

Generative AI for Dementia Care: Feasibility of AI-Powered Task Verification and Caregiver Support

Joy Lai, David Black, Kelly Beaton, Bing Ye, Alex Mihailidis

Submitted to: JMIR Aging
on: March 20, 2025

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on its website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressly prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Original Manuscript..... 5

Supplementary Files..... 27

 Figures 28

 Figure 1..... 29

 Figure 2..... 30

 Figure 3..... 31

 Figure 4..... 32

 Figure 5..... 33

 Figure 6..... 34

Generative AI for Dementia Care: Feasibility of AI-Powered Task Verification and Caregiver Support

Joy Lai¹; David Black²; Kelly Beaton²; Bing Ye^{1, 3}; Alex Mihailidis^{1, 3}

¹Institute of Biomedical Engineering Faculty of Applied Science and Engineering University of Toronto Toronto CA

²Engagement of People with Lived Experience of Dementia Consortium canadien en neurodégénérescence associée au vieillissement Toronto CA

³KITE Toronto Rehabilitation Institute Toronto CA

Corresponding Author:

Joy Lai
Institute of Biomedical Engineering
Faculty of Applied Science and Engineering
University of Toronto
Rosebrugh Bldg, 164 College St
Toronto
CA

Abstract

Background: Caregivers of people living with dementia (PLwD) face significant stress, particularly when verifying whether tasks are truly completed, despite the use of digital reminder systems. While PLwD may acknowledge reminders, caregivers often lack a reliable way to confirm task adherence. Generative AI, such as GPT-4, offers a potential solution by automating task verification through follow-up questioning and supporting caregiver decision-making.

Objective: This feasibility study evaluates an AI-powered task verification system integrated with a digital reminder framework for PLwD. Specifically, it examines (1) the effectiveness of GPT-4 in generating high-quality follow-up questions that help verify whether tasks were actually completed, (2) the accuracy of an AI-driven response flagging mechanism in identifying tasks requiring caregiver intervention, and (3) the role of caregiver feedback in refining system adaptability.

Methods: A theoretical AI-powered task verification pipeline was designed to enhance digital reminders by generating tailored follow-up questions, analyzing responses, and categorizing concerns. Each follow-up question corresponded to a specific reminder sent through the digital system, aiming to assess whether the task was genuinely completed. To test its feasibility, a simulated AI-powered pipeline was implemented using an anonymized dataset of 64 reminders. GPT-4 generated follow-up questions with and without additional contextual information about PLwD routines. A response classification system flagged task completion as High, Medium, or Low concern, based on response clarity and task urgency. Simulated caregiver feedback was incorporated to refine question quality and improve system adaptability over time.

Results: Contextual information and caregiver feedback significantly improved the clarity, specificity, and relevance of AI-generated follow-up questions. The response flagging mechanism demonstrated high accuracy, particularly for critical tasks such as safety-related reminders. However, subjective or non-urgent tasks posed classification challenges. Caregiver input iteratively enhanced system performance, ensuring a balance between automation and human oversight.

Conclusions: This study demonstrates the feasibility of integrating generative AI into dementia care for task verification and caregiver support. AI-generated follow-up questions provide a structured way to confirm whether tasks were truly completed after a digital reminder was acknowledged. The findings suggest that context-aware AI-generated prompts, combined with iterative caregiver feedback, improve task verification accuracy, reduce caregiver stress, and enhance PLwD support. Future research should focus on real-world implementation, longitudinal usability, and scalability to optimize AI-driven dementia care interventions.

(JMIR Preprints 20/03/2025:73641)

DOI: <https://doi.org/10.2196/preprints.73641>

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.

Only make the preprint title and abstract visible.

No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain visible to all users.

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <http://www.jmir.org/>

No. Please do not make my accepted manuscript PDF available to anyone.

Original Manuscript

Generative AI for Dementia Care: Feasibility of AI-Powered Task Verification and Caregiver Support

Abstract

Background: Caregivers of people living with dementia (PLwD) face significant stress, particularly when verifying whether tasks are truly completed, despite the use of digital reminder systems. While PLwD may acknowledge reminders, caregivers often lack a reliable way to confirm task adherence. Generative AI (Artificial Intelligence), such as GPT-4, offers a potential solution by automating task verification through follow-up questioning and supporting caregiver decision-making.

Objective: This feasibility study evaluates an AI-powered task verification system integrated with a digital reminder framework for PLwD. Specifically, it examines (1) the effectiveness of GPT-4 in generating high-quality follow-up questions that help verify whether tasks were actually completed, (2) the accuracy of an AI-driven response flagging mechanism in identifying tasks requiring caregiver intervention, and (3) the role of caregiver feedback in refining system adaptability.

Methods: A theoretical AI-powered task verification pipeline was designed to enhance digital reminders by generating tailored follow-up questions, analyzing responses, and categorizing concerns. Each follow-up question corresponded to a specific reminder sent through the digital system, aiming to assess whether the task was genuinely completed. To test its feasibility, a simulated AI-powered pipeline was implemented using an anonymized dataset of 64 reminders. GPT-4 generated follow-up questions with and without additional contextual information about PLwD routines. A response classification system flagged task completion as High, Medium, or Low concern, based on response clarity and task urgency. Simulated caregiver feedback was incorporated to refine question quality and improve system adaptability over time.

Results: Contextual information and caregiver feedback significantly improved the clarity, specificity, and relevance of AI-generated follow-up questions. The response flagging mechanism demonstrated high accuracy, particularly for critical tasks such as safety-related reminders. However, subjective or non-urgent tasks posed classification challenges. Caregiver input iteratively enhanced system performance, ensuring a balance between automation and human oversight.

Conclusions: This study demonstrates the feasibility of integrating generative AI into dementia care by enhancing task verification and decision support. Context-aware AI-generated prompts, combined with caregiver feedback, improve task verification accuracy, reduce caregiver stress, and strengthen PLwD support. Future research should focus on real-world validation and scalability to optimize caregiver workload reduction and AI-driven dementia care interventions.

Keywords: Dementia; Caregiving; Generative AI; Task verification; Artificial intelligence; Digital health; Reminder systems; Caregiver support; Aging-in-place; Assistive technology

Introduction

Home-based care allows people living with dementia (PLwD) to age in place and remain in their homes for as long as possible, but this often relies heavily on the commitment of informal caregivers, such as family members and friends [1,2]. These caregivers of PLwD experience significantly higher levels of distress and burnout compared to caregivers of individuals with other chronic diseases [3]. This stress poses a serious challenge to the long-term sustainability of home-based care, which is a critical goal for many families [4–6]. A large part of the caregiver's responsibilities involves supporting the PLwD to complete daily tasks, since cognitive impairments such as memory loss and impaired judgment make it difficult for PLwD to adhere to daily tasks [2].

Digital reminder systems are designed to support PLwD by helping them remember and complete daily tasks such as brushing teeth, meals, appointments, and chores, thereby reducing the caregiver's responsibility [7]. However, their effectiveness depends on verifying whether tasks are actually completed, a challenge that remains unresolved in most digital systems [8–11]. Insights from two Engagement of People with Lived Experience of Dementia (EPLED) members who contributed to this study emphasized that even if a PLwD views a reminder or indicates task completion, caregivers often have no way to confirm this without manually monitoring the task. This lack of verification undermines the system's purpose, potentially increasing workload and stress instead of alleviating it. Current solutions, such as wearable sensors, pose challenges related to privacy and affordability. Generative AI offers a novel, non-intrusive alternative by analyzing responses to follow-up questions, providing real-time caregiver support. This feasibility study investigates the use of generative artificial intelligence (AI) to enhance these systems by verifying task completion and serving as a decision-support tool, helping caregivers determine when intervention is needed to assist the PLwD.

Study Rationale

Previous research into digital reminder systems for PLwD underscores the importance of reliable methods for verifying task completion. Both researchers and caregivers have noted challenges with the accuracy of self-reported task completion—where PLwD indicate task completion through the digital system—and the added workload of caregivers manually verifying adherence [9,11–13]. Existing tools such as wearable devices, motion sensors, and environmental monitors, often function as activity recognition tools to determine whether tasks are completed [14–21]. However, these systems face challenges related to accuracy, affordability, privacy concerns, and user acceptance due to their complexity or intrusiveness. Additionally, many of these methods are incompatible with digital reminder systems, limiting their integration [20–23]. This highlights the need for a non-intrusive, adaptable alternative that works within existing digital reminder frameworks to verify task completion.

To address these limitations, generative AI offers a promising alternative for task verification within digital reminder systems. Unlike sensor-based approaches, generative AI leverages natural language understanding to provide personalized, privacy-preserving solutions [24–30]. Few-shot prompting allows these systems to dynamically incorporate caregiver feedback and tailor outputs to individual needs, enhancing user experience and accessibility [31].

One unexplored potential is using generative AI to generate tailored questions for PLwD to confirm task completion—offering a practical, non-intrusive alternative to sensor-based methods [25,32]. Additionally, generative AI can analyze response patterns to identify situations requiring caregiver intervention,

distinguishing between high-priority tasks (e.g., missed medical appointments) and lower-priority ones (e.g., watering plants). This approach could reduce caregiver stress, improve task adherence, and enhance care management[33,34]. This study focuses on early-stage dementia, where PLwD can still engage with prompts and provide responses [9,35,36]. As dementia progresses and cognitive engagement declines, the system's applicability diminishes, emphasizing its design for early-stage support.

To address these challenges, this study explores a generative AI-driven approach that seamlessly integrates with existing digital reminder systems. The system is designed to:

- Generate tailored follow-up questions
- Flag concerning responses
- Support caregivers in prioritizing interventions

Unlike fully automated solutions, this system maintains a balance between caregiver oversight and AI-driven automation, ensuring adaptability to individual needs. While the system components were developed as modular Python scripts and executed manually in this initial phase, this approach allowed for iterative refinement based on user feedback. Using anonymized historical datasets as a proof of concept, this study assessed the system's ability to:

1. Generate high-quality follow-up questions
2. Identify concerning response patterns
3. Enhance task adherence while reducing caregiver responsibility

Objectives

This feasibility study aims to develop and evaluate an AI-powered task verification system within a digital reminder framework. The study is guided by the following research questions:

1. Can we leverage OpenAI's Generative Pre-trained Transformer 4 (GPT-4) to generate high-quality follow-up questions that align with the cognitive and communication preferences of PLwD, and what factors contribute to improving the quality of these questions (e.g., added context, personalization, or reminder quality)?
2. Can the system effectively flag concerning responses to follow-up questions, helping caregivers identify when tasks may require intervention or professional assistance?
3. Can the system be guided to improve its performance through caregiver feedback?

Methods

Co-Designing with EPLED Members

The study design was centered around co-design with two EPLED members, whose feedback, provided during collaborative sessions, played a crucial role in ensuring the system addressed real-world caregiving needs [37]. Both EPLED members were caregivers for family members with dementia, bringing firsthand experience that deeply informed the design process.

They highlighted the importance of minimizing cognitive overload for PLwD, recommending observation-based prompts (e.g., “Is your breakfast plate empty?”) over memory-reliant ones (e.g., “What did you eat for breakfast?”). They also emphasized tailoring prompts to individual preferences and routines, such as using personalized questions like “Did you drink your orange juice?” instead of generalized queries like “Did you drink something?”.

To further enhance system personalization, they suggested conducting initial interviews with caregivers and PLwD to capture unique routines, preferences, and communication styles. For instance, caregivers could highlight sensitivities, such as avoiding time-focused prompts for tasks where PLwD might fixate on duration. Instead of “Brush your teeth for 2 minutes,” a less stress-inducing prompt like “Are your teeth clean?” would be preferred.

Additionally, EPLED members proposed features such as proactive reminders to enhance adherence, allowing the system to resend follow-up questions if the initial reminder is not acknowledged. These insights were incorporated into the system design explained below, ensuring the system could adapt to various caregiving contexts while maintaining accessibility, privacy, and ease of use.

Their feedback also played a crucial role in shaping how follow-up question quality was evaluated. The emphasis on minimizing cognitive overload directly informed the evaluation criteria, ensuring that AI-generated follow-up questions avoided memory-dependent phrasing, encouraged observation-based verification, and remained concise and easy to understand. For instance, instead of asking “Did you brush your teeth?”, which requires memory recall, a more effective prompt informed by EPLED guidance would be “Are your teeth clean?”, allowing for a verifiable response. Similarly, a vague medication reminder like “Did you take your medication?” was improved to “How many pills are left in your morning pillbox?”, encouraging a response that directly confirms task completion.

These insights were embedded into the evaluation criteria (Table 1), outlined in the Evaluation Checklist for Follow-up Question Quality section, ensuring follow-up questions aligned with real-world caregiving needs. The following section details how GPT-4 was prompted to generate these tailored questions, incorporating EPLED feedback into the AI system’s design.

Theoretical Pipeline Overview

This section outlines the ideal theoretical pipeline for integrating an AI-powered follow-up question generation and response analysis system into a digital reminder system for PLwD. Designed based on the guidelines of EPLED members, this pipeline focuses on empowering caregivers to review, refine, and customize follow-up questions while maintaining trust, usability, and oversight. The feasibility study evaluates how well this approach functions in practice.

The pipeline begins with the setup of a digital reminder, followed by the automated generation of follow-up questions tailored to the specific task and context. These follow-up questions are then reviewed and refined by caregivers to ensure they are understandable and accessible for PLwD. Once finalized, the system delivers these questions after the PLwD acknowledges a reminder. The PLwD’s response is then analyzed, and the system categorizes it into different levels of concern (e.g., high, medium, or low) to help caregivers determine whether intervention is needed.

A visual representation of this pipeline can be seen in Figure 1, illustrating the key components and flow of information. Additionally, Figure 2 provides a sample of how a reminder progresses through the system,

demonstrating the step-by-step interactions between the digital reminder, AI-generated follow-up questions, and caregiver/PLwD involvement.

By structuring this process, the pipeline aims to balance automation with human oversight, ensuring that caregivers remain in control while reducing their manual efforts. The actual tested prototype, which evaluates this theoretical framework in a real-world caregiving scenario, is discussed in Prototyping and Verification of the System section.

Context Collection through Interviews

Before PLwD and caregivers begin using the reminder system, they will participate in interviews or complete questionnaires to provide essential context and information. These interviews help understand the needs, communication styles, preferences, and daily routines of the PLwD. Key information collected includes identifying critical tasks, relationships between PLwD and caregivers, and any personalized tasks. This context is integrated into the prompts provided to GPT-4 to ensure tailored prompts and follow-up questions are relevant and aligned with each individual's unique caregiving situation.

Metadata Creation for Reminders

Caregivers create reminders through the digital reminder system, which are retained and processed by the task verification pipeline along with critical metadata. This metadata includes:

- Reminder type (e.g., mealtime, hygiene, appointments)
- Reminder content (e.g., "Take morning pills" or "Doctor's appointment at 2 PM")
- Priority level (e.g., high or low)

Follow-Up Question Generation with GPT-4

Once the metadata and contextual information are established, GPT-4 uses a few-shot prompting technique to generate follow-up questions tailored to each task. The prompts and few-shot examples were co-designed with EPLED members to encourage the generation of questions that align with the cognitive and communication abilities of PLwD, such as being concise and supportive. The prompt with examples is as follows:

"The following are examples of reminders and their associated follow-up questions:

- Reminder: Brush teeth after dinner.
 - Follow-up question: Is your toothbrush wet?
- Reminder: Drink a glass of water before starting your day.
 - Follow-up question: Is your water glass empty?
- Reminder: Wash your hands before lunch.
 - Follow-up question: Are your hands damp from washing?
- Reminder: Feed the cat in the morning.

- o Follow-up question: Is the cat's food bowl full?

Now generate a follow-up question for the new reminder. The follow-up question should:

- Be actionable and easy to understand.
- Use context clues from the reminder without making assumptions or fabricating information.
- Avoid relying on the person's memory.
- Be kind, guiding, and non-demanding in tone.
- Use clear, short, and concise language.
- Refer back to the reminder directly.
- Ensure the question aligns with the reminder and allows for immediate confirmation of the task."

Additionally, the system incorporates personalized context collected from the interviews into the prompts. For instance, caregivers can provide context such as "Sam is our family dog, not a person", which is incorporated into the prompt sent to GPT-4. This ensures the system generates accurate questions for ambiguous reminders, such as "Go on a walk with Sam."

Caregiver Approval and Feedback Loop

Once a follow-up question is generated, it is sent to the caregiver who can choose to approve the question as-is, modify it to better align with the PLwD's needs, or create entirely new questions. Modified or re-written questions are then incorporated into the few-shot prompting dataset, allowing the system to learn and improve over time. This iterative feedback loop ensures that caregivers can tailor the system as necessary to meet the PLwD's needs. Caregiver modifications would be automatically recorded by the system, enabling ongoing improvements without the need for repeated manual input.

Delivering and Responding to Follow-Up Questions

Once the follow-up question is finalized, it is delivered to the PLwD after they acknowledge the corresponding reminder. They can respond to the question either by typing a free-form answer, speaking their response using voice input, or selecting from multiple-choice options generated by GPT-4. The prompt to generate these options is as follows:

"Generate exactly 3 short, clear, and distinct responses for the following question. Here is an example:

Question: Can you check if the pillbox is empty?

Responses: Yes, it is empty. No, it is not empty. I don't want to check."

This multi-response mechanism is designed to accommodate varying communication abilities, promoting natural and accessible engagement.

Response Analysis and Flagging Levels of Concern

The system then analyzes the PLwD's response using GPT-4, taking into account both the response itself and the reminder information (e.g., reminder type, priority, and description) to categorize the level of concern as high, medium, or low. The prompt used is as follows:

"You are analyzing a response from a person living with dementia to a follow-up question. Consider the context of the reminder and the response to determine the level of concern. Different reminder types (e.g., Medication, Social Engagement) may require varying levels of sensitivity. Your task is to evaluate the level of

concern in the response and categorize it as follows:

- High: Immediate caregiver attention is needed.
- Medium: Follow-up is required, but it is not urgent.
- Low: No immediate concern or action is needed.

Context:

- Reminder Name: [Insert reminder name here]
- Reminder Type: [Insert reminder type here]
- Follow-Up Question: [Insert follow-up question here]
- Response: [Insert response here]

Based on this context, provide a single word categorization: High, Medium, or Low.”

Instead of following a predefined strict threshold, GPT-4 determines the concern level based on its understanding of the provided context. It is given general guidance that high concern means immediate caregiver intervention is needed (e.g., if a critical task like medication is missed), medium concern indicates that follow-up is required but not urgent (e.g., if a non-time-sensitive task is forgotten), and low concern suggests no immediate intervention is necessary, though caregivers can still review the response for transparency and decide whether to act.

It is important to note that these concern levels are generated based on GPT-4’s interpretation of the inputted context rather than rigid rules, and they are meant to support caregiver decision-making rather than serve as definitive alerts for emergency situations.

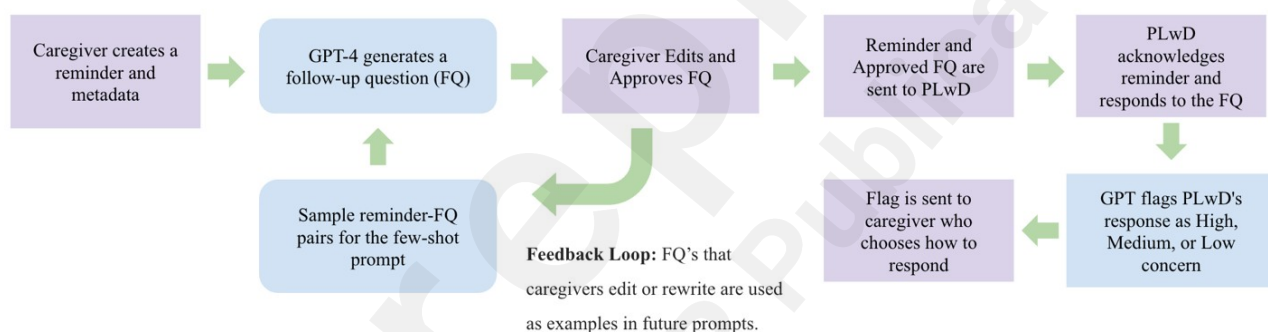


Figure 1 - Theoretical pipeline for the system

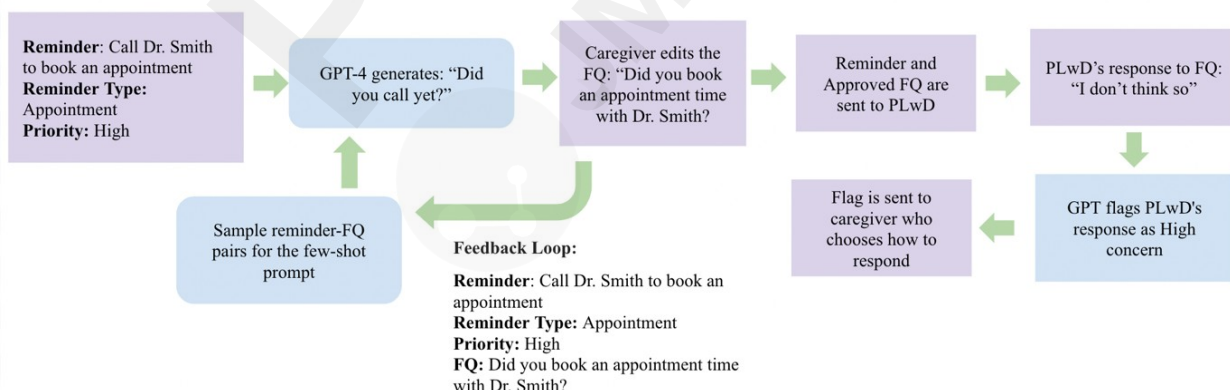


Figure 2 – Sample workflow of a reminder through the theoretical pipeline

Prototyping and Verification of the System

This feasibility study evaluates a prototyped version of the previously described theoretical pipeline by using

Python scripts with the GPT-4 API to replicate the automated flow through the digital reminder system and its interactions with the caregiver, as illustrated in Figure 3. This feasibility study utilized anonymized reminder metadata from a prior usability study involving two dyads of caregivers and PLwD, where caregivers used a digital reminder system to send reminders over two-week periods in the home. Dataset 1 had an average of 29 characters per reminder with 27 reminders, while Dataset 2 had an average of 11.5 characters per reminder with 37 reminders, both reflecting real-world caregiving scenarios, including appointments, personal hygiene, daily routines, social engagement, meals, and safety-related activities. To mitigate ethical risks, such as providing incorrect health-related advice, medication-related reminders were excluded. The removal of personal identifiers ensured participant privacy and adherence to ethical standards.

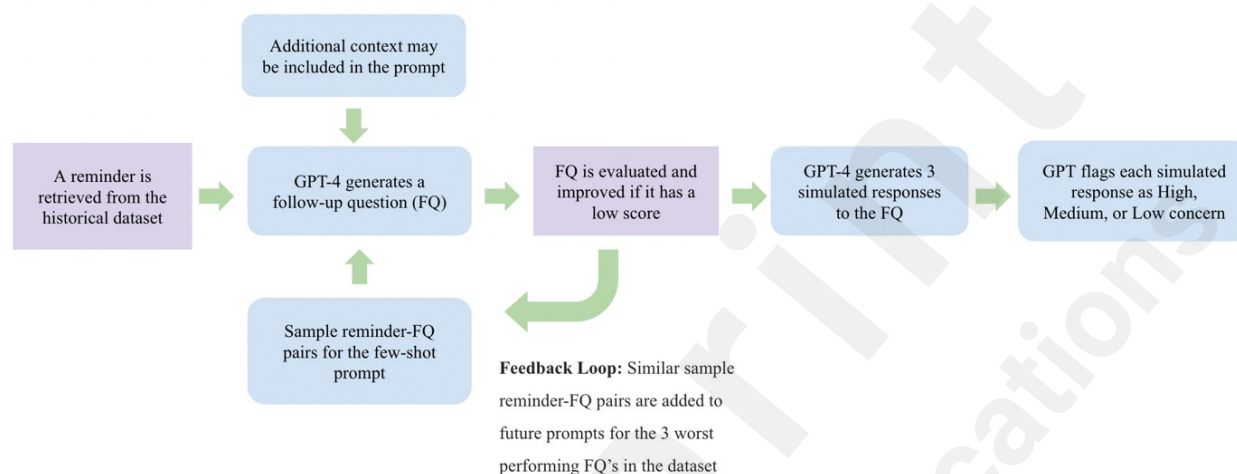


Figure 3 - Actual prototyped pipeline to simulate the theoretical pipeline

The historical datasets served as a foundation for simulating interactions between caregivers, PLwD, and the task verification system. Although caregivers and PLwD did not directly interact with the system, the datasets allowed for an evaluation of the system's ability to generate meaningful follow-up questions and verify task adherence. Contextual details known about the dyads' routines, relationships, and preferences were incorporated into the prompts to simulate context from interviews. This approach simulated realistic caregiving scenarios and tested the system's ability to generate relevant and personalized prompts.

To answer Research Question 1, the quality of follow-up questions was evaluated by comparing instances where context was provided with those where it was not. Additionally, the impact of varying levels of reminder detail and clarity was examined to determine if the style of writing influenced the quality of the generated follow-up questions. The specific method used to evaluate the quality of the follow-up questions is described in the Evaluation Checklist for Follow-up Question Quality section.

To address Research Question 2, which focused on the identification and flagging of potentially concerning responses, three types of responses were generated by GPT-4 for each follow-up question: one indicating task completion, one indicating non-completion, and one offering an ambiguous answer.

To assess flagging accuracy, responses were categorized into High, Medium, or Low concern levels, following a structured decision process informed by EPLED caregiver insights. As illustrated in Figure 4, responses were first evaluated for whether they were concerning, meaning they were unclear, unusual, or indicated non-completion. If a response was not concerning, it was classified as Low Concern, requiring no immediate caregiver action. If the response was concerning, the next step was to assess whether the task was time-

sensitive or critical.

- **High Concern:** Requires immediate caregiver intervention when a time-sensitive or safety-critical task is not completed. These tasks typically involve:
 - Time-sensitive commitments (e.g., virtual meetings at a set time).
 - Safety risks (e.g., ensuring the stove is off, locking doors).
 - Essential physical well-being (e.g., eating meals, maintaining hygiene).
- **Medium Concern:** Follow-up is recommended but not urgent, meaning the task is not time-sensitive but the response was unclear or indicated difficulty. These tasks often:
 - Affect daily health or routine but do not require immediate action (e.g., missing a doctor's appointment that can be rescheduled).
 - Have flexible timing (e.g., household chores, exercise).
- **Low Concern:** No immediate action is required, though caregivers can review the response. These tasks tend to:
 - Affect non-essential activities (e.g., forgetting to water plants or skipping a hobby).
 - Be subjective in urgency, depending on individual preferences (e.g., missing a social engagement).

Since task urgency varies by caregiver and context, future iterations of the system could enable customizable flagging thresholds, allowing caregivers to adjust concern levels based on their own priorities and the specific needs of the PLwD. For example, while some caregivers may classify missing a meal as High Concern due to health risks, others may consider it Medium Concern depending on the PLwD's dietary habits and overall condition. By integrating contextual flexibility, the system could better adapt to personalized caregiving preferences, ensuring that AI-generated alerts remain relevant and actionable.

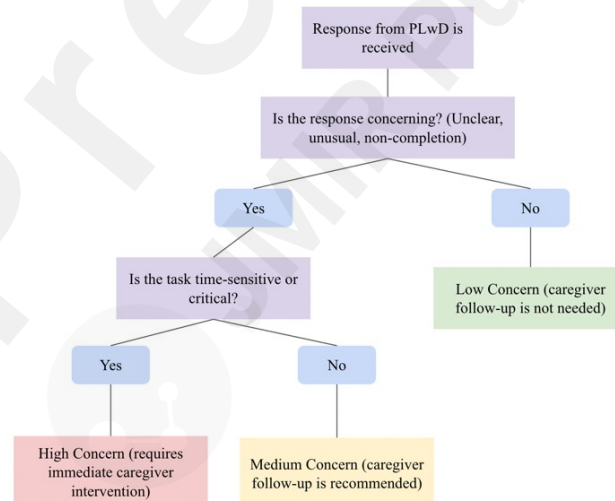


Figure 4 - Decision tree for response classification

To address Research Question 3, a simulated feedback loop with caregiver input was used to refine follow-up questions. After generating follow-up questions, they were evaluated, and the three lowest-scoring pairs were identified for improvement. Simulating caregiver input, these questions were refined and added to the few-shot prompting set. The updated dataset was then used to regenerate and re-evaluate the questions, comparing them to the originals. This iterative process demonstrated how caregiver input could enhance the system's follow-up questions while balancing reduced caregiver responsibility with maintaining control, transparency, and the option to override the system.

Evaluation Checklist for Follow-up Question Quality

The evaluation of AI-generated follow-up questions was conducted using a checklist (Table 1) informed by EPLED input, emphasizing clarity, contextual relevance, memory independence, and a supportive tone. Instead of rigid quality classifications, the number of criteria met was counted for each question, and this sum is referred to as the "score" in the results section to provide a general understanding of performance. A higher score indicates a better question, while a lower score suggests a weaker question. The best possible score is 12 (meeting all criteria), and the worst is 0 (meeting none). However, not all criteria contribute equally to question effectiveness, so the score serves as a general snapshot rather than an absolute measure of quality.

For example, a follow-up question like *"Did you brush your teeth?"* is weak because it is vague and allows for a simple yes/no answer. In contrast, *"Is your toothbrush wet from brushing?"* is a stronger alternative as it encourages a verifiable response. Similarly, *"Did you finish cooking?"* is unclear regarding actual task completion, whereas *"Is the stove knob in the off position?"* directly verifies that the task was done.

While the score provides a quick overview of question quality, we also break down performance based on individual criteria to gain more specific insights. This allows for a deeper understanding of which aspects were consistently met and where improvements may be needed.

Table 1 - Evaluation Criteria Checklist

Criteria	Questions for Evaluation
Clarity	Is the question easy to understand without ambiguity or overly complex phrasing?
Conciseness	Does the question communicate its purpose in as few words as necessary while maintaining full meaning?
Contextual Relevance	Does the question refer specifically to the task or details mentioned in the reminder?
Avoidance of Assumptions	Does the question avoid introducing or assuming details that were not part of the original reminder?
Memory Independence	Does the question include enough contextual clues (e.g., time, place, or actions) to minimize reliance on the user's memory?
Support for Recall	Does the question avoid requiring the user to recall details without support or prompts?
Reminder Specificity	Could a more detailed or specific reminder make the question clearer or more actionable?
Task Completion Focus	Is the question focused on determining whether the task from the reminder was completed?
Avoidance of Irrelevance	Does the question avoid straying into unrelated or irrelevant topics?
Supportive Tone	Does the question use a supportive and respectful tone?
Encouraging Engagement	Does the tone encourage engagement without being judgmental or condescending?

Results

The results are presented descriptively to highlight trends and guide future research. Findings are structured into distinct subsections, with key takeaways at the end of each.

Follow-Up Question Quality

To address Research Question 1, the quality of AI-generated follow-up questions was evaluated across two datasets—one with contextual information and one without. The results indicate that contextual information significantly improved key aspects of question quality, particularly Reminder Specificity and Contextual Relevance. In Dataset 1, which contained more detailed reminders, Reminder Specificity increased by 33.3% when context was added, while Dataset 2 saw a smaller 23.5% improvement. This suggests that Dataset 1 leveraged richer reminder metadata to generate specific and actionable prompts, whereas Dataset 2, which contained shorter and often vague reminders, showed limited responsiveness to added context.

Although context improved these specific dimensions, its impact on other metrics, such as Avoidance of Irrelevance and Task Completion Focus, was less pronounced. Factors such as conciseness, clarity, and supportive tone remained consistent across all conditions, suggesting that these qualities were less dependent on the presence of additional context. These findings highlight the importance of well-structured reminders in optimizing AI-generated follow-up questions, as context alone cannot fully compensate for vague or minimal reminder inputs.

Table 2 - Overview of the number of questions meeting checklist criteria for both datasets, comparing results with and without context

Checklist Question	Dataset 1 Without Context	Dataset 1 With Context	Dataset 2 Without Context	Dataset 2 with Context
Reminder Specificity	56%	74%	46%	57%
Avoidance of Assumptions	74%	74%	70%	86%
Support for Recall	67%	74%	57%	59%
Avoidance of Irrelevance	100%	93%	89%	97%
Conciseness	100%	100%	100%	100%
Memory Independence	57%	75%	62%	70%
Contextual Relevance	78%	89%	70%	78%
Clarity Without Context	85%	85%	81%	92%
Supportive Tone	100%	100%	100%	100%
Encouraging Engagement	100%	100%	100%	100%
Clarity	100%	100%	100%	100%
Task Completion Focus	74%	78%	65%	68%

Impact of Reminder Quality on Follow-up Question Quality

As shown in Figures 5 and 6, reminder quality played a crucial role in determining follow-up question scores, even after additional context was added. Dataset 1, which contained longer and more detailed reminders, exhibited substantial improvements in follow-up question quality when context was introduced. In contrast, Dataset 2, which contained short and often ambiguous reminders (e.g., one- or two-word reminders such as "Groceries" or "Meal Prep"), exhibited only minor improvements. This suggests that higher-quality reminders benefit more from additional context, while overly simple reminders provide limited information for the system to build upon, restricting the effectiveness of contextual enhancements.

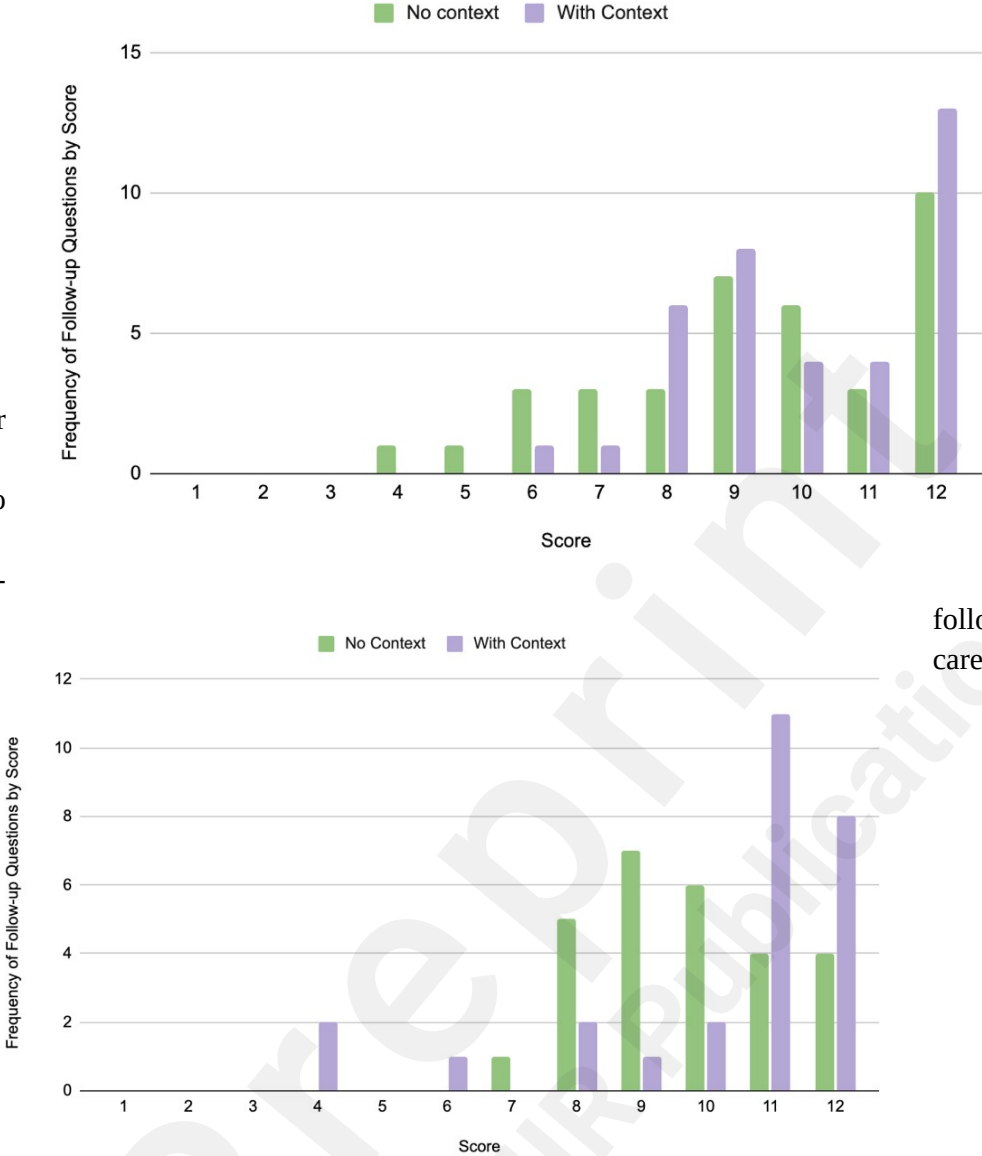
Interestingly, Dataset 1 exhibited a wider range of scores after context was added, whereas Dataset 2 showed greater variability before context was incorporated. This suggests that for reminders with some level of

context

refine
questions,
reminders
vague

results
added
findings
need for
crafted
inputs to

of AI-



specificity,
can
meaningfully
follow-up
whereas
that are too
may yield
unpredictable
regardless of
context. These
reinforce the
carefully
reminder
maximize the
effectiveness
generated
follow-up questions in
caregiving scenarios.

Figure 6 - Evaluation Checklist Score for Dataset 2, with and without added context

Impact of Caregiver Feedback

To address Research Question 3, caregiver input was integrated into the system through few-shot learning, allowing the model to refine its question generation based on prior caregiver modifications. This led to significant improvements in follow-up question quality, with an average score increase of 4 points across both datasets, as shown in Table 3. In Dataset 1, for example, the original follow-up question "Is your car ready?" was revised to "Is [redacted] with you now?", increasing its score from 4 to 11. Similarly, in Dataset 2, the vague reminder "Brunch" initially produced the question "Did you set the table for brunch?", which scored 7. After incorporating caregiver feedback, the revised question "Is the brunch table set and ready?" received a score of 12.

These improvements demonstrate that caregiver feedback plays a vital role in refining AI-generated follow-up questions. In Dataset 1, which contained more detailed reminders, caregiver modifications primarily enhanced emotional engagement and task specificity. In Dataset 2, caregiver feedback helped bridge gaps created by sparse reminders, ensuring that follow-up questions were more actionable. For instance, the reminder "Laundry" initially generated the vague question "Is the laundry basket empty?". This was later refined into "Is the laundry folded and put away?" or "Is the washing machine running with your clothes inside?", recognizing that an empty laundry basket does not necessarily indicate task completion. These findings underscore the importance of caregiver input in improving task focus, relevance, and emotional engagement in AI-generated prompts.

Table 3 - Summary of the change in scores for poor-performing follow-up questions after simulated caregiver feedback

Reminder	Original Follow-Up Question	Original Score	Revised Follow-Up Question (Few-Shot Learning Applied)	Revised Score	Change
Dataset 1					
[Redacted] coming tomorrow	Will you set out fresh towels for [Redacted]'s visit?	4	Are you excited for [Redacted]'s visit tomorrow	8	+4
Pick [Redacted] up from the GO station	Is your car ready for the drive to the GO station?	4	Is [Redacted] with you now?	11	+7
Home and Garden show	Is the TV turned on to the Home and Garden channel?	6	Are you watching the Home and Garden Show on TV?	8	+2
Dataset 2					
Leave Cambridge	Are you ready to head out for your activity?	6	Have you packed your bag to leave Cambridge?	11	+5
Brunch	Did you set the table for brunch?	7	Did you enjoy your brunch?	11	+4
Laundry	Is the laundry basket empty?	8	Are the clothes washed and put away?	10	+2

Evaluation of Flagging Performance

The AI system's ability to classify responses into High, Medium, or Low concern levels was evaluated across both datasets, with and without additional context, as shown in Table 4. The overall flagging accuracy was high, but certain misclassifications revealed areas for improvement, particularly in handling subjective or ambiguous responses, shown in Table 5.

In Dataset 1, errors decreased from two to one when context was added. For instance, without context, the reminder "Home and Garden show" was incorrectly flagged as Medium concern instead of Low concern. When context was added, clarifying that the reminder referred to a television program rather than an in-person event, the classification was corrected. Similarly, the reminder "Make an appointment for pedicure" was initially flagged as Low concern, despite its time-sensitive nature. Context had limited impact in resolving this error, indicating challenges in assessing urgency for personal care tasks.

In Dataset 2, misclassifications dropped from three to two with the addition of context. The reminder "Brunch" was initially flagged as Low concern when the response was "I wasn't hungry." While skipping a meal may not always be concerning, in dementia care, caregivers may still want to be informed. Another misclassification occurred with the reminder "Water flowers," which was mistakenly flagged as High priority instead of Medium, overestimating the task's importance. Although some errors were corrected with context, others, such as "Quiet time on balcony", were misclassified when context was added, demonstrating the challenge of categorizing subjective tasks.

EPLED members noted that a response such as "No, I'm not settled yet." could indicate High or Medium concern depending on caregiver experience, even though, under the current flagging criteria, it would more likely be categorized as Low concern. This highlights the difficulty of balancing false positives and false negatives, particularly for subjective or routine tasks.

Despite these challenges, high-risk tasks—such as safety-critical reminders—were consistently flagged correctly across all conditions. These results suggest that while context enhances the system's performance, its impact varies depending on task type, and further refinements are needed to address ambiguities in subjective or routine responses effectively.

Table 4 - Overview of correct and incorrect flagging for Datasets 1 and 2, comparing performance with and without context

Dataset	Number of Responses	Correct Flags	Incorrect Flags
Dataset 1 (with context)	81	80	1
Dataset 1 (without context)	81	79	2
Dataset 2 (with context)	111	109	2
Dataset 2 (without context)	111	108	3

Table 5 - Description of incorrectly flagged responses

Dataset	Reminder	Follow-Up Question	Response	Flagged Concern Level	Correct Concern Level
1 (Without Context)	Home and Garden show	Did you check the schedule for the Home and Garden show?	I forgot to check it.	Medium	Low
1 (Without Context)	Make appointment for pedicure	Have you booked your pedicure appointment?	No, I haven't booked it yet.	Low	Medium

1	(With Context)	Call radiology to get on cancellation list	Have you dialed the number for radiology to ask about the cancellation list?	No, I haven't dialed yet.	Medium	High
2	(Without Context)	Brunch	Did you finish your brunch plate?	I wasn't hungry.	Low	Medium
2	(Without Context)	Water flowers	Are the soil surfaces in the flower pots damp?	No, they are dry.	High	Medium
2	(With Context)	Quiet time on balcony	Are you comfortably settled on the balcony?	No, I'm not settled yet.	Medium	Low

Discussion

This feasibility study demonstrated the potential of AI-powered task verification systems to support caregivers and PLWD by addressing key research questions and identifying critical challenges.

For the first research question, the system successfully generated high-quality follow-up questions that aligned with the cognitive and communication preferences of PLWD. Contextual information and caregiver feedback played a crucial role in improving question quality. Context-enriched prompts helped refine vague reminders, such as transforming “Pick up [redacted]” from an assumptive query like “Is your car ready?” into a more relevant and actionable question such as “Is [redacted] with you now?”. Additionally, GPT-4's multilingual capabilities allow the system to be adapted for diverse caregiving settings, making it accessible in languages such as English, Spanish, French, and Chinese [30].

The second research question focused on the system's ability to flag responses as High, Medium, or Low concern. The results indicate that the system achieved high accuracy in classifying responses, particularly for high-priority tasks such as safety-critical reminders. Contextual information improved classification in nuanced cases, such as distinguishing between a physical event and a television program in the “Home and Garden show” reminder, which was reclassified from Medium to Low concern when additional context clarified its significance. However, subjective or routine tasks remained challenging, even when context was provided. For example, the reminder “Quiet time on balcony” was misclassified, illustrating the difficulty of assessing personal or non-urgent activities. These findings underscore the feasibility of using GPT-4 for response flagging while also highlighting the need for further refinements in handling ambiguous or less urgent tasks.

To address the third research question, a simulated feedback loop demonstrated how caregiver input could iteratively enhance follow-up question quality and improve system adaptability. Through this process, simulated caregiver input shifted the focus from preparatory steps to direct task outcomes. For instance, in response to the reminder “[Redacted] is coming today,” the system originally generated “Are towels set out?”, which was later refined into a more engaging and meaningful question such as “Are you looking forward to [Redacted]’s visit today?”. This highlights an important function of generative AI—GPT-4 can be guided to create emotionally engaging prompts that reinforce awareness of upcoming events rather than assuming that every reminder necessitates a completed task. By leveraging context and few-shot learning, the system avoided unnecessary assumptions while still prompting relevant responses.

However, this iterative caregiver involvement raises a fundamental question: Does AI task verification truly alleviate caregiver stress, or does it shift responsibility toward system monitoring and refinement? While AI automation reduces the need for manual task verification, it also requires ongoing training, adjustments, and oversight by caregivers to ensure effectiveness. In the short term, this adaptation process could increase caregiver workload, creating a trade-off between automation and manual intervention. Further research is

needed to determine whether the long-term benefits outweigh the short-term effort caregivers must invest in training the system and learning how to use it.

In summary, this feasibility study highlights the promise of integrating generative AI into task verification systems for dementia care. Contextual information and caregiver feedback significantly enhanced the quality of follow-up questions, making them more actionable and relevant to the caregiving context. Additionally, the system demonstrated strong performance in accurately flagging high-priority tasks, particularly those involving safety. Nevertheless, challenges with subjective or routine tasks, as well as the balance between caregiver workload and automation, underscore the need for ongoing refinement and further research.

Insights from Co-Design

Collaboration with EPLED members played a pivotal role in designing a system that addressed real-world caregiving needs. By recommending observation-based prompts instead of memory-reliant ones, EPLED members helped address the cognitive stress for PLWD. Personalization was emphasized throughout, with tailored prompts shown to improve task adherence and engagement. Features like proactive reminders and caregiver visibility into flagged responses reinforced the importance of transparency and caregiver trust.

However, members also highlighted that the datasets used to train the system may not represent the diverse backgrounds and socioeconomic contexts of all caregiving dyads. For example, one member noted that "brunch" was not relevant in their caregiving experience, as the PLWD they cared for would not even understand the term. This raised important questions about whether GPT could appropriately handle class-specific or culture-specific terms.

Additionally, EPLED members noted that PLWD's conditions worsen at varying rates, and their needs can change rapidly. This underscored a critical challenge: determining when the tool is no longer sufficient to support PLWD as their condition progresses. These insights highlighted the need for ongoing adaptability and robust mechanisms to evaluate the tool's continued effectiveness over time.

Limitations

This study had several limitations. The small dataset size and reliance on simulated responses limit generalizability and preclude statistical validation. Variability in reminder quality influenced outcomes, as vague reminders generated weaker follow-up questions, even with added context. Additionally, the system was not tested in real-world caregiving settings, where dynamic interactions with PLWD may introduce new challenges.

The system was designed for digital reminder verification and did not include medication-related tasks. Future research should explore its integration into clinical care, particularly for medication adherence tracking and mobile health platforms. While context improved follow-up question quality, its impact on response flagging was inconsistent, particularly for subjective tasks like "Quiet time on balcony". Further studies should examine how different types of contextual inputs affect AI classification accuracy.

GPT-4 struggled with ambiguous responses, such as "I don't remember," potentially misclassifying them as Low, Medium, or High concern. Future refinements should incorporate clarification prompts, escalation mechanisms, and hybrid AI-human oversight to enhance reliability. Additionally, GPT-4's training data may

contain biases, potentially leading to inaccurate or misleading responses. Caregivers must validate AI-generated content to prevent incorrect task verification, emphasizing the need for manual oversight and correction mechanisms.

Finally, evaluation methods relied on subjective interpretation of task urgency, limiting reproducibility. Despite these constraints, this study demonstrates feasibility and highlights key areas for future real-world testing and refinement.

Future Direction

Future research will focus on validating the system in real-world caregiving environments using larger, more diverse datasets and longitudinal studies to assess its adaptability to evolving caregiving needs and its long-term impact on caregiver stress and task adherence. Expanding user involvement beyond EPLED members will provide broader insights into caregiver preferences, enabling further system refinement.

A key area for exploration is scalability, particularly in training the system on a larger caregiver dataset to improve personalization settings and ensure context-enriched prompts function effectively across different caregiving styles and reminder formats. Research should also investigate GPT's capacity for handling increasing personalization and guide strategies for generating clear, targeted prompts without overwhelming the system with excessive contextual information.

Understanding caregiver preferences for reminders, sensitivity to false alarms, and trust in AI-driven assistance will be essential for optimizing the balance between automation and caregiver oversight. Future studies should explore acceptable levels of AI intervention, ensuring that caregivers remain in control while minimizing unnecessary stress or system over-reliance. Interviews with caregivers and PLwD will help refine expectations for how AI can best support caregiving routines while maintaining a user-centered design.

Beyond home-based care, this system could be explored for potential clinical integration, particularly in telehealth services and electronic health records (EHRs). Investigating whether AI-powered task verification could complement clinical workflows, medication adherence tracking, or remote patient monitoring presents an important avenue for future development. By addressing these considerations, future research can ensure that the system remains scalable, adaptable, and beneficial across diverse caregiving and healthcare settings.

Conclusion

This feasibility study demonstrates the potential of AI-powered task verification systems to address critical challenges in dementia care. By generating personalized follow-up questions, accurately flagging concerning responses, and supporting caregiver decision-making, the system provides a foundation for improving task adherence and reducing caregiver stress. While further research is needed to validate and scale the system, these findings lay the groundwork for integrating AI into home-based dementia care, enhancing the well-being of both caregivers and PLwD.

Acknowledgements

This study was supported by AGE-WELL NCE and the University of Toronto. The authors thank KB and DB for their valuable contributions as lived experience experts, helping to shape the study design, evaluation criteria, and review. The authors also acknowledge BY for editorial assistance and AM for supervision and guidance throughout the research process.

Conflicts of Interest

None declared.

Abbreviations

AI: Artificial Intelligence

PLwD: People Living with Dementia

GPT-4: Generative Pre-trained Transformer 4

EPLED: Engagement of People with Lived Experience of Dementia

GO: GO Transit (regional public transit service)

FQ: Follow-up Question

References

1. Curnow E, Rush R, Maciver D, Górska S, Forsyth K. Exploring the needs of people with dementia living at home reported by people with dementia and informal caregivers: a systematic review and Meta-analysis. *Aging Ment Health*. Routledge; 2021. p. 397–407. PMID:31791140
2. Frias CE, Cabrera E, Zabalegui A. Informal caregivers' roles in dementia: The impact on their quality of life. *Life MDPI AG*; 2020 Nov 1;10(11):1–12. doi: 10.3390/life10110251
3. Armstrong J. NAVIGATING THE PATH FORWARD FOR DEMENTIA IN CANADA REPORT. 2022. Available from: https://alzheimer.ca/sites/default/files/documents/Landmark-Study-Report-1-Path_Alzheimer-Society-Canada_0.pdf [accessed Nov 29, 2024]
4. Zwerling JL, Cohen JA, Verghese J. Dementia and caregiver stress. *Neurodegener Dis Manag*. 2016. p. 69–72. PMID:27027883
5. Channer NS, Hartt M, Biglieri S. Aging-in-place and the spatial distribution of older adult vulnerability in Canada. *Applied Geography Elsevier Ltd*; 2020 Dec 1;125. doi: 10.1016/j.apgeog.2020.102357
6. Wiles JL, Leibing A, Guberman N, Reeve J, Allen RES. The meaning of “aging in place” to older people. *Gerontologist* 2012 Jun;52(3):357–366. PMID:21983126
7. Peres B, Campos PF. A systematic review of reminder and guidance systems for Alzheimer's Disease and Related Dementias patients: context, barriers and facilitators. *Disabil Rehabil Assist Technol Taylor and Francis Ltd*; 2023; doi: 10.1080/17483107.2023.2277821
8. Mettouris C, Yeratziotis A, Fixl L, Parker S, Starosta-Sztuczka J, Clarke V, Rosenauer G, Daxberger W, Koumou S, Panayides M, Zampa G, Theodorou C, Papadopoulos GA. eSticky: An Advanced Remote Reminder System for People with Early Dementia. *SN Comput Sci Springer*; 2023 Jul 1;4(4). doi: 10.1007/s42979-023-01768-3
9. McGoldrick C, Crawford S, Evans JJ. MindMate: A single case experimental design study of a

- reminder system for people with dementia. *Neuropsychol Rehabil* Routledge; 2021;31(1):18–38. PMID:31429370
10. Jönsson KE, Ornstein K, Christensen J, Eriksson J. A reminder system for independence in dementia care: A case study in an assisted living facility. *ACM International Conference Proceeding Series Association for Computing Machinery*; 2019. p. 176–185. doi: 10.1145/3316782.3321530
 11. Sanchez AA, Lai J, Ye B, Mihailidis A. Enhancing communication and autonomy in dementia through technology: Navigating home challenges and memory aid usage. *Gerontechnology International Society for Gerontechnology*; 2024;23(1):1–11. doi: 10.4017/gt.2024.23.1.880.06
 12. Jönsson KE, Ornstein K, Christensen J, Eriksson J. A reminder system for independence in dementia care: A case study in an assisted living facility. *ACM International Conference Proceeding Series Association for Computing Machinery*; 2019. p. 176–185. doi: 10.1145/3316782.3321530
 13. Mettouris C, Yeratziotis A, Fixl L, Parker S, Starosta-Sztuczka J, Clarke V, Rosenauer G, Daxberger W, Koumou S, Panayides M, Zampa G, Theodorou C, Papadopoulos GA. eSticky: An Advanced Remote Reminder System for People with Early Dementia. *SN Comput Sci Springer*; 2023 Jul 1;4(4). doi: 10.1007/s42979-023-01768-3
 14. Shahid ZK, Saguna S, Åhlund C. Unsupervised Forecasting and Anomaly Detection of ADLs in single-resident elderly smart homes. *Proceedings of the ACM Symposium on Applied Computing Association for Computing Machinery*; 2023. p. 607–610. doi: 10.1145/3555776.3577822
 15. Arifoglu D, Bouchachia A. Activity Recognition and Abnormal Behaviour Detection with Recurrent Neural Networks. *Procedia Comput Sci Elsevier B.V.*; 2017. p. 86–93. doi: 10.1016/j.procs.2017.06.121
 16. Sim K, Phua C, Yap G-E, Biswas J, Mokhtari M. Activity recognition using correlated pattern mining for people with dementia. *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society 2011*. p. 7593–7597. doi: 10.1109/IEMBS.2011.6091872
 17. Lobanova V, Bezdetnyy D, Anishchenko L. Human Activity Recognition Based on Radar and Video Surveillance Sensor Fusion. *2023 IEEE Ural-Siberian Conference on Biomedical Engineering, Radioelectronics and Information Technology (USBREIT) 2023*. p. 25–28. doi: 10.1109/USBREIT58508.2023.10158846
 18. Na SH, Kim KJ, Huh EN. Wandering Detection and Activity Recognition for Dementia Patients Using Wireless Sensor Networks. *JOURNAL OF INTERNET TECHNOLOGY* 2012;13(1):115–125. PMID:WOS:000300137800012
 19. Civitarese G. Human Activity Recognition in Smart-Home Environments for Health-Care Applications. *2019 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops) 2019*. p. 1. doi: 10.1109/PERCOMW.2019.8730719
 20. Simonsson S, Casagrande FD, Zouganeli E. Use of Clustering Algorithms for Sensor Placement and Activity Recognition in Smart Homes. *IEEE Access* 2023;11:9415–9430. doi: 10.1109/ACCESS.2023.3239265
 21. Hao J, Bouchard B, Bouzouane A, Gaboury S. Real-Time Activity Prediction and Recognition in Smart Homes by Formal Concept Analysis. *2016 12th International Conference on Intelligent Environments (IE) 2016*. p. 103–110. doi: 10.1109/IE.2016.24
 22. Machot FA, Ranasinghe S, Plattner J, Jnoub N. Human Activity Recognition based on Real Life Scenarios. *2018 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops) 2018*. p. 3–8. doi: 10.1109/PERCOMW.2018.8480138
 23. Fortin-Simard D, Bilodeau JS, Gaboury S, Bouchard B, Bouzouane A. Method of Recognition and Assistance Combining Passive RFID and Electrical Load Analysis That Handles Cognitive Errors. *Int J Distrib Sens Netw* 2015; PMID:WOS:000364665000001
 24. Zaman KT, Hasan WU, Li J, Tao C. Empowering Caregivers of Alzheimer’s Disease and Related Dementias (ADRD) with a GPT-Powered Voice Assistant: Leveraging Peer Insights from Social Media. *Proc IEEE Symp Comput Commun Institute of Electrical and Electronics Engineers Inc.*; 2023. doi: 10.1109/ISCC58397.2023.10218142

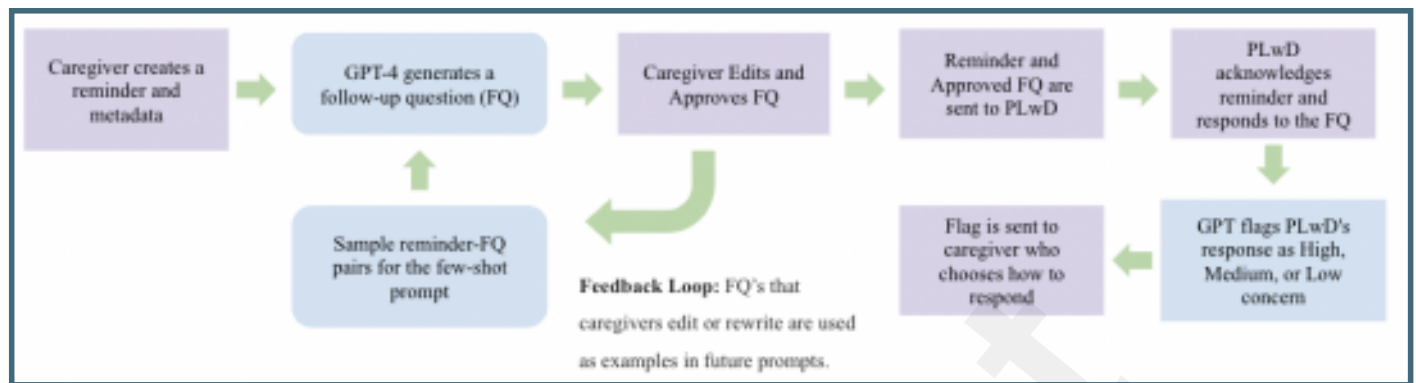
25. Chen X, Fan G, Roberts E, Howell S. A ChatGPT-enabled Intelligent Assistive Technology for Homecare to Support Activities of Daily Living for People with Mild Dementia. *ACM International Conference Proceeding Series Association for Computing Machinery*; 2024. p. 1–7. doi: 10.1145/3678429.3678430
26. Hossain G, Pomare ZS, Prybutok G. ChatGPT: A Companion for Dementia Care. *Digest of Technical Papers - IEEE International Conference on Consumer Electronics Institute of Electrical and Electronics Engineers Inc.*; 2024. doi: 10.1109/ICCE59016.2024.10444253
27. Gilman ES, Kot S, Engineer M, Dixon E. Training Adults with Mild to Moderate Dementia in ChatGPT: Exploring Best Practices. *Association for Computing Machinery (ACM)*; 2024. p. 101–106. doi: 10.1145/3640544.3645230
28. Mao R, Chen G, Zhang X, Guerin F, Cambria E. GPTeval: A Survey on Assessments of ChatGPT and GPT-4. 2023 Aug 23; Available from: <http://arxiv.org/abs/2308.12488>
29. Zhang Z, Rossi RA, Kveton B, Shao Y, Yang D, Zamani H, Dernoncourt F, Barrow J, Yu T, Kim S, Zhang R, Gu J, Derr T, Chen H, Wu J, Chen X, Wang Z, Mitra S, Lipka N, Ahmed N, Wang Y. Personalization of Large Language Models: A Survey. 2024 Oct 29; Available from: <http://arxiv.org/abs/2411.00027>
30. OpenAI, Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman FL, Almeida D, Altenschmidt J, Altman S, Anadkat S, Avila R, Babuschkin I, Balaji S, Balcom V, Baltescu P, Bao H, Bavarian M, Belgum J, Bello I, Berdine J, Bernadett-Shapiro G, Berner C, Bogdonoff L, Boiko O, Boyd M, Brakman A-L, Brockman G, Brooks T, Brundage M, Button K, Cai T, Campbell R, Cann A, Carey B, Carlson C, Carmichael R, Chan B, Chang C, Chantzis F, Chen D, Chen S, Chen R, Chen J, Chen M, Chess B, Cho C, Chu C, Chung HW, Cummings D, Currier J, Dai Y, Decareaux C, Degry T, Deutsch N, Deville D, Dhar A, Dohan D, Dowling S, Dunning S, Ecoffet A, Eleti A, Eloundou T, Farhi D, Fedus L, Felix N, Fishman SP, Forte J, Fulford I, Gao L, Georges E, Gibson C, Goel V, Gogineni T, Goh G, Gontijo-Lopes R, Gordon J, Grafstein M, Gray S, Greene R, Gross J, Gu SS, Guo Y, Hallacy C, Han J, Harris J, He Y, Heaton M, Heidecke J, Hesse C, Hickey A, Hickey W, Hoeschele P, Houghton B, Hsu K, Hu S, Hu X, Huizinga J, Jain S, Jain S, Jang J, Jiang A, Jiang R, Jin H, Jin D, Jomoto S, Jonn B, Jun H, Kaftan T, Kaiser Ł, Kamali A, Kanitscheider I, Keskar NS, Khan T, Kilpatrick L, Kim JW, Kim C, Kim Y, Kirchner JH, Kiros J, Knight M, Kokotajlo D, Kondraciuk Ł, Kondrich A, Konstantinidis A, Kosic K, Krueger G, Kuo V, Lampe M, Lan I, Lee T, Leike J, Leung J, Levy D, Li CM, Lim R, Lin M, Lin S, Litwin M, Lopez T, Lowe R, Lue P, Makanju A, Malfacini K, Manning S, Markov T, Markovski Y, Martin B, Mayer K, Mayne A, McGrew B, McKinney SM, McLeavey C, McMillan P, McNeil J, Medina D, Mehta A, Menick J, Metz L, Mishchenko A, Mishkin P, Monaco V, Morikawa E, Mossing D, Mu T, Murati M, Murk O, Mély D, Nair A, Nakano R, Nayak R, Neelakantan A, Ngo R, Noh H, Ouyang L, O’Keefe C, Pachocki J, Paino A, Palermo J, Pantuliano A, Parascandolo G, Parish J, Parparita E, Passos A, Pavlov M, Peng A, Perelman A, Peres F de AB, Petrov M, Pinto HP de O, Michael, Pokorny, Pokrass M, Pong VH, Powell T, Power A, Power B, Proehl E, Puri R, Radford A, Rae J, Ramesh A, Raymond C, Real F, Rimbach K, Ross C, Rotsted B, Roussez H, Ryder N, Saltarelli M, Sanders T, Santurkar S, Sastry G, Schmidt H, Schnurr D, Schulman J, Selsam D, Sheppard K, Sherbakov T, Shieh J, Shoker S, Shyam P, Sidor S, Sigler E, Simens M, Sitkin J, Slama K, Sohl I, Sokolowsky B, Song Y, Staudacher N, Such FP, Summers N, Sutskever I, Tang J, Tezak N, Thompson MB, Tillet P, Tootoonchian A, Tseng E, Tuggle P, Turley N, Tworek J, Uribe JFC, Vallone A, Vijayvergiya A, Voss C, Wainwright C, Wang JJ, Wang A, Wang B, Ward J, Wei J, Weinmann C, Welihinda A, Welinder P, Weng J, Weng L, Wiethoff M, Willner D, Winter C, Wolrich S, Wong H, Workman L, Wu S, Wu J, Wu M, Xiao K, Xu T, Yoo S, Yu K, Yuan Q, Zaremba W, Zellers R, Zhang C, Zhang M, Zhao S, Zheng T, Zhuang J, Zhuk W, Zoph B. GPT-4 Technical Report. 2023 Mar 15; Available from: <http://arxiv.org/abs/2303.08774>
31. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S, Herbert-Voss A, Krueger G, Henighan T, Child R, Ramesh A, Ziegler DM, Wu J, Winter C, Hesse C, Chen M, Sigler E, Litwin M, Gray S, Chess B, Clark J, Berner C, Mccandlish

- S, Radford A, Sutskever I, Amodei D. Language Models are Few-Shot Learners. Available from: <https://commoncrawl.org/the-data/>
32. Milella F, Donato Russo D, Bandini S. AI-Powered Solutions to Support Informal Caregivers in Their Decision- Making: A Systematic Review of the Literature. *OBM Geriatr* 2023;7(4). doi: 10.21926/obm.geriatr.2304262
 33. Borna S, Maniaci MJ, Haider CR, Gomez-Cabello CA, Pressman SM, Haider SA, Demaerschalk BM, Cowart JB, Forte AJ. Artificial Intelligence Support for Informal Patient Caregivers: A Systematic Review. *Bioengineering*. Multidisciplinary Digital Publishing Institute (MDPI); 2024. doi: 10.3390/bioengineering11050483
 34. Lindeman DA, Kim KK, Gladstone C, Apesoa-Varano EC, Hepburn K. Technology and Caregiving: Emerging Interventions and Directions for Research. *Gerontologist*. Gerontological Society of America; 2020. p. S41–S49. PMID:32057082
 35. Norton LE, Malloy PF, Salloway S. The Impact of Behavioral Symptoms on Activities of Daily Living in Patients With Dementia. *Am J Geriatr Psychiatry*. 2001.
 36. Alzheimer's Society. The progression and stages of dementia. 2021.
 37. Snowball E, Aiken C, Norman M, Hykaway W, Dempster Z, Itzhak I, McLellan E, McGilton KS, Bethell J. Engaging people with lived experience of dementia in research meetings and events: insights from multiple perspectives. *Frontiers in Dementia* Frontiers Media SA; 2024 Jul 5;3. doi: 10.3389/frdem.2024.1421737

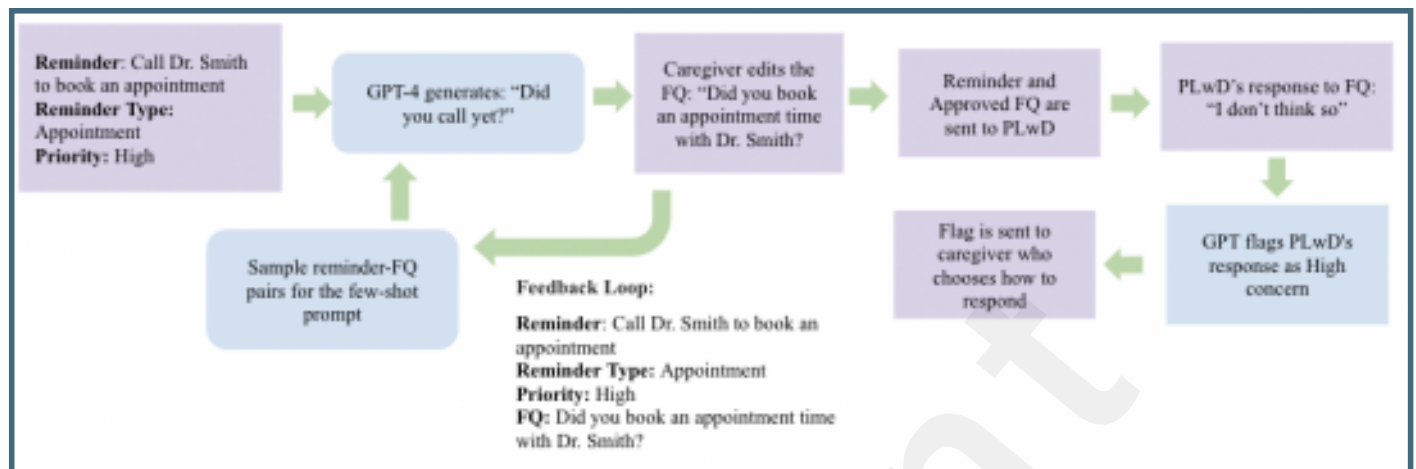
Supplementary Files

Figures

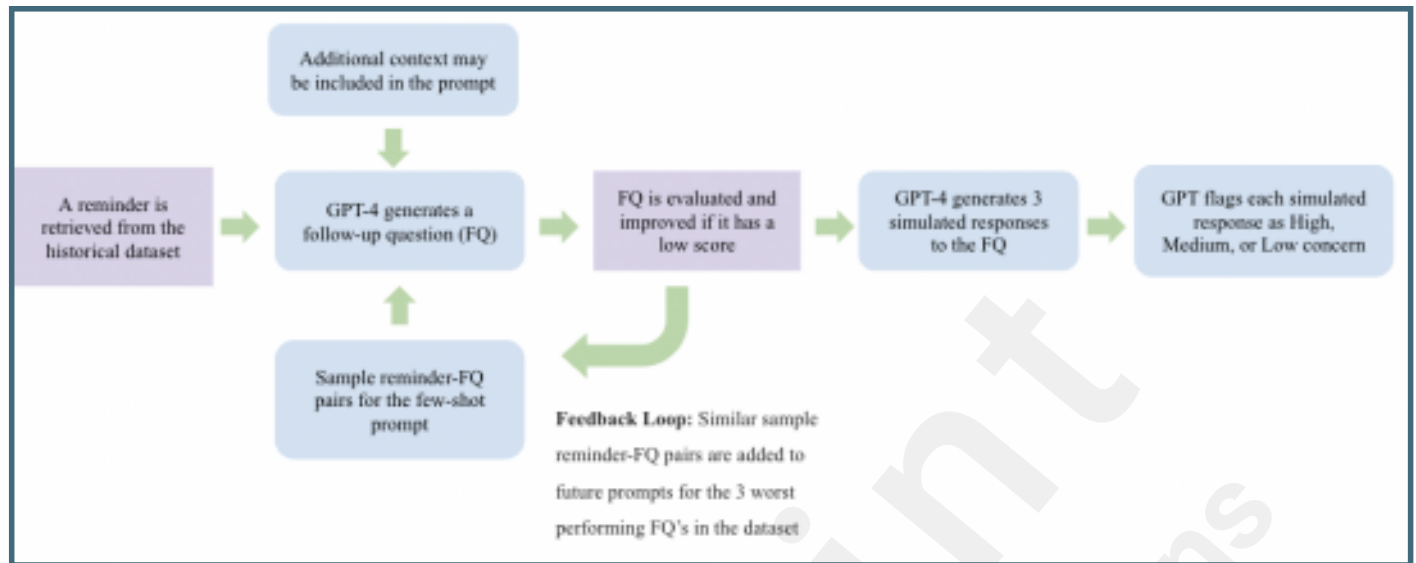
Theoretical pipeline for the system.



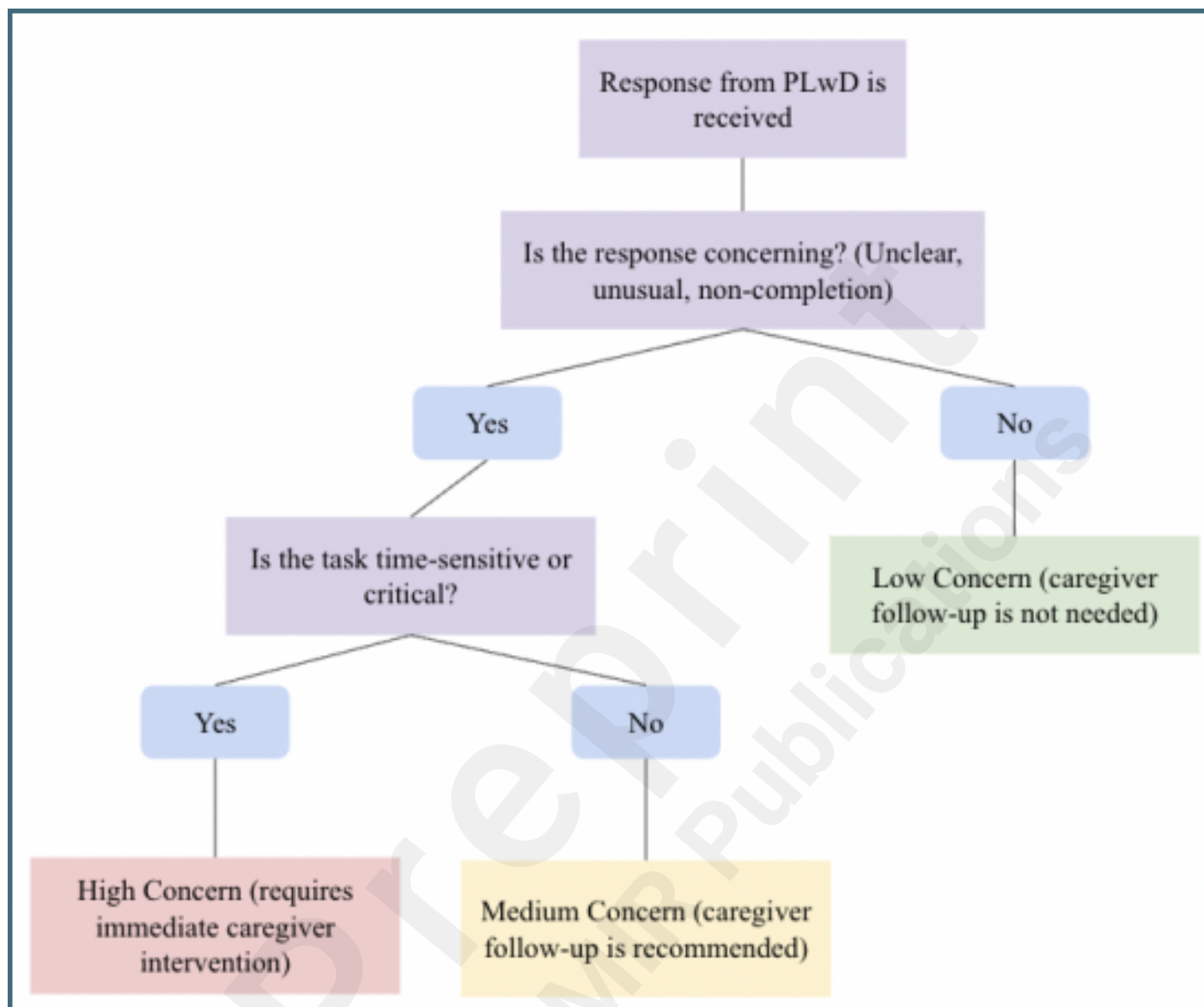
Sample workflow of a reminder through the theoretical pipeline.



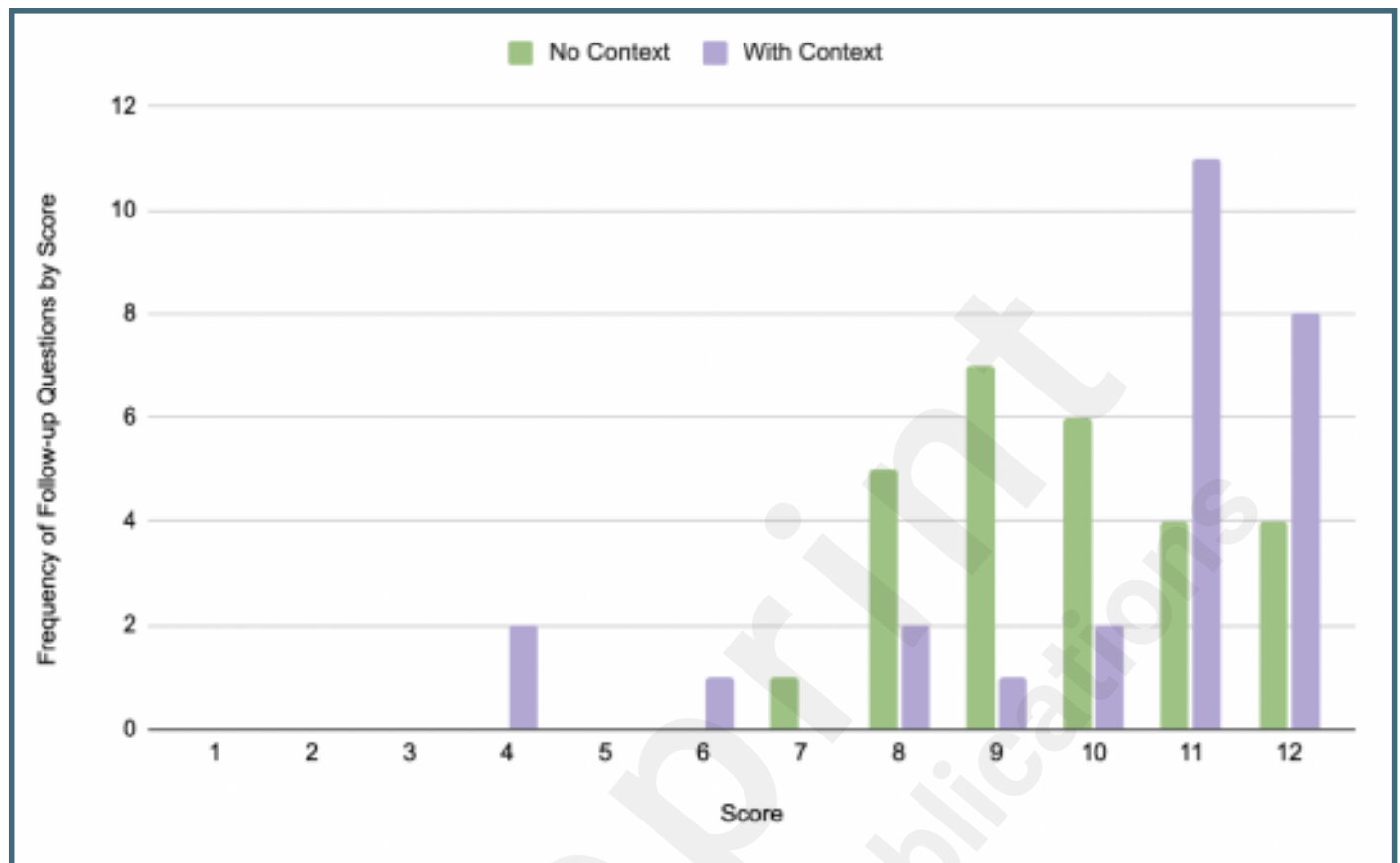
Actual prototyped pipeline to simulate the theoretical pipeline.



Decision tree for response classification.



Evaluation Checklist Score for Dataset 1, with and without added context.



Evaluation Checklist Score for Dataset 2, with and without added context.

