. 1, no. 2, p. 173–178, Jan. 1986.
[5] U. of Michigan Medical School, "Self-serve data tools", accessed: 2025-04-04. [Online]. Available: https://medresearch.umich.edu/office-research/about-office-research/our-units/data-office-clinical-translational-research/self-serve-data-tools
Acknowledgment: Image under the Motivation section was generated using Midjourney (https://www.midjourney.com/)