Appendix A: REACH-LC (Relational Evaluation of AI in Clinical Horizons – Long Conversation) Proposal

Introduction

REACH-LC (Relational Evaluation of AI in Clinical Horizons – Long Conversation) is a proposed research/evaluation protocol that aims to assess the safety and trustworthiness of artificial intelligence systems throughout extended, multi-session dialogues. This protocol is designed to address the limitations of traditional short-turn evaluations, which typically focus on brief exchanges or conversations. REACH-LC uses failure type notation (e.g., FAB/MISATTR/OMIT/DECON/OVER/CAL/AGREE) and escalation logging approaches consistent with Table 5 and Box 3.

Key Features

- Longitudinal Safety Monitoring: The protocol emphasizes ongoing safety monitoring across multi-turn arcs, meaning it evaluates AI performance over multiple steps and extended conversations.
- Comprehensive Metrics: REACH-LC introduces metrics to evaluate compliance with approved methods, the timeliness of escalation when necessary, and the maintenance of relational safety throughout the interaction.
- Consensus Building and Pilot Testing: The approach utilizes Delphi-style consensus methods[51] and pilot testing, with an initial focus on geriatric psychiatry and other high-stakes domains, to refine evaluation criteria and processes.

Motivation and Scope

The proposal is motivated by public recognition of the weaknesses observed in long-conversation settings and by early empirical evidence suggesting that AI performance and safety-relevant reliability can degrade as context and interaction length increase.[48] Rather than presenting a finalized standard, REACH-LC is offered as a collaborative research/evaluation protocol, inviting further development and refinement through engagement with clinical, technical, and ethical experts. The intended focus is on applications in geriatric psychiatry and other domains where safety and trustworthiness are critical.