

Helping the Helper: LLM-Assisted Problem Articulation for Older Adults Seeking Technology Support

Hasti Sharifi
University of Illinois Chicago
Chicago, Illinois, USA
hshari3@uic.edu

Homaira Huda Shomee
University of Illinois Chicago
Chicago, Illinois, USA
hshome2@uic.edu

Melissa Lamar
Rush Alzheimer's Disease Center
Chicago, Illinois, USA
melissa_lamar@rush.edu

Sourav Medya
University of Illinois Chicago
Chicago, Illinois, USA
medya@uic.edu

Debaleena Chattopadhyay
University of Illinois Chicago
Chicago, Illinois, USA
debchatt@uic.edu

Abstract

Older adults often struggle to articulate technology support needs due to unfamiliar technical terminology and age-related cognitive changes. We explore how large language models (LLMs) can facilitate this problem articulation process. Through a diary study ($n = 27$), we identified four communication barriers in older adults' queries: *verbosity*, *incompleteness*, *over-specification*, and *under-specification*. To mitigate these, we developed an LLM pipeline that clarifies context and paraphrases unstructured queries. LLM-rephrased queries significantly improved automated solution accuracy (69% vs. 35%). Furthermore, younger adults ($n = 48$) acting as technology helpers understood LLM-rephrased queries better (93.7% vs. 65.8%) and reported greater ease in providing support. Older adults ($n = 34$) also found the resulting solutions highly actionable (94.7%). Finally, we contribute the first synthetic dataset of older adults' technology assistance queries (STAQ). This work demonstrates how LLMs can improve technology support seeking for older adults by addressing age-related communication barriers.

CCS Concepts

• **Social and professional topics** → **Seniors**; • **Human-centered computing** → **Empirical studies in accessibility**; *Accessibility systems and tools*; *Interactive systems and tools*.

Keywords

Technology support, digital support, digital skills, artificial intelligence, large language models, foundation models, older adults, aging

ACM Reference Format:

Hasti Sharifi, Homaira Huda Shomee, Melissa Lamar, Sourav Medya, and Debaleena Chattopadhyay. 2026. Helping the Helper: LLM-Assisted Problem Articulation for Older Adults Seeking Technology Support. In *The 28th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '26)*, October 25–28, 2026, Vila Nova de Gaia, Portugal. ACM, New York, NY, USA, 22 pages. <https://doi.org/10.1145/3797867.3829031>



This work is licensed under a Creative Commons Attribution 4.0 International License. *ASSETS '26, Vila Nova de Gaia, Portugal*
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2521-0/2026/10
<https://doi.org/10.1145/3797867.3829031>

1 Introduction

Digital technologies have become essential for accessing information, obtaining services, maintaining social connections, and participating fully in civic and economic life. Consequently, there is a growing need among older adults to engage effectively with these tools [1, 32, 41]. While a recent survey reports that nearly all U.S. adults aged 50 and older now own at least one device—with smartphone ownership rates comparable to younger adults—device ownership has not inherently translated into broad or sustained use [1, 36]. To bridge the gap between owning digital devices and using them effectively, older adults often require reliable assistance [27, 46, 58].

While limited technology uptake can reflect personal preferences [33, 54], many older adults cite the lack of high-quality technology support as a primary barrier to continued use [1, 58]. This need for effective technology support is compounded by several age-related factors. Many older adults encounter contemporary digital systems later in life, without the structured learning or informal exposure that younger generations experience through school, work, and social contexts [17]. Furthermore, age-related cognitive changes can make it more difficult to learn and retain new information [6, 46, 58], frequently leading to a lack of digital confidence [5, 12]. These challenges and the fluidity of information to be processed by older adults are exacerbated by the fast-evolving nature of software. As applications frequently update and new interaction paradigms emerge, older adults are forced to acquire entirely new mental models or schema—internal representations of how systems function [24]. Forming these new models requires significant cognitive effort, which may tax an already aging working memory capacity [63], particularly during moments of frustration or urgency (e.g., making repeated unsuccessful attempts [76] or rushing to cancel a rideshare before a fee is charged [58]). Consequently, even minor system changes can disrupt an older adult's ability to navigate a familiar tool or explain what went wrong.

When these disruptions occur, older adults typically prefer to seek help from friends, family members, or trusted community resources [4, 19, 30, 40, 57, 58]. Alternatively, some, with higher digital skills, attempt to resolve issues independently by turning to search engines or artificial intelligence (AI)-powered chatbots [29, 58]. In either scenario, whether interacting with a human helper or an automated system, the first step toward resolving a technology

problem is being able to communicate it clearly. However, articulating a technology problem requires more than simply stating that an error occurred; it demands accessing specific vocabulary, applying a mental model of the system, and doing so under cognitive constraints [60].

Age-related changes in processing speed and working memory capacity can complicate this task [10, 25, 37, 44]. For example, reductions in these cognitive domains that naturally occur with aging may make real-time problem-solving difficult and the articulation of the underlying issue much more challenging [31]. Furthermore, while older adults often maintain broader general vocabulary knowledge than their younger counterparts [10], this advantage rarely extends to technology-related terms that are typically acquired later in life. Moreover, common cognitive conditions such as mild cognitive impairment (MCI) can further impair cognitive abilities beyond expected age-related alterations, making it significantly more difficult to learn a new skill, adapt to system changes, or articulate questions to acquire appropriate solutions [16, 65].

Prior research highlights that older adults frequently experience uncertainty and a lack of confidence when attempting to phrase technology questions effectively [19, 57, 58]. This uncertainty makes technology support problematic on both ends; technology helpers frequently report difficulty understanding older adults' problem descriptions, particularly during remote or asynchronous communication [57]. Despite these documented challenges, most AI-powered support systems assume older adults can pose well-formed queries and anchor them to a specific interface or interaction context [21, 68, 73]. Consequently, there remains a critical gap between how systems expect users to ask for help and how older adults naturally communicate their technical struggles.

In this paper, we examine how English-speaking, community-dwelling older adults communicate technology problems in unstructured contexts, and how large language models (LLMs) can facilitate the problem articulation process by translating natural descriptions into actionable queries. By reducing the cognitive and verbal burden on older adults, we aim to better align technology support systems with real-world help-seeking practices. Specifically, we contribute: (1) a characterization of organically occurring technology support queries from older adults; (2) a few-shot prompt-chaining approach that clarifies, paraphrases, and generates solutions for these queries; (3) an evaluation of this pipeline with both younger and (4) older adults; and (5) the development of STAQ (pronounced "stack"), a dataset of synthetic technology assistance queries designed to support the training of age-inclusive AI systems. Our findings underscore the necessity of designing technology support tools that reflect the realities of aging, communication, and cognition within sociotechnical ecosystems.

2 Background

To situate our work, we review prior literature across three areas: older adults' help-seeking practices, age-related barriers in technology problem articulation, and recent AI applications in technology support.

2.1 Older Adults and Technology Support

The definition of an "older adult" is not fixed to a specific chronological age [26, 67]. In the literature, definitions vary widely, ranging from individuals aged 50 and above to those 70 and older [49, 71]. Aging is often experienced relative to specific environments, objects, or rapid technological changes [8]. In this work, we define older adults as individuals aged 60 or over and adopt an anti-deficit model of aging. This perspective avoids treating older users as a "problematic other" or viewing aging as inherently disabling [34]. Instead, it recognizes that older adults are active, discriminating users who evaluate technology based on perceived risk, personal values, and cultural expectations [33, 57]. Notably, the need for technology support is not universal among this demographic. Many older adults possess high digital proficiency and frequently act as technology helpers for their peers or other family members [22, 57, 58].

However, when older adults do struggle with technology use—for instance, when encountering unfamiliar applications or system updates—effective technology support remains critical to prevent abandonment [19, 46, 57, 58, 74]. While early technology acceptance models emphasized facilitating conditions like instruction manuals [6, 52, 66], contemporary research indicates a strong preference for interactive and socially situated learning [46, 59]. During initial onboarding, older adults may rely on professional instructors or structured classes [22, 45]. However, during continued use, older adults often turn to informal social networks, seeking step-by-step guidance from trusted friends, family members, or peers [19, 57, 58], or they engage in independent troubleshooting through trial-and-error, online searches, or AI tools [19, 21, 75]. In either case, the success of the intervention depends heavily on the older adult's ability to clearly communicate the problem to their technology helpers or support tools.

2.2 Age-Related Barriers in Technology Problem Articulation

Although older adults' technology use and help-seeking behaviors have been extensively studied [6, 40, 46, 58, 64], far less attention has been given to *how* they communicate their technology problems. Articulating any complex problem is an inherently demanding cognitive task. It requires working memory to hold the problem's parameters and executive function to retrieve the precise language to describe it [55, 56]. While older adults typically maintain their general vocabulary (crystallized intelligence), they frequently experience age-related declines in fluid cognitive abilities, such as processing speed and working memory [48, 55]. During language production, these fluid changes manifest as real-time word-finding difficulties, tip-of-the-tongue states, and specific declines in action verbal fluency—the ability to rapidly generate verbs [11, 31]. Consequently, when explaining an unfamiliar situation, the high cognitive load of navigating the problem space frequently disrupts efficient word retrieval [63]. This cognitive bottleneck can force individuals to substitute precise terminology with generalized phrasing, vague referents, or lengthy narratives, creating a fundamental barrier to technology problem communication.

Rather than formulating precise questions, older adults frequently produce ambiguous queries that lack necessary context or rely on vague descriptions of symptoms [19, 60]. The cognitive burden of

translating a system error into an actionable question can be so high that older adults may struggle silently or abandon tasks altogether, a challenge that has prompted recent research into detecting user frustration implicitly through interaction logs rather than relying on explicit queries [27]. Recent reflexive research also suggests that difficulties in problem articulation should not be viewed merely as individual cognitive deficits; they are symptoms of rapidly evolving sociotechnical systems that regularly outpace the mental models of all users, but particularly affect those without continuous technology exposure [36]. In structured user studies with well-defined tasks and think-aloud protocols, prior work has found that older adults tend to use specific question words, such as “what,” “why,” and “how” when encountering interaction issues [18] and ask distinct types of questions during help-seeking from technology helpers [76].

A limitation across prior studies examining these communication barriers, is their reliance on specific applications, structured tasks, and controlled settings [19, 21, 68, 73, 75, 76]. In organic contexts, problem communication may be significantly more unstructured, lacking in context, and far removed from the precise terminology expected by automated technology support tools or technology helpers.

2.3 Modern Support Tools using Foundation Models

The emergence of large language models (LLMs) and vision-language models (VLMs) has shifted technology support from static knowledge bases to dynamic, conversational interventions. Contemporary systems increasingly leverage these foundation models across three primary paradigms: automated customer support [62, 69], tool learning—where models autonomously execute software tasks [50]—and conversational assistants designed to guide software navigation [28, 29]. To ground these conversational assistants in the user’s specific context, developers employ different methods for parsing graphical user interfaces (GUIs). Some tools extract the application’s underlying view hierarchy and convert it into a structural representation (e.g., HTML) [68], whereas others rely on VLMs to directly encode the visual and structural elements of the screen [19, 73].

In technology support tools designed specifically for older adults, prior work has leveraged LLMs to mitigate visual and cognitive load, such as assisting users in finding a particular GUI element or remembering a list of task steps. Techniques like dynamically reducing an interface’s search space [75] demonstrate how intelligent systems can simplify the navigation of complex applications. Recent systems such as EasyAsk [21] and GuideMe [19] use foundation models to generate contextual, in-app troubleshooting guidance for independent use. EasyAsk employs a multimodal approach, combining voice queries and physical screen touches with an application’s underlying view hierarchy to generate interactive tutorials [21]. GuideMe uses VLMs to process raw screenshots alongside natural language queries, generating step-by-step visual instructions overlaid directly on the interface [19]. However, the effectiveness of these assistants depends strictly on the availability of rich contextual inputs—specifically, direct access to the system’s screen state or physical interaction with the interface [19, 21].

Despite these advancements, the burden of problem articulation remains a critical bottleneck. Relying on granular screen access [21] or manual screenshots [19] introduces additional effort and privacy concerns for older users. Even when AI generates accurate instructions, non-expert users may struggle to map that text back to the interface [29], yet strongly prefer guided, “do-it-with-me” assistance over completely autonomous execution [28]. Consequently, while foundation models offer unprecedented troubleshooting capabilities, their practical effectiveness is fundamentally limited by the initial problem articulation phase. If older adults cannot effectively communicate their contextual intent, the resulting guidance—whether AI-generated or human-provided—will inevitably fail to align with their actual needs.

3 How Older Adults Communicate Technology Problems

In this section, we examine how English-speaking, community-dwelling older adults describe technology-related problems as they naturally occur in their everyday life. We first detail the methodology of our formative diary study, and then present a qualitative analysis of the unstructured queries collected from our participants.

3.1 Methods

We conducted an online diary study to examine how older adults communicate technology problems in real-world contexts. Eligible participants were aged 60 or older, used a mobile device (e.g., smartphone or tablet) at least once per week, and were proficient in English. We focused specifically on mobile devices, as they are currently the primary computing platforms used by older adults [1]. Participants were recruited through community flyers, mailing lists, social media platforms, and word-of-mouth. The diary study lasted up to eight weeks, though participants could withdraw at any time. No monetary incentives were provided. After initial contact, a researcher (first author) met with participants to introduce the research study. Participants were asked to message the research team with any mobile device–related technology questions using their preferred communication method (e.g., email, text, or WhatsApp). This approach aligns naturally with asynchronous remote help-seeking, such as via text or email, as well as with independent queries to search engines or AI agents [58, 59]. Participants could communicate via text, screenshots, voice messages, or video recordings of their screens. During the first meeting, each participant received both written instructions and video tutorials on how to record audio, screen, and video using their device. They were asked not to share any personal information at any point during the study. To encourage participation, automated reminders were sent every two days, and members of the research team responded with technical assistance as an incentive. All communication, aside from the initial onboarding, occurred asynchronously.

Midway through the study, we observed low engagement, with participants submitting approximately one query per week on average. Follow-up interviews revealed that many older adults typically reached out to family members or trusted technology helpers first and only contacted the research team if the issue remained unresolved. To address this, we expanded recruitment to include adult technology helpers (aged 18 or older) who had previously assisted

Table 1: Key Characteristics of Older Adults' Technology Queries ($N = 57$)

Characteristic	Count	Example(s)
Verbosity	9	“These are the START SCREENS screenshots on [name]’s computer. One for searching, the usual purpose for booting up, and one for Click Bait. My computer combines the two and I do spend some wasted time clicking on the icons. Any way to combine them? Or better, just dispense with the clickbait. And BTW, the favorites are not necessarily the same. She hasn’t complained about hitting Facebook and getting my Facebook account — lately.” — 80, man “My wife [name] has been advised by trusted source that... Actually, it sounds like something we should both do from time to time... I’m also confused.” — 70, man
Over-specification	12	“Argh... I have had this problem for a while and don’t even know how or why it happened. I want to watch my wolf utubes with my phone or iPad in the horizontal position. All of a sudden it won’t turn when I turn my phone. What happened?!! How do I fix it?!!” — 66, woman “Need mobile recharge statement for reimbursement but not able to download from app. My son is out of station and he asked for recharge statement for his company for reimbursement.” — 60, woman
Under-specification	13	“I’m trying to watch a cooking video on YouTube, but I can’t understand what they’re saying. Can you help me?” — 62, woman “I’m really having trouble with my Kindle app. It’s been acting weird — the books I read daily aren’t syncing between my phone and tablet, and I can’t download any new ones. Also, the text is too small, and I have no idea how to fix it.” — 60, man
Incompleteness	24	“I come back to phone and it’s blown up so huge I can’t move it. [name] helped me resolve it by just clicking on it. There’s gonna be one every day.” — 61, woman

older adults. These helpers were asked to document asynchronous help interactions by submitting message transcripts, summarizing the issue in their own words, and reporting whether the issue was successfully resolved. As is common in longitudinal research with older adults, not all participants remained engaged for the full eight weeks [2, 15].

3.2 Results

We collected diary entries from 27 older adults (13 women; Mdn age = 66, age range = 60–87). Of these participants, 17 were recruited indirectly through their technology helpers. Over the study period, the number of queries submitted by each participant ranged from one to ten. All queries were unique; no participant asked about the same issue more than once. Participants reported using a variety of mobile devices, including iPhones, iPads, Android smartphones, and Amazon Fire tablets.

We collected 57 queries in text format, nine of which included a screenshot. No voice or video recordings were provided. These messages spanned 34 distinct mobile applications. Although all queries addressed different problems, two remained unresolved due to participant attrition. The dataset included both the original query and the full transcript of the asynchronous conversation that followed, either with a technology helper or a researcher. Query length varied substantially, ranging from 5 to 223 words ($Mdn = 65$, $IQR = 27$).

To categorize the *types* of questions, we first applied an existing taxonomy of older adults’ technology-related queries [76], which classifies problem articulation into five distinct categories: validation (verifying if an action is correct), directed informational (asking

how to execute a specific step), undirected informational (lacking intuition on what UI features to use next), navigational (locating a feature), and conceptual (understanding the system’s underlying model). Two researchers independently coded the dataset using a coding reliability thematic approach [9], achieving substantial interrater agreement (Cohen’s $\kappa = 0.72$). Following consensus, the final classification yielded four validation, 14 directed informational, 23 undirected informational, five navigational, and 11 conceptual questions. Although undirected informational queries dominated our corpus, prior research using this taxonomy observed that older adults primarily asked validation questions [76]. This likely reflects methodological differences. Earlier tasks were synchronous and conducted with a researcher present, likely prompting users to seek reassurance—which was not feasible in our asynchronous, naturalistic setting.

In addition to question type, we conducted a reflexive thematic analysis [9] to examine how older adults articulated their technology problems. The first author analyzed the queries inductively using open coding, referencing the entire conversation as needed. Then, the codes were discussed and iterated in group data analysis sessions until key themes were identified. To identify the open codes, we grounded our analysis in the aging literature about verbal knowledge, verbal fluency, and age-related changes in fluid and crystallized intelligence. We identified four recurring characteristics in how queries were articulated: excessive detail that obscured the core issue (*verbosity*), inclusion of non-essential specifics (*over-specification*), omission of relevant information (*under-specification*), and a lack of critical context (*incompleteness*).

The technology queries often suffered from verbosity, incorporating superfluous details that delayed a clear understanding of the

underlying issue (e.g., “My wife [name] has been advised by trusted source that...Actually, it sounds like something we should both do from time to time...I’m also confused...” —70, man). In contrast, other requests were concise but filled with highly specific situational details, complicating the identification of the root problem (e.g., “Need mobile recharge statement for reimbursement but not able to download from app. My son is out of station and he asked for recharge statement for his company for reimbursement” —60, woman). Queries also frequently lacked granular but critical context—such as whether a user was trying to cancel an ongoing or a scheduled rideshare—or they compounded multiple issues without defining the scope of each sub-problem (e.g., “I’m really having trouble with my Kindle app. It’s been acting weird - the books I read daily aren’t syncing between my phone and tablet, and I can’t download any new ones. Also, the text is too small, and I have no idea how to fix it.” —60, man). In this example, it remains unclear if the text size issue applies to the tablet, the phone, or just the newly downloaded books. Ultimately, the most prevalent characteristic of these queries was the absence of necessary diagnostic information, specifically the device type, operating system, or application in question. Furthermore, participants occasionally used unconventional or metaphorical language that required interpretation—for example, referring to voice commands as ‘tongue,’ describing a locked phone as ‘power going off,’ or calling dark mode ‘screen going dark.’ Table 1 provides illustrative examples of each characteristic. Note that a query may exhibit multiple characteristics, such as both verbosity and incompleteness.

In summary, our analysis identified several characteristics in older adults’ technology problem articulation that may hinder problem resolution. These patterns, along with occasional use of unconventional vocabulary, point to a need for interventions that support clearer, more effective communication. To address this, we next explore how large language models can be used to reformulate older adults’ queries so that they can be more readily resolved—whether by asking someone, searching online, or interacting with an AI-based support tool.

4 Reformulating Technology Problems with Large Language Models

After examining the different characteristics of older adults’ technology queries, we set out to reformulate them so that they can be resolved independently, for example, by searching online or asking a chatbot, or communicated effectively to technology helpers. When designing a solution, we chose to address two age-associated characteristics of our dataset: (1) 9/57 entries included a screenshot of the relevant user interface (UI), implying screen-related information is a rarity, and (2) a lack of essential contextual information that the technology helpers had to elicit via follow up questions in order to resolve the problem. These patterns challenge the assumptions of existing question-answering solutions that pertinent UI information is always available, [19, 21, 68, 73]. To address these issues, we next explore how large language models (LLMs) can be used to reformulate older adults’ original queries—making them easier to understand, and ultimately easier to resolve.

4.1 Methods

We selected two models, GPT-4o [3] and OS-ATLAS [73], to develop our approach. GPT-4o is an LLM, and OS-ATLAS is a foundational action model developed for graphical user interface (GUI) grounding and agentic tasks beyond its training distribution. We included OS-ATLAS due to its ability to operate without relying on external APIs and its strength in GUI grounding. GPT-4o was chosen as a general-purpose language model for its strong generalization capabilities and versatility [73]. Compared to other LLMs, such as Gemini, LLaMA, and Claude, GPT-4o demonstrates superior reasoning ability, causal coherence, and overall accuracy [39].

For our evaluation pipeline, GPT-4o was specified as OpenAI’s Chat Completions API with the model identifier `gpt-4o`. Because we did not pin a dated snapshot in the initial API configuration, calls used the dynamic production version active during the evaluation window, December 2024–March 2025. OS-ATLAS was specified as OS-Copilot/OS-Atlas-Base-7B, built on the open-source Qwen2VLForConditionalGeneration architecture with a Qwen2-VL-7B backbone, and evaluated locally using the Hugging Face Transformers library [72].

We used a subset of queries collected in the diary study (described in Section 3) for prompt engineering and model evaluation. Only queries that were understandable without screenshots and had been successfully resolved were included. Two researchers independently reviewed the 51 eligible queries and, when available, their follow-up messages to identify the underlying technology problem. Interrater reliability was high (Cohen’s $\kappa = .89$). The final dataset consisted of these 51 queries, their associated conversations, and expert annotations of the problems.

We developed confidence-aware prompts for GPT-4o and OS-ATLAS using an iterative few-shot prompting strategy, with separate iterations tailored to each model’s architecture. Although the final prompts varied slightly, the same three example queries from our dataset were used in the workflow across both models to ensure consistency. Each prompt guided the model through four tasks: (1) analyze the original query to identify missing contextual details; (2) generate up to three follow-up questions to elicit that information; (3) paraphrase the query into one or more concise, complete questions suitable for a Google search; and (4) provide a single, confident solution—defined as at least 90% confidence for GPT-4o and 80% for OS-ATLAS (Figure 1).

Throughout the iterations, we tested the prompts using representative examples covering diverse query types, including step-by-step instructions, explanations of technology concepts, and handling of impossible asks. Feedback from these tests was used to refine the prompts, addressing issues such as redundancy in follow-up questions, absence of important details in the paraphrased question, or use of overly technical terminology in the solution. Instructions were added to allow the models to break down complex requests into multiple paraphrased questions. Both models were explicitly directed to prioritize clarity and simplicity in their solutions, avoiding multiple alternatives to ensure usability for older adults. The iterative refinement process ensured that each model’s prompt achieved good performance while aligning with the design objectives. To reduce hallucination risk, we implemented confidence thresholds. Although language model “confidence” is a heuristic rather than a

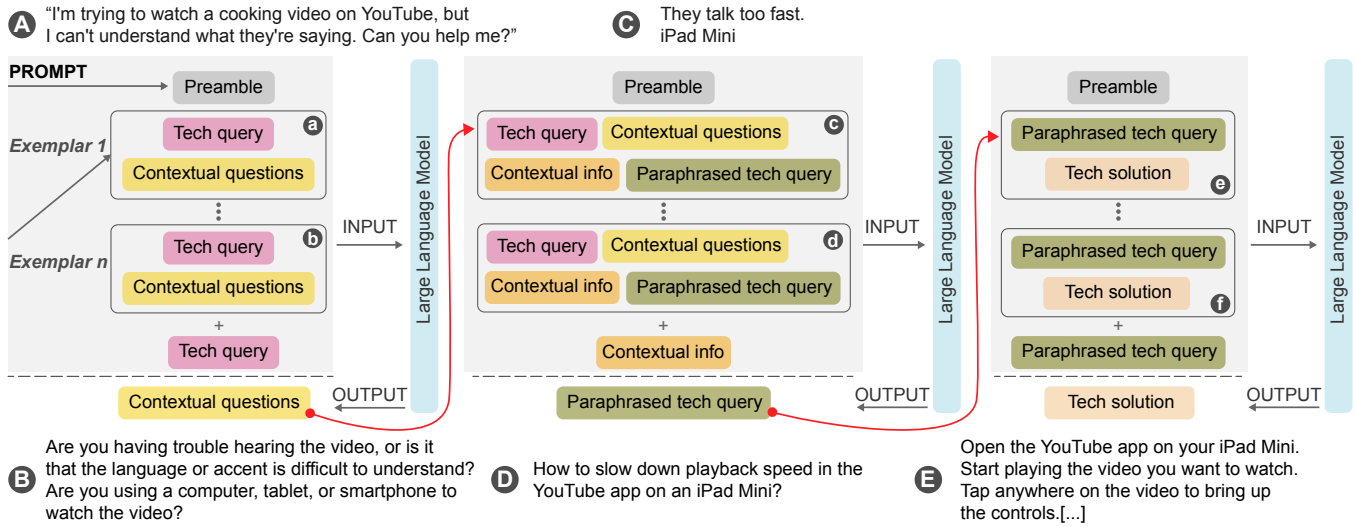


Figure 1: Our *few-shot prompt chaining* approach with large language models operates as follows: First, a query from an older adult is input into the model (A), which is then instructed to generate follow-up contextual questions (B). The model is provided with a few examples (a, b) for guidance. Older adults would then answer these questions to supply contextual information (C). Next, the model takes the original query, contextual questions, and provided contextual information as inputs to generate a paraphrased query (D), with additional examples (c, d) included for reference. Finally, using another set of examples (e, f), the model is asked to generate a solution to the paraphrased query (E).

statistical measure, these thresholds ensured that solutions were only generated when the model demonstrated sufficient comprehension, thereby improving response reliability. A lower threshold was necessary for OS-ATLAS, because it is a specialized action model rather than a general-purpose LLM; applying the 90% threshold caused it to over-trigger safety instructions and frequently return “I don’t know.” The 80% threshold ultimately provided a better balance between hallucination mitigation and response yield. The final prompts are available in the Appendix A.

4.2 Results

4.2.1 Model Workflow. GPT-4o was accessed through the OpenAI API, while OS-ATLAS was run locally. For each of the 48 original queries included in the evaluation (3/51 was used in prompt engineering), we followed a standardized workflow:

- (1) The original query is input into the model, which is then instructed to generate follow-up contextual questions.
- (2) Responses to these generated questions, i.e., contextual information, are provided to the model from our corpus collected through the diary study. If no relevant response is available, the model receives “I do not know” as input. Note that our corpus included not only the queries but also the conversations between technology helpers and older adults leading up to problem resolution. Most resolutions required eliciting additional contextual information.
- (3) The model is subsequently instructed to paraphrase the original query into a concise, complete format suitable for a simple Google search.
- (4) Finally, the model is asked to provide a single solution to the paraphrased query. Solutions may include step-by-step

instructions, conceptual explanations, or a brief rationale for why the query could not be resolved.

4.2.2 Model Evaluation. Model outputs were evaluated for *paraphrasing* and *solution-giving* effectiveness. Paraphrasing effectiveness was assessed based on expert judgment and categorized into three levels: correct, partially correct, and incorrect (see Appendix A for examples). Partially correct paraphrasing either omitted essential parts of the original query or introduced hallucinated details. Two experts (the first and second authors) independently rated the paraphrasing effectiveness of both models, achieving inter-rater agreement (Cohen’s κ) of .75 for OS-ATLAS and .85 for GPT-4o. Disagreements were discussed and resolved before final categorization (Table 2).

Expert evaluation of paraphrasing effectiveness revealed that GPT-4o substantially outperformed OS-ATLAS. GPT-4o produced correct paraphrases for 68.8% of the queries, compared to only 12.6% for OS-ATLAS. OS-ATLAS outputs were more frequently rated as partially correct (43.7%) or incorrect (43.7%), whereas GPT-4o had far fewer partially correct (25%) and incorrect (6.2%) responses. We did not evaluate the paraphrased queries with older adults for two reasons. First, the goal of paraphrasing was to improve the likelihood of obtaining a timely and accurate solution, rather than to assess how older adults perceived the paraphrased queries. Second, the same challenges that led older adults to generate problematic original queries would likely prevent them from accurately judging the effectiveness of the paraphrased queries [42, 70].

Solution-giving effectiveness was assessed based on the perceived accuracy of the solution, categorized into three levels: correct, partially correct, and incorrect (see Appendix A). The ground truth

Table 2: Expert evaluation of paraphrasing effectiveness (%)

Evaluation	GPT-4o	OS-Atlas
Correct	68.8	12.6
Partially correct	25	43.7
Incorrect	6.2	43.7

for this evaluation was available because all those technology problems had been resolved during the diary study with older adults. A correct solution for each technology query was created by the first author, who extracted the correct steps from the diary study conversations and structured them following the same instructions given to the models in the prompts. These diary-study resolutions served as a reliable baseline by confirming that each problem was solvable and showing one successful workflow. However, during evaluation, experts assessed whether each model output was functionally accurate, ensuring that technically valid alternative solutions were scored as correct if they resolved the core issue. A comparison of the two models' effectiveness is presented in Table 3. In the Google search condition, all correct solutions appeared within the first *five* results, excluding AI-generated content. Only the top 20 Google search results were reviewed to identify a solution. Both models responded with "I do not know" to three queries, one of which was common to both. Additional quantitative evaluation of the model outputs is available in the Appendix A.

Our findings demonstrate that LLMs—particularly GPT-4o—can effectively reformulate older adults' technology-related queries, even when those queries are vague, imprecise, or lacking critical context. We next examine how LLM-rephrased queries are perceived and interpreted by technology helpers, who play a critical role in resolving older adults' technology issues.

5 Evaluation of LLM-Rephrased Queries with Technology Helpers

Older adults often rely on younger individuals for technology support. However, prior research suggests that younger technology helpers frequently struggle to understand the problem before they can offer effective assistance [19, 21, 57, 58]. Ideally, LLM-rephrased queries should help mitigate this communication barrier. To evaluate this possibility, we conducted a controlled experiment with younger adults to assess whether LLM-generated paraphrases of older adults' technology queries improved perceived understanding, confidence, and ease of interpretation relative to the original queries.

5.1 Methods

Adults aged 18 to 50 were recruited to participate in a controlled experiment evaluating LLM-rephrased technology queries. Eligibility criteria included age and demonstration of at least professional working proficiency in English, verified through a brief language screener. No participants from our diary study (Section 3) was part of this evaluation. Participants were recruited via social media platforms and word-of-mouth referrals. No monetary compensation was provided. The experimental materials consisted of query pairs: original queries collected from older adults (as described in Section

3) and their corresponding LLM-reformulated versions generated by GPT-4o using the pipeline outlined in Section 4. To ensure content quality, only query pairs that had received a *correct* paraphrasing rating during expert evaluation were included (69% of 48 queries). A total of 33 query pairs were included in the study. Each participant rated five of these pairs, which were selected using a Latin square design to ensure balanced and systematic coverage across participants.

Query pairs were presented sequentially—original followed by rephrased—with the order of the pairs randomized for each participant. Participants were not informed which version had been reformulated by an LLM or whether either query was drawn from real users. For each query, participants were first asked whether they understood the query (response options: yes, no, maybe, I don't know). If they selected any response other than no, they were then asked to rate their confidence in understanding and ease of understanding, both on 5-point Likert scales.

In addition to these comprehension-related measures, participants completed demographic questions and an adapted version of the Digital Literacy Scale [7] (Appendix B) to assess their own digital proficiency. All procedures were conducted online. Following informed consent and eligibility screening, participants reviewed a randomized set of query pairs. All study materials were administered using the Qualtrics platform.

With query type (original vs. LLM-rephrased) as the independent variable, and comprehension, perceived confidence, and perceived ease of understanding as dependent variables, we tested the following hypotheses:

H1: Participants will understand significantly more rephrased queries than original queries.

H2: Participants will report significantly greater confidence in understanding rephrased queries than original queries.

H3: Participants will find rephrased queries significantly easier to understand than original queries.

5.2 Results

5.2.1 Participants. Forty-eight participants (*Mdn* age = 28 years, *IQR* = 3.25,8 range = 19–47) completed the study. Each of the 33 unique query pairs was evaluated by between 3 and 6 participants (*Mdn* = 5.0, *IQR* = 1.0). The sample included 25 men, 18 women, 3 participants who identified as genderqueer or gender non-conforming, 1 trans male/trans man, and 1 participant who preferred not to disclose their gender. Participants represented a diverse range of ethnic backgrounds: 19 identified as White, 11 as Asian, 9 as Middle Eastern or North African, 4 as Hispanic/Latino, 2 as Black or African American, 2 as Other, and 1 declined to report their ethnicity. Participants reported high digital literacy (*Mdn* = 4.88, *IQR* = 0.31, on a 5-point scale). Most participants (*n* = 44) indicated either native or full professional proficiency in English. Additionally, a majority (*n* = 31) had experience helping older adults with technology, primarily in a personal capacity; among these, 23 participants (74%) had been providing technology support for more than five years.

5.2.2 Hypotheses Testing. Participants reported understanding the rephrased queries significantly more often (*n* = 149, 93.7%) than the original queries (*n* = 106, 65.8%). A McNemar's test confirmed that

Table 3: Expert evaluation of solution effectiveness (%)

Paraphrasing by Solution by	None Google	GPT-4o Google	OS-Atlas Google	OS-Atlas OS-Atlas	GPT-4o GPT-4o
Correct	35.4	68.8	35.4	10.4	68.8
Partially correct	2.1	22.9	10.4	10.4	14.6
Incorrect	62.5	8.3	54.2	79.2	16.7

this difference was statistically significant, $\chi^2(1, N = 55) = 35.20$, $p < .001$, with a large effect size ($OR = 10.0$). Thus, H1 was supported.

A Wilcoxon signed-rank test revealed that participants expressed significantly greater confidence in understanding the rephrased queries ($Mdn = 5.00$, $IQR = 0.00$) compared to the original queries ($Mdn = 4.00$, $IQR = 1.50$), $V = 287.00$, $p < .001$, with a medium effect size ($r = .50$), supporting H2 (Figure 2).

Participants also rated the rephrased queries as significantly easier to understand ($Mdn = 5.00$, $IQR = 0.50$) than the original queries ($Mdn = 4.00$, $IQR = 1.00$), $V = 395.00$, $p < .001$, with a medium effect size ($r = .56$), supporting H3 (Figure 2).

5.2.3 Additional Analysis. Spearman's rank-order correlations were conducted to examine the relationship between participants' self-reported digital literacy and understanding, confidence, and ease when responding to the original and rephrased queries. For original queries, digital literacy was weakly positively correlated with understanding, $\rho = .12$, $p = .041$, and weakly negatively correlated with confidence, $\rho = -.12$, $p = .038$. No significant correlations were observed for rephrased queries: digital literacy was not significantly associated with understanding, confidence, or ease. These results suggest that participants with higher digital literacy were slightly more likely to understand the original queries but paradoxically felt less confident in their responses.

To further explore the impact of rephrasing, an improvement score was calculated as the difference in understanding between rephrased and original queries. A Spearman correlation revealed a significant negative association between digital literacy and improvement in understanding, $\rho = -.33$, $p = .022$. This indicates that participants with lower digital literacy benefited more from rephrased queries, reporting greater gains in understanding than their more digitally literate counterparts.

We also tested whether prior experience as a technology helper to older adults influenced comprehension. A Mann-Whitney U test indicated no significant difference in comprehension of original queries between technology helpers and non-helpers, $U = 3243.00$, $p = .193$. To determine whether technology helpers and non-helpers reported equivalent understanding of rephrased queries, a Two One-Sided Tests (TOST) procedure was conducted [35]. Equivalence bounds were set at ± 0.50 standardized units, corresponding to a moderate effect size (Cohen's d). The observed mean difference in understanding was small (M difference = 0.08), and both one-sided tests were statistically significant, $ps = .030$ and $< .001$. These results support the conclusion that technology helpers and non-helpers reported statistically equivalent levels of understanding for rephrased queries, within a moderate equivalence margin.

In this section, we showed that LLM-rephrased queries improved comprehension, confidence, and clarity among younger adults, including those with prior experience supporting older adults with technology. However, communicating the problem is only one part of the support process. In the next section, we assess whether the LLM-generated contextual questions and solutions are understandable and useful to older adults themselves.

6 Evaluation of LLM-Generated Solutions with Older Adults

In Section 4, we introduced our few-shot prompt chaining approach (Figure 1), which rephrases the original technology query and generates a corresponding solution. However, the success of this pipeline hinges on two critical factors: first, whether older adults feel comfortable and capable of answering the contextual follow-up questions generated by the model; and second, whether the solutions themselves are understandable, actionable, and elicit confidence. Although expert raters found the GPT-4o-generated solutions to be accurate 69% of the time, their practical value depends on how well older adults can understand and follow them. In this section, we assess those aspects with older adults.

6.1 Methods

We conducted a controlled experiment to evaluate older adults' perceptions of LLM-generated follow-up questions and solutions. Participants were eligible if they were aged 60 years or older and demonstrated at least limited working proficiency in English, verified through a self-assessment screener. No participants from our diary study (Section 3) was part of this evaluation. Recruitment was conducted via social media platforms and word-of-mouth referrals. No monetary compensation was offered.

Participants evaluated LLM-generated follow-up questions based on three criteria: (1) whether they felt able to answer the question (response options: yes, no, maybe), (2) their confidence in their answer, and (3) the perceived ease of answering it. They also evaluated LLM-generated solutions using a parallel set of measures: (1) whether they felt able to follow the solution (response options: yes, no, maybe), (2) their confidence in following it, and (3) the perceived ease of doing so. If participants selected any response other than no, they were asked to rate their confidence and ease using 5-point Likert scales.

Eighty-five unique follow-up questions and 36 unique solutions generated by GPT-4o were included for evaluation. Only solutions that had received a correct rating from expert evaluators in Section 3 were included to ensure content quality. To personalize the study experience, participants were first asked which mobile devices they currently use (e.g., smartphone, tablet, Kindle). Survey content was

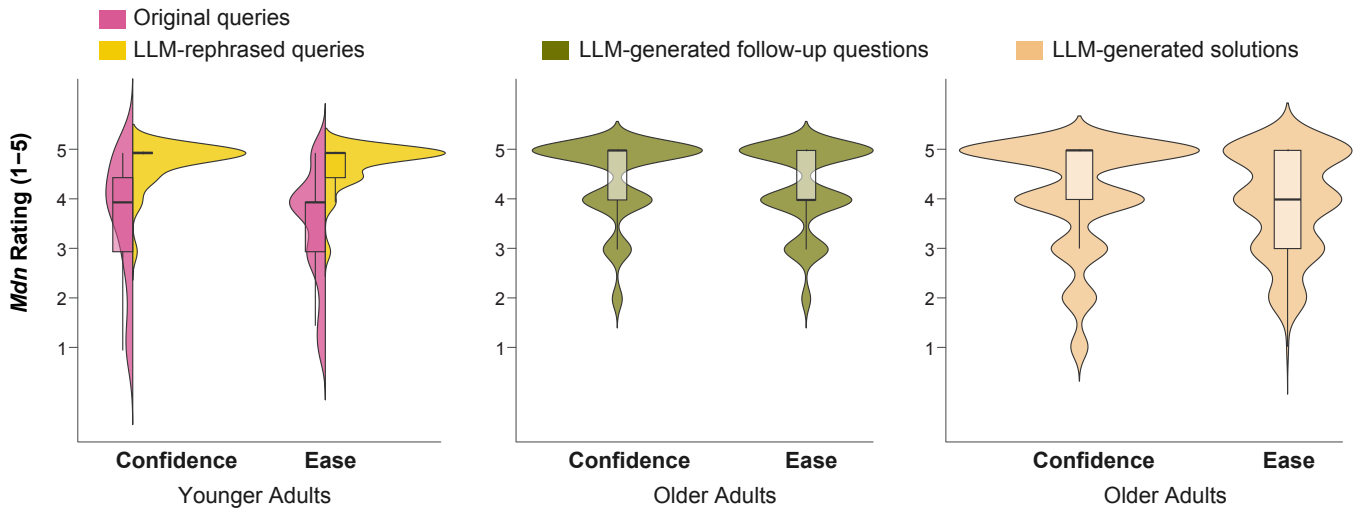


Figure 2: Younger adults reported significantly greater confidence and ease in understanding the LLM-rephrased queries (left). Older adults rated their confidence in responding to LLM-generated follow-up questions as high, but their ease of answering as moderate (center). Similarly, older adults reported high confidence in their ability to follow the solutions, but rated the ease of doing so as moderate (right).

then dynamically filtered so that participants only reviewed content relevant to the devices they reported using. Three follow-up and two solutions were not rated because none of our participants reported using the related mobile devices.

After completing eligibility screening and providing informed consent, participants were shown a randomized subset of follow-up questions and solutions. They then rated each item using the measures described above. All materials were administered online using the Qualtrics platform. At the end of the study, participants completed a brief demographic questionnaire and the 14-item Mobile Device Proficiency Questionnaire (MDPQ-14; range: 14 to 70 [53]).

6.2 Results

6.2.1 Participants. Thirty-four older adults (Mdn age = 72 years, IQR = 10.75, range = 63–91) participated in the study. Of those who provided demographic information, 19 identified as women and 7 as men. The majority of participants identified as White (n = 25). Educational attainment varied: seven participants reported holding a high school diploma, three had a 2-year degree, seven held a 3- or 4-year degree, seven had a master's degree, and three held a doctoral degree. In terms of living arrangements, 18 participants lived independently, four lived with family, and four resided in a senior living community. English proficiency was generally high, with 17 participants reporting full professional proficiency, eight reporting native or bilingual proficiency, five reporting professional working proficiency, and four reporting limited working proficiency. Self-reported mobile device proficiency was moderate to high (Mdn = 62.0, IQR = 2.15).

Participants rated 82 unique LLM-generated follow-up questions and 34 unique LLM-generated solutions. The number of follow-up questions rated per participant ranged from 1 to 20 (Mdn = 16, IQR = 3.75), with most participants rating between 14 and 18 items. For

the solution items, participants rated between 1 and 12 solutions (Mdn = 6, IQR = 5.0), with most providing responses for 6 to 10 solutions.

6.2.2 LLM-Generated Follow-up Questions. Participants reported high levels of perceived ability to respond to the LLM-generated follow-up questions. Across all ratings, 79.8% of responses indicated 'Yes' and 10.0% indicated 'Maybe' when asked if they could answer the question, suggesting strong engagement with some uncertainty. Confidence in their ability to respond was also high (Mdn = 5.0, IQR = 1.0), while ease of answering was rated as moderate (Mdn = 4.0, IQR = 1.0; Figure 2).

Spearman's rank-order correlation revealed that mobile device proficiency was positively associated with both confidence (ρ = .40, p < .001) and ease of answering (ρ = .56, p < .001). There was also a small but statistically significant correlation between digital proficiency and likelihood of responding 'Yes' to whether they could answer the question (ρ = .18, p = .003), indicating that participants with higher technology proficiency felt more capable of responding to the contextual questions.

6.2.3 LLM-Generated Technology Solutions. Regarding the LLM-generated solutions, 81.6% of responses indicated participants could follow the solution, and an additional 13.1% responded 'Maybe.' These results suggest that most participants found the solutions actionable, with a subset expressing partial confidence or uncertainty. Participants reported high confidence in their ability to follow the solutions (Mdn = 5.0, IQR = 1.0) and generally rated the ease of following the instructions as moderate (Mdn = 4.0, IQR = 2.0; Figure 2).

Spearman rank-order correlations were also conducted to examine whether digital proficiency and education level were associated with participants' confidence, ease, and perceived ability to follow proposed technology solutions. Digital proficiency showed strong,

statistically significant associations with all three outcomes: confidence ($\rho = .53, p < .001$), ease ($\rho = .47, p < .001$), and ability to follow the solution ($\rho = .44, p < .001$). In contrast, education level was moderately associated with confidence ($\rho = .34, p < .001$) and perceived ability to follow the solution ($\rho = .31, p < .001$), but not significantly related to ease.

In sum, our results show the promise of LLM-generated query reformulation and technology support in improving older adults' technology experiences. However, scalable development remains challenging. Most models are trained on data from younger, digitally fluent users, often overlooking the communication styles and cognitive needs of aging populations [68, 73]. To address this, next we created a synthetic dataset that reflects these overlooked characteristics.

7 STAQ: A Synthetic Technology Assistance Queries Dataset for Training Age-Inclusive AI Systems

Developing equitable AI systems that support older adults requires training data that reflect the realities of aging. Within HCI, there is a rightful consensus that empirical, human-centered data collection cannot be replaced. However, requiring older adults to participate in the exhaustive, high-volume data generation necessary to pre-train or fine-tune modern foundation models is both practically prohibitive and ethically burdensome. Consequently, most large-scale training datasets default to younger demographics, leaving older adults structurally underrepresented. We acknowledge that synthetic data should never replace participatory design or human evaluation; rather, it can serve as a necessary, scalable bridge for model training. By using large language models to systematically amplify empirically observed behaviors—rather than relying on stereotypical assumptions—researchers can simulate underrepresented communication barriers without extracting endless labor from vulnerable populations. To this end, this section introduces STAQ (Synthetic Technology Assistance Queries), pronounced “stack”: a synthetic training dataset grounded in the real-world, unstructured queries collected during our formative diary study (Section 3). Our dataset is publicly available in the Zenodo repository¹.

7.1 Dataset Development

We generated synthetic data using few-shot prompting with GPT-4o, guided by the communication characteristics of older adults' technology queries identified in Section 3: verbosity, over-specification, under-specification, and incompleteness. Each prompt included three example pairs—an original query exhibiting one of these characteristics and a corresponding expert-generated paraphrase (see Appendix C). To ensure diversity while maintaining realism, the model was instructed to vary the types of technical issues. To prevent the model from generating ageist stereotypes, we did not instruct the LLM to emulate a generic “older adult.” Instead, the generation was strictly anchored in the empirical data, requiring the model to mirror the specific conversational phrasing, expressions of uncertainty, and informal tone observed in our real-world dataset.

To examine linguistic characteristics and validate the realism of the generated text, we analyzed token length and lexical diversity. On average, the synthetic original queries contained $M = 58.89$ tokens, whereas the paraphrases averaged $M = 23.75$ tokens (range: 13 to 174 tokens). We also calculated the type-token ratio (TTR) to assess lexical repetition. In the synthetic dataset, queries written in the older adult style yielded a TTR of 0.858, compared to 0.937 for expert paraphrases. A similar pattern was observed in the real-world dataset (TTR = 0.842 for older adults vs. 0.890 for experts), demonstrating that the synthetic queries successfully mirrored authentic communication patterns.

7.2 Dataset Composition

The STAQ dataset comprises 514 instances, distributed across four categories: verbosity ($n = 120$), over-specification ($n = 120$), under-specification ($n = 154$), and incompleteness ($n = 120$). Each instance includes a technology-related query written in the communication style of older adults, paired with an LLM-generated paraphrase that clarifies the intended meaning. Illustrative examples are provided in Table 4.

7.3 Dataset Evaluation

We evaluated the STAQ dataset on two key dimensions: fidelity and face validity. Fidelity refers to the extent to which the synthetic data preserves the statistical properties and latent semantic patterns of the original dataset. To assess this, we compared vector representations of synthetic and real-world data using Principal Component Analysis (PCA) applied to paragraph-level embeddings generated by a Sentence Transformers (SBERT) model. Unlike standard BERT, which is optimized for token-level tasks, SBERT uses a Siamese network architecture to produce sentence-level embeddings that capture semantic similarity more effectively [51]. As shown in Figure 3, the PCA visualization demonstrated a high degree of overlap between the distributions, confirming that the synthetic queries successfully mirror the underlying semantic structure of the original data and indicating strong fidelity.

To assess face validity, we conducted a qualitative review with two domain experts in aging and technology. A random sample of 50 synthetic queries was independently evaluated by each expert. Reviewers classified each query based on how likely it was to have been communicated by an older adult, using the following categories: (1) likely, (2) possibly, or (3) unlikely. Of the 50 queries, 40 were rated as ‘likely,’ indicating strong plausibility (nine were rated ‘unlikely,’ one ‘possibly’). Inter-rater reliability was high (Cohen's $\kappa = 0.83$). These results support the face validity of the STAQ dataset as a realistic representation of how older adults communicate when seeking technology support.

8 Discussion

Despite AI's potential to enhance older adults' well-being and independence, it also risks reinforcing digital ageism [43]. This bias may widen the digital divide not just between generations, but also within older populations along lines of education, income, and access [20, 23]. Although device ownership is rising, many older adults still struggle with tasks beyond basic use [1, 13]. Mainstream

¹<https://doi.org/10.5281/zenodo.21323107>

Table 4: Example queries from the STAQ dataset and their characteristics.

Characteristic	Synthetic Query	AI-Rephrased Query
Verbosity	You know, ever since my granddaughter gave me this tablet thingie, I've been having a hard time trying to post photos on my Facebook. Every time I try to post a photo, it just gets stuck or something and then the tablet just freezes up. I can't even turn the darn thing off then. I try pressing the button thing to turn it off, but nothing happens. It used to be easier with my old computer. I've tried waiting it out and it still doesn't work right. Do you think you could help with that?	Why does my tablet freeze and become unresponsive when I try to post photos on Facebook, and how can I fix it?
Over-specification	I've been trying to connect to that zoom thing, you know, for the church meeting. But it's been quite a headache. I finally got the app on my computer, but now it's asking for a password or something. I just don't quite understand why or how to get that. Can you help explain?	How do I join a Zoom meeting and where do I find the password for the meeting?
Under-specification	I want to send pictures to my grandson but my Gmail is just not doing it.	How can I successfully send images through email using Gmail?
Incompleteness	Hey, my Facebook is not working like before. How do I make it normal?	How can I revert to the previous version or settings of my Facebook application as it is not behaving as expected currently?

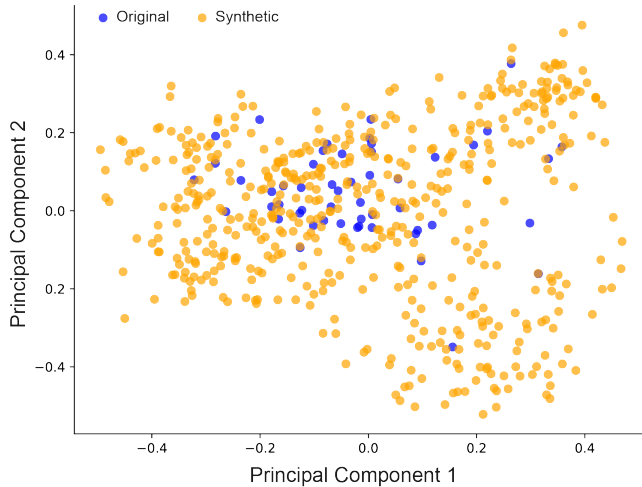


Figure 3: PCA visualization of sentence embeddings generated by Sentence Transformers (SBERT) for original (blue) and synthetic (orange) queries. The overlapping distributions indicate that the synthetic queries in the STAQ dataset capture semantic patterns similar to those in the original data.

AI systems, often designed for clear and concise inputs, may exclude users whose communication styles are different.

This work focuses on technology support, which lies at the intersection of growing need and increasing automation. While older adults require help with continuously evolving digital systems

[36], support is increasingly delegated to AI-based agents. These systems often assume users can engage with advanced tools at the very moment they experience confusion—risking what might be called algorithmic ageism [73].

As a result, technology support often falls to family and friends, who adopt caregiving roles long before physical support is needed [57]. This “technology caregiving” also presents challenges, especially when older adults struggle to articulate problems. Queries that are vague, overly detailed, or use unfamiliar terms can be hard to interpret. This study deepens our understanding of these barriers and introduces a age-friendly approach to make support systems more inclusive.

8.1 Primary Findings

Our study offers empirical insight into how older adults communicate technology problems in unstructured contexts. We documented how help-seeking queries are often incomplete, vague, or overly detailed (Table 1), directly conflicting with current AI tools that assume contextual information is already available [21, 68]. To address this, we developed an LLM-based approach that mimics a human helper: asking contextual questions, paraphrasing the query, and generating a solution. Using few-shot prompt chaining, we found that GPT-4o effectively reformulated unstructured queries, substantially increasing the retrieval of correct solutions compared to the original queries. Importantly, this LLM-mediated articulation benefited both sides of the support interaction. LLM-rephrased queries were more easily understood by younger adults, who often serve as informal technology helpers, improving their comprehension and confidence. Simultaneously, older adults evaluated the

LLM-generated solutions as easy to understand and follow. These findings highlight the potential for LLMs to reduce the communicative burden of help-seeking and support diverse aging populations.

The specific magnitudes we report—the solution-accuracy and comprehension gains from query rephrasing—are a snapshot of current models. As language models continue to improve, we anticipate that future models will handle unstructured input better and may supersede our specific workflow. Our contribution, however, is not the LLM pipeline *per se*, but the knowledge of how older adults articulate technology problems, where the breakdown happens, and how LLMs can mitigate it. Our experiments further demonstrate that mitigating these breakdowns improves technology support on both sides of the interaction—for the older adults receiving it and the technology helpers providing it.

8.2 Limitations

Our findings should be interpreted in light of several limitations. First, the older adult participants across our diary study and evaluation experiments comprised a relatively small, self-selected group of community-dwelling, English-speaking individuals. All possessed baseline experience with smart devices, meaning our sample lacked digital novices. Consequently, while their queries reflect common frustrations, these findings may not generalize to users with minimal technology exposure or significant cognitive impairments. Second, while our formative diary study captured naturalistic help-seeking, our subsequent evaluation experiments were conducted in controlled settings; rating static queries and solutions may not fully reflect the dynamic, conversational complexities of real-world support interactions. Finally, our approach relied on a state-of-the-art, proprietary large language model (GPT-4o). This performance may not be replicable with smaller or open-source models, potentially limiting scalability in resource-constrained environments. Furthermore, reliance on a proprietary API means that future changes to model behavior could affect output consistency and reliability.

8.3 Implications for Aging Research and Practice

This work advances aging research by advancing understanding of how older adults articulate technology problems—an often-overlooked aspect of aging with technology. While prior studies have largely focused on whether older adults adopt certain technologies or why they may resist them [14], our study shifts attention to what happens after adoption—when older adults encounter difficulties and seek help. The communication patterns we identified offer a vocabulary for describing common pitfalls in help-seeking, which can inform both research and practice.

Digital literacy programs, for instance, could incorporate modules that teach older adults how to formulate more effective help requests—such as how to emphasize the core issue and provide relevant details. However, it may be unrealistic to expect all older adults to adopt concise, technically precise language, particularly given age-related cognitive changes and rapidly changing technology. This reinforces the need for support systems that adapt to users, rather than forcing users to adapt to systems.

Our LLM-based approach offers a model for such age-friendly design. By beginning with real-world diary entries and designing around older adults’ actual communication styles—rather than defaulting to younger user norms—we treat older adults not as deficient, but as capable users with distinct interaction patterns. Our diary study revealed a strong reliance on text, with only a small fraction of queries including screenshots—and none utilizing screen or video recording options. This limited use of multimodal context may stem from the cognitive overhead required to capture and attach visual information, unfamiliarity or difficulty with those actions, or underlying privacy concerns. Future iterations of the pipeline could address this barrier by dynamically generating instructions during the follow-up phase, proactively guiding older adults through capturing and sharing screenshots when visual context is necessary.

AI-powered technology support could also extend the reach of existing technology education efforts. For example, community centers or libraries could deploy AI chatbots to provide just-in-time assistance for older patrons—especially those without digitally skilled family members or immediate access to help. These systems could foster autonomy, allowing older adults to solve problems on their own schedules rather than waiting for in-person assistance. While our evaluation provides a cross-sectional view of help-seeking, longitudinal deployments are necessary to understand how such a system impacts problem articulation over time. Observing repeated interactions with the system could reveal whether the tool exerts a pedagogical effect, gradually training older adults to independently formulate more structured and complete technology support queries.

Finally, our study emphasizes the diversity within the older adult population—across cognitive abilities, digital proficiency, and communication styles. Some individuals prefer narrative, detailed explanations; others are brief and to the point. Our LLM pipeline was able to handle both over- and under-specified queries, suggesting its potential to support a range of help-seeking behaviors. For aging services practitioners, this flexibility enables more tailored support. AI tools could triage requests by prompting more detail from minimal input or distilling key information from verbose descriptions. In both cases, the experience becomes more personalized.

8.4 Implications for AI Research and Practice

Our findings have broad implications for the future of AI research and practice, particularly in the development of human-centered systems that serve diverse populations. In addressing the technology support needs of older adults, our study highlights a persistent disconnect between users’ real-world help-seeking behaviors and the assumptions built into many current AI systems. Most AI models are trained on data from younger, digitally fluent populations, resulting in interaction norms that often misalign with older adults’ communication styles. These biases can lead to systems that misinterpret older users’ queries or produce responses that require digital proficiency many do not possess.

Our study challenges these assumptions by showing that large language models (LLMs) can be adapted to accommodate older adults’ natural language through targeted prompt engineering. By centering the real-world communication patterns of older adults,

we demonstrate that inclusive design can lead to more accessible and effective support experiences—not only for older users, but for everyone. Clarity, context sensitivity, and adaptability are not niche accommodations; they are universal design principles for equitable AI.

While this work focuses on characterizing and supporting older adults' problem articulation, future research should explore model-centric applications for STAQ. To our knowledge, it is the first structured dataset focused specifically on this form of help-seeking behavior. STAQ can serve as an auditing benchmark to evaluate whether commercial LLMs exhibit systematic performance drops on older adult-style queries compared to standard technical support corpora. Comparing performance on STAQ against standard, crowd-sourced data (which typically skew younger) can reveal algorithmic disparities and directly inform the refinement of inclusive models. However, deploying LLM-assisted pipelines in real-world technology caregiving contexts necessitates a critical evaluation of privacy and trust. Processing potentially sensitive, personal technology queries through cloud-based APIs introduces privacy concerns for both older adults and their helpers, underscoring the need to explore secure, localized, or on-device model processing in future system designs.

This work also surfaces key methodological insights for the AI research community. First, we demonstrate that synthetic datasets can successfully capture users' lived experiences when grounded in empirical, user-centered data, offering a scalable path toward inclusive design. However, as prompt-based data generation becomes increasingly popular, it inherently risks producing stereotypical or hallucinated outputs if not anchored to real user needs. We mitigated this by strictly integrating domain-informed and user-centered strategies into our generation pipeline. As synthetic generation expands, researchers must continue to balance computational scalability with rigorous, human-centered evaluation frameworks to ensure AI tools reflect true real-world diversity rather than algorithmic assumptions.

9 Conclusion

When seeking support to resolve technology problems, older adults typically either search for answers online or turn to a trusted person in their social network. In either case, they must articulate the problem effectively enough to reach a resolution. In this paper, we identified the communication challenges older adults face when describing technology problems and examined how these challenges affect both the quality of support they receive and their ability to find solutions independently (e.g., via a Google search).

To address these challenges, we developed an LLM pipeline that reformulates unstructured queries into clearer, more actionable requests and generates targeted solutions—reducing the cognitive burden on technology helpers and improving the likelihood of successful independent resolution through online resources. To support further research in this space, we introduced STAQ, a synthetic dataset of older adults' technology problems designed to enable more realistic evaluation of AI-assisted support systems.

Ultimately, this work emphasizes that older adults are not deficient users, but communicators with distinct interaction styles that current AI systems often overlook. Embedding these authentic

styles into AI training, evaluation, and deployment processes is essential for building equitable, age-inclusive technologies.

Acknowledgments

This work was supported in part by the National Institute on Aging of the National Institutes of Health under Award No. P30AG083255 to DC. We thank our study participants for their time and Tasneem Mubashshira for her assistance with data analysis.

Ethical Considerations

All study procedures were approved by the university's Institutional Review Board (IRB). A secure version of ChatGPT [3] was used to assist with language editing and article formatting. No AI-generated content was included without human review, revision, and approval. The authors retain full responsibility for the accuracy and integrity of the manuscript.

References

- [1] AARP Research. 2026. *2026 Technology Trends and Older Adults*. Technical Report. AARP. doi:10.26419/res.01020.001
- [2] Martha Abshire, Victor D Dinglas, Maan Isabella A Cajita, Michelle N Eakin, Dale M Needham, and Cheryl Dennison Himmelfarb. 2017. Participant retention practices in longitudinal clinical research studies with high retention rates. *BMC medical research methodology* 17, 1 (2017), 30.
- [3] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altmenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).
- [4] Lan An, Christel Huwald, Diego Muñoz, Sonja Pedell, and Leon Sterling. 2023. Weisst Du wieviel Sternlein stehen?—Building older adults' confidence in technology use through co-designing digital storytelling. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–7.
- [5] Lan An, Diego Muñoz, Sonja Pedell, and Leon Sterling. 2022. Understanding confidence of older adults for embracing mobile technologies. In *Proceedings of the 34th Australian Conference on Human-Computer Interaction*. 38–50.
- [6] Yvonne Barnard, Mike D Bradley, Frances Hodgson, and Ashley D Lloyd. 2013. Learning to use new technologies by older adults: Perceived difficulties, experimentation behaviour and usability. *Computers in human behavior* 29, 4 (2013), 1715–1724.
- [7] Serkan Bayrakcı and Haldun Narmanlioğlu. 2021. Digital Literacy as Whole of Digital Competences: Scale Development Study. *Düşünce ve Toplum Sosyal Bilimler Dergisi* 4 (2021), 1–30. <https://izlik.org/JA43HH78MJ>
- [8] Eva Brandt, Thomas Binder, Lone Malmberg, and Tomas Sokoler. 2010. Communities of everyday practice and situated elderliness as an approach to co-design for senior interaction. In *Proceedings of the 22nd Conference of the Computer-Human Interaction Special Interest Group of Australia on Computer-Human Interaction*. 400–403.
- [9] Virginia Braun and Victoria Clarke. 2021. One size fits all? What counts as quality practice in (reflexive) thematic analysis? *Qualitative research in psychology* 18, 3 (2021), 328–352.
- [10] Janet Bryan, Mary A Luszcz, and John Robertson Crawford. 1997. Verbal knowledge and speed of information processing as mediators of age differences in verbal fluency performance among older adults. *Psychology and aging* 12, 3 (1997), 473.
- [11] Deborah M Burke, Donald G MacKay, Joanna S Worthley, and Elizabeth Wade. 1991. On the tip of the tongue: What causes word finding failures in young and older adults? *Journal of memory and language* 30, 5 (1991), 542–579.
- [12] Adeline Chu and Beth Mastel-Smith. 2010. The outcomes of anxiety, confidence, and self-efficacy with Internet health information retrieval in older adults: a pilot study. *CIN: Computers, Informatics, Nursing* 28, 4 (2010), 222–228.
- [13] Sara J. Czaja. 2019. Usability of technology for older adults: where are we and where do we need to be. *J. Usability Studies* 14, 2 (Feb. 2019), 61–64.
- [14] Sara J Czaja, Neil Charness, Arthur D Fisk, Christopher Hertzog, Sankaran N Nair, Wendy A Rogers, and Joseph Sharit. 2006. Factors predicting the use of technology: findings from the Center for Research and Education on Aging and Technology Enhancement (CREATE). *Psychology and aging* 21, 2 (2006), 333.
- [15] Karen Davies, Andrew Kingston, Louise Robinson, Joan Hughes, Judith M Hunt, Sally AH Barker, June Edwards, Joanna Collerton, Carol Jagger, and Thomas BL Kirkwood. 2014. Improving retention of very old participants in longitudinal

- research: experiences from the Newcastle 85+ study. *PLoS One* 9, 10 (2014), e108370.
- [16] Eleni Demetriou and Roe Holtzer. 2017. Mild cognitive impairments moderate the effect of time on verbal fluency performance. *Journal of the International Neuropsychological Society* 23, 1 (2017), 44–55.
 - [17] Glen H Elder Jr. 1994. Time, human agency, and social change: Perspectives on the life course. *Social psychology quarterly* (1994), 4–15.
 - [18] Mingming Fan, Qiwen Zhao, and Vinita Tibdewal. 2021. Older adults' think-aloud verbalizations and speech features for identifying user experience problems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13.
 - [19] Kairong Fang, Jiesi Zhang, Shi-Ting Ni, Pan Hui, and Yuyang Wang. 2026. GuideMe: A VLM-Based System Assisting Independent Smartphone Learning for Older Adults. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*. 1–19.
 - [20] Thomas N Friemel. 2016. The digital divide has grown old: Determinants of a digital divide among seniors. *New media & society* 18, 2 (2016), 313–331.
 - [21] Weiwei Gao, Kexin Du, Yujia Luo, Weinan Shi, Chun Yu, and Yuanchun Shi. 2024. Easyask: An in-app contextual tutorial search assistant for older adults with voice and touch inputs. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 3 (2024), 1–27.
 - [22] Nelly Geerts, Werner Schirmer, Anina Vercruyssen, and Ignace Glorieux. 2024. Exploring older adults' ICT support: A mismatch between needs and provision. *New Media & Society* 26, 12 (2024), 7314–7330.
 - [23] Nancy M Gell, Dori E Rosenberg, George Demiris, Andrea Z LaCroix, and Kushang V Patel. 2015. Patterns of technology use among older adults with and without disabilities. *The Gerontologist* 55, 3 (2015), 412–421.
 - [24] D Kristen Gilbert and Wendy A Rogers. 1999. Age-related differences in the acquisition, utilization, and extension of a spatial mental model. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences* 54, 4 (1999), P246–P255.
 - [25] Jean K Gordon, Megan Young, and Carly Garcia. 2018. Why do older adults have difficulty with semantic fluency? *Aging, Neuropsychology, and Cognition* 25, 6 (2018), 803–828.
 - [26] Elizabeth D Hutchison. 2010. A life course perspective. *Dimensions of human behavior: The changing life course* 4 (2010), 1–38.
 - [27] Yixin Jiang, Manushree Singhania, Teerapaun Tanprasert, and Joanna McGrenere. 2026. When To Help? Toward Non-Intrusive Struggle Detection to Support Older Adults Learning New Computer Tasks. In *Proceedings of the Extended Abstracts of the 2026 CHI Conference on Human Factors in Computing Systems*. 1–7.
 - [28] Anjali Khurana, Xiaotian Su, April Yi Wang, and Parmit K Chilana. 2025. Do it for me vs. do it with me: Investigating user perceptions of different paradigms of automation in copilots for feature-rich software. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–18.
 - [29] Anjali Khurana, Hariharan Subramanyam, and Parmit K Chilana. 2024. Why and when llm-based assistants can go wrong: Investigating the effectiveness of prompt-based interactions for software help-seeking. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*. 288–303.
 - [30] Sunyoung Kim, Krzysztof Z Gajos, Michael Muller, and Barbara J Grosz. 2016. Acceptance of mobile technology by older adults: a preliminary study. In *Proceedings of the 18th international conference on human-computer interaction with mobile devices and services*. 147–157.
 - [31] Soomin Kim, Hyuna Jang, Sou Jin Choi, Hyung-Ji Kim, Jae-Hong Lee, and Miseon Kwon. 2021. Quantitative and qualitative differences of action verbal fluency between young and older adults. *Dementia and geriatric cognitive disorders* 50, 6 (2021), 585–591.
 - [32] Sunyoung Kim, Willow Yao, and Xiaotong Du. 2022. Exploring older adults' adoption and use of a tablet computer during COVID-19: longitudinal qualitative study. *JMIR aging* 5, 1 (2022), e32957.
 - [33] Bran Knowles and Vicki L Hanson. 2018. The wisdom of older technology (non) users. *Commun. ACM* 61, 3 (2018), 72–77.
 - [34] Bran Knowles, Vicki L Hanson, Yvonne Rogers, Anne Marie Piper, Jenny Waycott, Nigel Davies, Aloha Hufana Ambe, Robin N Brewer, Debaleena Chattopadhyay, Marianne Dee, et al. 2021. The harm in conflating aging with accessibility. *Commun. ACM* 64, 7 (2021), 66–71.
 - [35] Daniël Lakens. 2017. Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social psychological and personality science* 8, 4 (2017), 355–362.
 - [36] Amanda Lazar, Elissa Carpio, Ruipu Hu, Beth Barnett, Kibron Tesfatsion, and Sheena Erete. 2026. Interrogating the “Us” Versus “Them” Dichotomy in Technology Research with Older Adults. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*. 1–15.
 - [37] Terry Levitt, Jonathan Fugelsang, and Margaret Crossley. 2006. Processing speed, attentional capacity, and age-related memory change. *Experimental Aging Research* 32, 3 (2006), 263–295.
 - [38] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*. 74–81.
 - [39] Chaochao Lu, Chen Qian, Guodong Zheng, Hongxing Fan, Hongzhi Gao, Jie Zhang, Jing Shao, Jingyi Deng, Jinlan Fu, Kexin Huang, et al. 2024. From gpt-4 to gemini and beyond: Assessing the landscape of mlms on generalizability, trustworthiness and causality through four modalities. *arXiv preprint arXiv:2401.15071* (2024).
 - [40] Tamir Mendel, Debin Gao, David Lo, and Eran Toch. 2021. An exploratory study of social support systems to help older adults in managing mobile safety. In *Proceedings of the 23rd International Conference on Mobile Human-Computer Interaction*. 1–13.
 - [41] George Mois and Wendy A Rogers. 2024. Developing a Framework for Digital Activities of Daily Living. *The Gerontologist* 64, 10 (2024), gnae110. doi:10.1093/geront/gnae110
 - [42] Benjamin Morrison, Lynne Coventry, and Pam Briggs. 2021. How do Older Adults feel about engaging with Cyber-Security? *Human behavior and emerging technologies* 3, 5 (2021), 1033–1049.
 - [43] Louis Neven and Alexander Peine. 2017. From triple win to triple sin: How a problematic future discourse is shaping the way people age with technology. *Societies* 7, 3 (2017), 26.
 - [44] Susan Nittrouer, Joanna H Lowenstein, Taylor Wucinich, and Aaron C Moberly. 2016. Verbal working memory in older adults: The roles of phonological capacities and processing speed. *Journal of Speech, Language, and Hearing Research* 59, 6 (2016), 1520–1532.
 - [45] Tobias Olsson and Dino Viscovi. 2018. Warm experts for elderly users: Who are they and what do they do? *Human Technology* 14, 3 (2018). Publisher: University of Jyväskylä.
 - [46] Carolyn Pang, Zhiqin Collin Wang, Joanna McGrenere, Rock Leung, Jiamin Dai, and Karyn Moffatt. 2021. Technology adoption and learning preferences for older adults: evolving perceptions, ongoing challenges, and emerging design opportunities. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–13.
 - [47] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 311–318.
 - [48] Denise C Park, Gary Lautenschlager, Trey Hedden, Natalie S Davidson, Anderson D Smith, and Pamela K Smith. 2002. Models of visuospatial and verbal memory across the adult life span. *Psychology and aging* 17, 2 (2002), 299.
 - [49] Anne Marie Piper, Raymundo Cornejo Garcia, and Robin N Brewer. 2016. Understanding the challenges and opportunities of smart mobile devices among the oldest old. *International Journal of Mobile Human Computer Interaction (IJMHCI)* 8, 2 (2016), 83–98.
 - [50] Yujia Qin, Shengding Hu, Yankai Lin, Weize Chen, Ning Ding, Ganqu Cui, Zheni Zeng, Xuanhe Zhou, Yufei Huang, Chaojun Xiao, et al. 2024. Tool Learning with Foundation Models. *ACM Comput. Surv.* 57, 4, Article 101 (Dec. 2024), 40 pages. doi:10.1145/3704435
 - [51] Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084* (2019).
 - [52] Karen Renaud and Judy Van Biljon. 2008. Predicting technology acceptance and adoption by the elderly: a qualitative study. In *Proceedings of the 2008 annual research conference of the South African Institute of Computer Scientists and Information Technologists on IT research in developing countries: riding the wave of technology*. 210–219.
 - [53] Nelson A Roque and Walter R Boot. 2018. A new tool for assessing mobile device proficiency in older adults: the mobile device proficiency questionnaire. *Journal of Applied Gerontology* 37, 2 (2018), 131–156.
 - [54] Nina Sakhnini and Debaleena Chattopadhyay. 2022. A review of smartphone fact-checking apps and their (non) use among older adults. In *adjunct publication of the 24th international conference on human-computer interaction with mobile devices and services*. 1–8.
 - [55] Timothy A Salthouse. 1996. The processing-speed theory of adult age differences in cognition. *Psychological review* 103, 3 (1996), 403.
 - [56] Timothy A Salthouse and Renee L Babcock. 1991. Decomposing adult age differences in working memory. *Developmental psychology* 27, 5 (1991), 763.
 - [57] Hasti Sharifi and Debaleena Chattopadhyay. 2023. A cross-cultural study of relational maintenance in tech caregiving. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–10.
 - [58] Hasti Sharifi and Debaleena Chattopadhyay. 2023. Senior technology learning preferences model for mobile technology. *Proceedings of the ACM on Human-Computer Interaction* 7, MHCI (2023), 1–39.
 - [59] Hasti Sharifi, Joseph E Michaelis, and Debaleena Chattopadhyay. 2024. Development and Evaluation of the Mobile Tech Support Questionnaire for Older Adults. In *Proceedings of the 26th International ACM SIGACCESS Conference on Computers and Accessibility* (St. John's, NL, Canada) (ASSETS '24). Association for Computing Machinery, New York, NY, USA, Article 74, 18 pages. doi:10.1145/3663548.3675661
 - [60] Hasti Sharifi, Homaira Huda Shomee, Sourav Medya, and Debaleena Chattopadhyay. 2025. How older adults communicate their technology problems: Challenges and design opportunities. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–13.

- [61] Lingfeng Shen, Lema Liu, Haiyun Jiang, and Shuming Shi. 2022. On the evaluation metrics for paraphrase generation. *Proceedings of the 2022 conference on empirical methods in natural language processing (2022)*, 3178–3190.
- [62] Hanchen Su, Wei Luo, Yashar Mehdad, Wei Han, Elaine Liu, Wayne Zhang, Mia Zhao, and Joy Zhang. 2025. LLM-friendly knowledge representation for customer support. In *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*. 496–504.
- [63] John Sweller. 1988. Cognitive load during problem solving: Effects on learning. *Cognitive science* 12, 2 (1988), 257–285.
- [64] Teerapaun Tanprasert, Jiamin Dai, and Joanna McGrenere. 2024. Helpcall: Designing informal technology assistance for older adults via videoconferencing. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–23.
- [65] Konstantinos Vadikolias, Anna Tsiakiri-Vatamidis, Grigorios Tripsianis, Georgios Tsigoulis, Panagiotis Ioannidis, Aspasia Serdari, John Heliopoulos, Miltos Livaditis, and Charitomeni Piperidou. 2012. Mild cognitive impairment: effect of education on the verbal and nonverbal tasks performance decline. *Brain and behavior* 2, 5 (2012), 620–627.
- [66] Viswanath Venkatesh, James YL Thong, and Xin Xu. 2012. Consumer acceptance and use of information technology: Extending the Unified Theory of Acceptance and Use of Technology1. *MIS quarterly* 36, 1 (2012), 157–178.
- [67] John Vines, Gary Pritchard, Peter Wright, Patrick Olivier, and Katie Brittain. 2015. An age-old problem: Examining the discourses of ageing in HCI and strategies for future research. *ACM Transactions on Computer-Human Interaction (TOCHI)* 22, 1 (2015), 1–27.
- [68] Bryan Wang, Gang Li, and Yang Li. 2023. Enabling conversational interaction with mobile ui using large language models. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–17.
- [69] Haoxin Wang, Xianhan Peng, Huang Cheng, Yizhe Huang, Ming Gong, Chenghan Yang, Yang Liu, and Jiang Lin. 2025. ECom-Bench: Can LLM Agent Resolve Real-World E-commerce Customer Support Issues?. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*. 276–284.
- [70] Shengzhi Wang, Khalisa Bolling, Wenlin Mao, Jennifer Reichstadt, Dilip Jeste, Ho-Cheol Kim, and Camille Nebeker. 2019. Technology to support aging in place: Older adults' perspectives. In *Healthcare*, Vol. 7. MDPI, 60.
- [71] Leigh Ann Wilson, Kate O'Loughlin, Robyn Hector, and Deborah Ann Black. 2021. Using a smartphone application (App) to assist older adults to navigate their local area during extreme weather events and changing environmental conditions: A qualitative study. *Australasian Journal on Ageing* 40, 1 (2021), e95–e99.
- [72] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771* (2019).
- [73] Zhiyong Wu, Zhenyu Wu, Fangzhi Xu, Yian Wang, Qiushi Sun, Chengyou Jia, Kanzhi Cheng, Zichen Ding, Liheng Chen, Paul Pu Liang, et al. 2024. Os-atlas: A foundation action model for generalist gui agents. *arXiv preprint arXiv:2410.23218* (2024).
- [74] Ja Eun Yu and Debaleena Chattopadhyay. 2020. “Maps are hard for me”: identifying how older adults struggle with mobile maps. In *Proceedings of the 22nd International ACM SIGACCESS Conference on Computers and Accessibility*. 1–8.
- [75] Ja Eun Yu and Debaleena Chattopadhyay. 2024. Reducing the Search Space on demand helps Older Adults find Mobile UI Features quickly, on par with Younger Adults. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–22.
- [76] Ja Eun Yu, Natalie Parde, and Debaleena Chattopadhyay. 2023. “Where is history”: Toward designing a voice assistant to help older adults locate interface features quickly. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–19.
- [77] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675* (2019).

A LLM-Generated Paraphrasing and Solutions

In this section, we provide detailed supplementary material for the prompts and examples used to generate contextual questions, paraphrased queries, and solutions.

A.1 Prompts used with GPT-4o

Figures 4, 5, and 6 show the prompts corresponding to contextual question generation, query paraphrasing, and solution generation, respectively.

A.2 Examples of LLM-Generated Paraphrasing

Figures 7 and 8 illustrate cases where the model produced correct and incorrect paraphrases, respectively.

A.3 Examples of LLM-Generated Solutions

Figures 9 and 10 show correct and incorrect solution outputs.

A.4 Quantitative Evaluation of Model Outputs

To complement our qualitative analyses, we conducted a quantitative evaluation of model-generated paraphrases and solutions using established natural language processing (NLP) metrics. Our goal was to assess how closely the LLM-generated content aligned with expert-written references, especially in the context of older adults' help-seeking behaviors. The following measures were used.

BLEU [47] is a precision-based metric originally developed for evaluating machine translation. It measures how many words or short sequences of words (n-grams) in the model output match those in a reference text. Higher scores indicate greater similarity to the reference. In this study, BLEU was used to assess how well LLM-generated paraphrases and solutions matched expert-written versions.

ROUGE-L [38] measures the longest common subsequence between the model output and a reference text, capturing both precision and recall. It is often used to evaluate the quality of summaries. Here, it was used to assess whether model-generated solutions preserved the key steps and sequence of expert solutions.

BERTScore [77] compares the contextual embeddings of words in the model output and reference text using a pretrained language model (BERT). It goes beyond exact word matches to assess semantic similarity. Higher scores suggest closer alignment in meaning, even if different wording is used.

Cosine similarity measures the angle between two vectors—in this case, the semantic representations of the model output and reference text. It ranges from -1 to 1 , where 1 indicates identical semantic orientation. This metric was used to quantify how similar the model-generated solutions were to the expert-written solutions at a conceptual level.

ParaScorer [61] is a paraphrase evaluation framework that combines reference-based and reference-free metrics. Reference-based metrics compare model output to an expert-written paraphrase, while reference-free metrics assess semantic similarity and fluency without relying on a reference. This tool helps evaluate how well LLM systems reformulate user queries for improved clarity and problem-solving effectiveness.

Paraphrasing quality was evaluated using ParaScorer [61], which provides both reference-based and reference-free metrics. Reference-based metrics assess the degree of similarity between model outputs and expert-generated paraphrases, while reference-free metrics evaluate semantic similarity and fluency without a predefined reference. GPT-4o outperformed OS-ATLAS on reference-based evaluation (0.95 vs. 0.85), indicating stronger alignment with expert rewordings. In contrast, OS-ATLAS performed better on reference-free evaluation (0.70 vs. 0.60). However, paraphrasing older adults' queries often requires adding contextual information not present in the original message, which increases lexical divergence and may penalize models in reference-free evaluation. For our use case—clarifying real-world, underspecified queries—reference-based alignment was a more relevant indicator of effectiveness. Accordingly, GPT-4o's performance suggests greater suitability for supporting older adults through LLM-powered rewording of technology queries.

To further assess paraphrasing quality, we compared model outputs to expert-generated paraphrases using widely accepted natural language processing metrics. These included contextual similarity [77], machine translation quality [47], summarization accuracy [38], and cosine-based semantic similarity. As shown in Table 5, GPT-4o consistently outperformed OS-ATLAS across all metrics. We applied these same metrics to evaluate model-generated solutions against expert-written responses (Table 6). GPT-4o again demonstrated stronger alignment, achieving a BLEU score of 0.62 and a ROUGE-L score of 0.72, compared to OS-ATLAS scores of 0.04 and 0.30, respectively. Because BLEU and ROUGE measure n-gram overlap and sequence fidelity, they are particularly well suited for evaluating multi-step technology support instructions—where both accuracy and clarity of sequence are essential for usability. This is especially important for older users, who often benefit from step-by-step, easy-to-follow guidance when resolving technology issues.

Table 5: Quantitative evaluation of query paraphrasing.

Model	BERTScore	BLEU	ROUGE-L	Cosine
GPT-4o	.98	.75	.87	.88
OS-ATLAS	.92	.09	.40	.36

Table 6: Quantitative evaluation of solutions generated.

Model	BERTScore	BLEU	ROUGE-L	Cosine
GPT-4o	.94	.62	.72	.79
OS-ATLAS	.87	.04	.30	.38

B Study Instruments

How much do you agree or disagree with the following statements? (5-point scale: 1 = Strongly Disagree, 5 = Strongly Agree)

- I know how to protect others and my own personal data (e.g., photos, addresses, family information) online.

An older adult has described the following problem:
{Original Query}
 Your task is to assist them in solving their issue. For that, the first step is:
 Determine if the problem description contains all the necessary information to resolve the issue. If all information is provided: Print that you are ready to proceed to the next step. If additional information is needed: Ask up to three specific and concise follow-up questions to clarify any missing details. Only ask questions that are essential to solving the problem. Avoid unnecessary or overly technical questions. You don't need to ask 3 questions, you can ask less. Provide only the questions without additional text.

Example 1:
 User Query: "On my desktop always have the option of reverting to a prior date's settings or configurations. Is there a similar feature for Android phones?"
 Model Response: No questions, I can solve it.

Example 2:
 User Query: "Argh.. I have had this problem for a while and don't even know how or why it happened. I want to watch my wolf utubes with my phone or iPad in the horizontal position. All of a sudden it won't turn when I turn my phone. What happened?!! How do I fix it?!!"
 Model Response: 1. What type of phone do you have? (For example, iPhone or Android) 2. Have you noticed this happening with applications other than YouTube?

Example 3:
 User Query: "It seems that "System Apps" can't be deleted, but can be stripped of updates and remain functioning. Is this my insurance against deletion?"
 Model Response: 1. What is the device type? (For example, Laptop, Tablet, or Phone) 2. Do you know if it is an Apple product or not?

--End of examples--

Figure 4: The prompt used for generating contextual questions.

- I can verify online information using multiple sources.
- I understand what hardware and software technologies are.
- I can install software on my computer or other electronic devices.
- I understand the terms "Internet" and "World Wide Web (WWW)."
- I can effectively use e-government applications, such as on-line tax filing, driver's license applications, and voter registration.
- I can use cloud computing technologies (Google Drive, iCloud, Dropbox, etc.) effectively in daily life.
- I can use the calendar on mobile devices not only just to look at dates but also to check dates, set reminders, and create events.
- I can upload and broadcast videos online.
- I can use digital technologies effectively in daily practice such as reservation, shopping, address finding, etc.
- I can add a web page that I use to bookmarks or favorites.
- I know how to restrict apps' access to my personal information (location, contacts, camera, etc.)
- I can recognize and block unwanted / spam emails and phishing messages.
- I can change the privacy/security settings on my social media posts and profile.
- I know how to create a strong and secure password.
- I can effectively use at least one software program related to my field (e.g., Photoshop, SPSS, Premiere, Microsoft Word).

The final digital literacy score was computed using the mean of all ratings [7] (Bayrakci et al., 2021).

C STAQ

C.1 Prompt used with GPT-4o

Figure 11 shows the prompt used to generate synthetic data with few-shot examples.

Now, I will send you the user's answers to the questions you asked in the previous turn:
{Answers to the follow-up questions, if available}

After reviewing the answers to the follow-up questions, rephrase the question in a way that is clear enough for the user to copy and paste into a Google Search Bar and be able to find the answer. If you think there are multiple questions with different solutions being asked, separately rephrase them. If you are less than 90% confident in your understanding of the user's intent and problem, just say, 'I don't know.' Provide only the rephrased questions without additional text.

Example 1:

Technology question: On my desktop always have the option of reverting to a prior date's settings or configurations. Is there a similar feature for Android phones?

Model Response: No questions, I can solve it. Can I revert back my settings and configurations to a prior date on an Android device?

Example 2:

Technology question: Argh.. I have had this problem for a while and don't even know how or why it happened. I want to watch my wolf utubes with my phone or iPad in the horizontal position. All of a sudden it won't turn when I turn my phone. What happened?!! How do I fix it?!!

Model Response: 1. What type of phone do you have? (For example, iPhone or Android)

2. Have you noticed this happening with applications other than YouTube?

User: 1. iPhone 2. I don't know

Model Response: How to turn off Portrait Orientation Lock on an iPhone?

Example 3:

Technology question: "It seems that 'System Apps' can't be deleted, but can be stripped of updates and remain functioning. Is this my insurance against deletion?"

Model Response: 1. What is the device type? (For example, Laptop, Tablet, or Phone) 2. Do you know if it is an Apple product or not?

User: It's a Samsung phone."

Model Response: Can I delete system apps on an Android phone accidentally?

--End of examples--

Figure 5: The prompt used for paraphrasing a technology query.

{Paraphrased Question generated by the model}

Now, provide step-by-step instructions on how to solve the tech issue(s) in a language that is understandable for older adults. Keep each step short, and the instructions concise. If you are less than 90% confident that the solution would work for the user, just say, 'I don't know.' Providing incorrect solutions will frustrate our users. If the user's request is impossible, say, 'You cannot do that' with a brief explanation of why. If no action from the user is required, say, 'No action needed' with a brief explanation of why. Do not offer alternative solutions. Only give the one solution you are most confident in. Do not ask any more questions from the user.

Example 1:

Paraphrased question: Can I revert back my settings and configurations to a prior date on an Android device?

Model response: Unfortunately, that is not possible on an Android phone.

Example 2:

Paraphrased question: How to turn off Portrait Orientation Lock on an iPhone?

Model response: Here is step-by-step instructions:

1. Swipe down from the top-right corner of your screen to open Control Center.
2. Tap the button that has a lock inside and make sure that it's off.
3. Go back to your video and turn your iPhone sideways.

Example 3:

Paraphrased question: Can I delete system apps on an Android phone accidentally?

Model response: You can't delete system apps that came pre-installed on your Android phone.

--End of examples--

Figure 6: The prompt used for generating a technology solution.

User: When I go into Wi-Fi on Settings on my iPad, it states 'Privacy Warning' under the network I'm connected to. It doesn't state that on my iPhone though. Do you know what that means and if I need to do something about it?

Model: Is the Wi-Fi network you're connected to at home, work, or a public place?

User: Home

Model: Why does my home Wi-Fi network show a "Privacy Warning" on my iPad but not on my iPhone?

Figure 7: Example of a correct paraphrasing.

User: Occasionally when working on my laptop the screen I'm working on will just slide off the right. And I'm left with just my desktop. If it was a document I was working then I know that I can go to the Microsoft Word icon on my toolbar, click it, and the homescreen will "shift" again and my document will reappear. But I don't know what I'm doing that would have caused the document to slide off screen to begin with. This doesn't happen frequently, just occasionally as I said. And frustrates me when it does.

Model: 1. What type of laptop do you have? (For example, Dell, HP, or Lenovo)

2. What operating system is installed on your laptop? (For example, Windows, macOS, or Linux)

3. Are you using a specific application that causes the screen to slide off the right?

User: 1. MacBook 2. MacOS Ventura 3. Word, but it can happen with other apps too.

Model: Why does my laptop screen slide off the right occasionally and how can I fix it?

Correct paraphrasing: Why does my current window slide off-screen on macOS Ventura, and how do I get it back?

Figure 8: Example of an incorrect paraphrasing.

Paraphrased query: How do I make everything on my iPhone, including app icons and text, appear larger?

Model: Here are step-by-step instructions:

1. Open the Settings app on your iPhone.
 2. Scroll down and tap Display & Brightness.
 3. Tap Text Size to make text larger. Use the slider to adjust the size.
 4. If you need even larger text, go back and tap Accessibility.
 5. Select Display & Text Size and then tap Larger Text.
 6. Turn on Larger Accessibility Sizes and move the slider to make everything bigger.
 7. To make icons and buttons larger, go back to Accessibility and tap Zoom.
 8. Turn on Zoom and follow the on-screen instructions to zoom in and out.
- This should make apps, text, and icons easier to see.

Figure 9: Example of a correct technology solution.

Paraphrased query: How do I cancel a scheduled Lyft ride?

Model: Here are step-by-step instructions:

1. Open the Lyft app on your phone.
 2. Tap on the menu icon (three lines) in the top-left corner.
 3. Select "Ride History" from the menu.
 4. Find the scheduled ride you want to cancel.
 5. Tap on the ride, and then tap "Cancel Ride."
 6. Confirm your cancellation when prompted.
- That's it! Your ride should now be canceled.

Correct solution: Here are step-by-step instructions:

1. Open the Lyft app
2. Tap "Trips" on the bottom of the screen.
3. Locate the trip you want to cancel by scrolling through the trips if necessary.
4. Next to the trip details, find the "Cancel Trip" button.
5. Tap on "Cancel Trip."
6. Confirm your cancellation if prompted.

Figure 10: Example of an incorrect technology solution.

Several older adults have described the following tech problems in their own words. Below each example, an expert paraphrase of their true intention is provided:

Example1:

My iPad is not loading websites. It keeps stalling. I am connected to WiFi. I turned it off and back on. Still not loading

Paraphrased Question (Expert Interpretation): Why are websites not loading on both my iPhone and iPad even though they are both connected to my home WiFi?

Example2:

It appears that Netflix is an undeletable file. Can that be???

Paraphrased Question (Expert Interpretation): How do I uninstall or disable the Netflix app on a Samsung phone?

Example3:

I have added a credit card to Apple Pay but I don't know how to open Apple Pay when I want to

Paraphrased Question (Expert Interpretation): How do I open Apple Pay on an iPhone with Touch ID?

Your task is to generate one synthetic example similar to these.

Provide both: The original question in the casual, conversational style typical of older adults and A paraphrased expert interpretation that clarifies the true intent of the question.

Do not mention that it is synthetic.

Ensure diversity in the type of tech problem described. It does not need to exactly follow the issues above.

Follow the sentence structure and tone that older adults typically use, including hesitations, casual phrasing, and a conversational style. Output the result in JSON format with the following structure in JSON format:

```
{
  "original_query": "The raw query from the older adult",
  "paraphrased_query": "The expert rephrased version of the query"
}
```

Figure 11: Prompt used to generate synthetic data.