Generative AI Evaluation in Healthcare Systems and Implementation Standard GENAISIS Overview

INTRODUCTION

Large Language Models like GPT hold immense potential for application in the healthcare domain [1]. One hope is that they can alleviate administrative burden by automatically generating summaries based on patient records [2] [1], answering patient questions, and transcribing patient-physician consultations. To transform these use cases into viable clinical applications, guidance is needed on the key considerations and evaluation criteria across the LLM-development cycle.

OBJECTIVES

The primary objective of GENAISIS (**GEN**erative **AI** Evaluation in Healthcare **S**ystems and **I**mplementation **S**tandard) is to establish the key considerations and criteria for responsibly integrating generative AI, and notably Large Language Models (LLMs), into healthcare settings. This framework will define the different phases of the LLM lifecycle in healthcare and identify key criteria to consider for each phase. The outcome will be a comprehensive quality framework for the development, evaluation, implementation, and monitoring of LLMs in healthcare. The intended audience for this framework is researchers, developers, and clinicians who are developing LLM tooling or want to assess the validity of (commercially) available tooling for their own clinical setting.

METHOD

This framework will be developed using a modified Delphi method. The Delphi method compiles individual opinions to create a consensus using collective intelligence [3]. A list of candidate items, serving as key considerations for best practices within each phase, is developed based on literature reviews of LLMs or consortium members' recommendations. A diverse participant list is assembled to ensure the generalization of the guidance to all stakeholders.

At each Delphi round, participants are presented with items for best practice consideration in each phase. They are invited to rate their agreement level with the item's inclusion and comment on it. Items with high agreement (80% of experts agree) are retained. Items with low agreement are removed or rephrased. If necessary, an optional review meeting will be held to discuss and finalize the set of guidelines. The study result will be a publishable paper.

Reference

1. Lee, P., S. Bubeck, and J. Petro, *Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine.* N Engl J Med, 2023. **388**(13): p. 1233-1239.
2. Ali, S.R., et al., *Using ChatGPT to write patient clinic letters.* The Lancet Digital Health, 2023. **5**(4): p. e179-e181.
3. Nasa, P., R. Jain, and D. Juneja, *Delphi methodology in healthcare research: How to decide its appropriateness.* World J Methodol, 2021. **11**(4): p. 116-129.