Mass General Brigham

Large Language Models in Dementia Care

Ipsit V. Vahia
Corrigan-Minehan Endowed Chair in Psychiatry
Chief, Division of Geriatric Psychiatry
Medical Director, Institute for Technology in Psychiatry
Director, Technology and Aging Lab

McLean Hospital
Harvard Medical School

www.geropsychtech.org
www.mcleangprp.org
LinkedIn: Ipsit Vahia

McLean Hospital | Confidential—do not copy or distribute

Background

Detecting Cognitive Decline

From speech

- GPT-3 embeddings separated Alzheimer's disease from controls and estimated cognitive scores from speech alone

From clinical notes

- In 6,945 MGB notes, GPT-4 beat Llama 2 but not models trained on local data; combining them did best

Agbavor & Liang, PLOS Digit Health, 2022; Du et al, eBioMedicine, 2024



Answering Caregivers' Questions

ChatGPT vs Google

- ChatGPT gave more relevant answers to 60 dementia queries; Google more often gave dates and sources

Real caregiver posts

- ChatGPT-3.5 answered 60 Reddit caregiver questions well on next steps but weakly on completeness

Hristidis et al, J Med Internet Res, 2023; Aguirre et al, JMIR Aging, 2024

New Tools

Caregiving model (CaLM)

- Small models grounded in caregiving knowledge performed reliably for ADRD caregivers

Clinical interviewer

- A voice agent detected symptoms with 82% sensitivity and 91% specificity in 25 patients (preprint)

Parmanto et al, JMIR Form Res, 2024; arXiv 2509.11478, 2025



The CARE-ALL Study

Why Caregivers, Why Now

The problem

- Dementia caregivers are highly burdened; an estimated 31–34% have depression
- Geriatric care managers (GCMs) are the standard of care, but scarce, costly, and not covered by Medicare or Medicaid
- Caregivers are already turning to chatbots

Study goals

- Compare a dementia-specific LLM with an experienced GCM in answering caregiver questions
- Test whether a GCM working with an LLM does better than either alone
- Develop an ethics framework and best practices for clinicians and caregivers

CARE-ALL Study Protocol

Main study

LLM vs geriatric care manager

1. Focus groups

16 family caregivers

6 groups on needs and challenges

2. Question bank

100+ real questions

Randomly cut to 75:

25 factual

25 problem-solving

25 emotional

3. Three answers

LLM · 13 min

GCM · ~6 wks

GCM + LLM · ~2 wks

4. Blinded ranking

Each caregiver ranks 15 question sets in REDCap

Guesses which source wrote each answer

5. Outcomes

Preference by question type

Source identification

Burden and prior LLM use

Ethics study

Modified Delphi + caregiver interviews

Expert panel

22 stakeholders: caregivers, clinicians, ethicists, researchers, engineers

Delphi round 1

Open-ended exploration

86 candidate items

Delphi round 2

Rate importance 1–5

35 items reach consensus

Delphi round 3

Re-rate and prioritize

+11 items = 46 total

Caregiver interviews

16 caregivers · scenario-based · thematic analysis

CARE-ALL Ethical Framework

7 domains

Pre-launch questions for developers, clinicians, researchers, IRBs

Caregivers and Their Questions

Who and how

- 16 caregivers, aged 22–83 (mean 51)
- Six baseline focus groups on experiences, challenges, and needs
- 100+ common questions, randomly narrowed to 75: 25 factual, 25 problem-solving, 25 emotional

Example questions

- **Factual:** “What tracking devices can I use to protect my loved one?”
- **Problem-solving:** “How do I get my care recipient to take their medication?”
- **Emotional:** “Is it OK to lie to my loved one?”

Three Ways to Answer 75 Questions

LLM alone

- Rippl's dementia-specific LLM
- 13 minutes for all 75
- Bulleted, instructional, action-focused

GCM alone

- McLean's geriatric care manager
- About 6 weeks alongside a full workload; 20–30 minutes per question
- Narrative and conversational, with her own insights

GCM + LLM

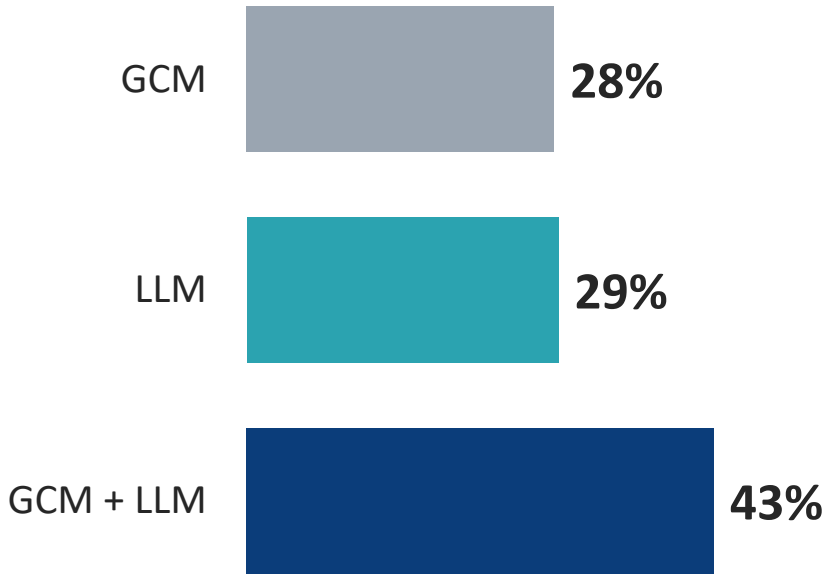
- The GCM started from the LLM's draft
- About 2 weeks
- She added insight and nuance to the LLM response

Blinded Ranking by Caregivers

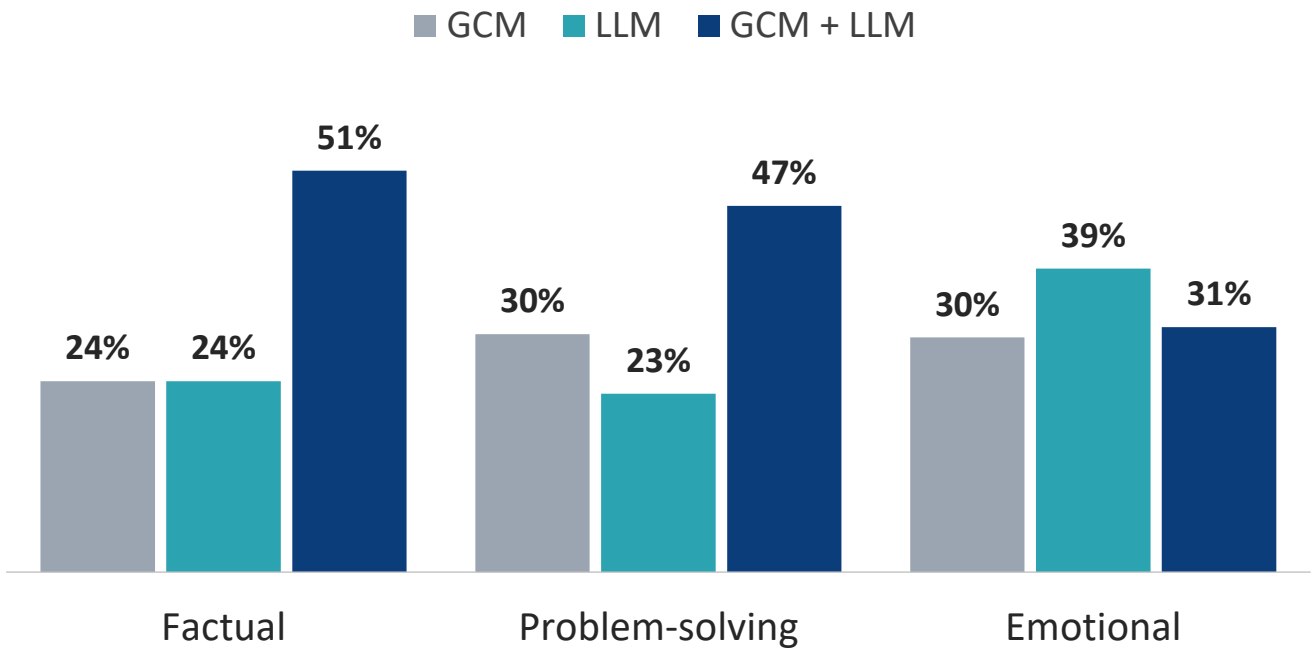
- Each caregiver saw 15 questions in REDCap, 5 from each theme
- Three responses per question, source hidden
- Ranked the responses from least to most helpful
- Guessed which response came from the LLM, the GCM, or both
- Rated confidence in spotting AI-written responses before and after

Caregivers Preferred the Hybrid

Overall



By question type



Share of rankings placing each response first. **vs GCM alone:** GCM + LLM OR 1.85 (1.46–2.33); LLM OR 1.46 (1.15–1.87)

237 rankings from 16 caregivers (preliminary). Analysis: Boyu Ren, PhD

The Time + Efficiency Factor

Our LLM generated responses **465,130% faster** than the GCM, highlighting substantial efficiency gains.

What Else We Learned

- **Hard to tell apart:** caregivers identified the source correctly only about 55% of the time
- **Confidence misled:** self-rated skill at spotting AI did not match actual accuracy
- **Burden shifted preference:** more burdened caregivers favored LLM answers for factual questions and GCM + LLM for problem-solving
- **Experience mattered:** caregivers new to LLMs preferred the GCM's answers
- **Style may drive preference:** bulleted LLM text is easier to read than narrative prose

Lessons from the Field

- The grant was written with one chatbot company; after funding and before IRB approval, it was acquired
- McLean held equity in the acquiring company, so a conflict-of-interest review delayed the study about 4 months
- After the study ended, the chatbot's owner was acquired again, and the bot is no longer available
- **Lesson:** commercial LLM partners can disappear mid-program; plan for it from the start

Building an Ethical Framework

Modified Delphi (3 rounds)

- 22 stakeholders: caregivers, clinicians, ethicists, researchers, engineers
- 86 candidate items → 35 reached consensus, then 11 more = 46
- Inclusion: median ≥ 4 and ≥ 70% rating “very” or “critically” important

Caregiver interviews

- 16 caregivers, including 2 professional caregivers
- Scenario-based: autonomous medication management; behavioral guidance
- Thematic analysis coded to ethics domains

Silverman, Yip, et al. (CARE-ALL)



What Stakeholders Prioritized

Delphi: top priorities

- Mitigate inaccurate or false information (45.5%)
- Regular review of outputs by human experts (45.5%)
- Clear line between supportive advice and clinical decisions (40.9%)

Caregivers: what surprised us

- Little trust in autonomous medication management; much more in behavioral guidance, especially with oversight
- 78.6% would share more personal data for more accurate advice
- Consent items did not reach consensus; caregivers likened LLM use to a support group

The CARE-ALL Ethical Framework

- **Clinical boundaries and crisis pathways:** advice, not clinical decisions; clear escalation
- **Accuracy and evidence:** evidence-based design; communicate uncertainty
- **Review and monitoring:** regular human expert review with defined triggers
- **Equity, bias, and accessibility:** reflect diverse caregiving realities
- **Autonomy and consent:** respect the dignity of the person with dementia; seek assent
- **Data protection:** treat data as protected health information
- **Long-term impacts:** track dependency, lost human contact, displaced services

Silverman, Yip, et al. (CARE-ALL)



Key Lessons from CARE-ALL

- Caregivers preferred LLM-assisted human answers over either alone
- LLMs did best on factual questions and least well on emotional ones
- Caregivers want accuracy, human oversight, and clear clinical boundaries
- The CARE-ALL framework gives developers, clinicians, researchers, and IRBs concrete questions to ask before launch

Summary – LLMs and Dementia

- The easy availability and rapidly-growing use of LLMs for application in medicine means that they are highly likely to have utility in dementia care.
- The available studies thus far point to a range of uses
- They may be particularly relevant for supporting caregivers.
- They seem too be most impactful when used in combination with a human
- Their speed of response exceeds human response.
- Accuracy of responses seems to be of most import.

THANK YOU

Mass General Brigham

ivahia@mclean.harvard.edu
617 855 3267

LinkedIn: Ipsit Vahia

The Technology and Aging Lab @ McLean: www.geropsychtech.org

McLean Geriatric Psychiatry Research Program: www.mcleangprp.org