

One Question, Four Voices: How Advice for Alzheimer's Caregiving Differs Between Caregivers, Clinicians, and Large Language Models

CONGNING NI, Vanderbilt University, USA and Vanderbilt University Medical Center, USA

JINKYUNG KATIE PARK, Clemson University, USA

YANG LI, Vanderbilt University, USA

SARVECH QADIR, Vanderbilt University, USA

SEAN S. HUANG, Vanderbilt University Medical Center, USA

JAMES S. POWERS, Vanderbilt University Medical Center, USA

JULIA A. HINER, University of Texas Health Science Center at Houston, USA

VANIA LEUNG, University of Illinois Chicago, USA

PATRICIA COMMISKEY, Vanderbilt University Medical Center, USA

LIJUN SONG, Vanderbilt University, USA

QINGXIA CHEN, Vanderbilt University Medical Center, USA

PAMELA J. WISNIEWSKI, International Computer Science Institute, USA

BRADLEY A. MALIN, Vanderbilt University Medical Center, USA and Vanderbilt University, USA

ZHIJUN YIN, Vanderbilt University Medical Center, USA and Vanderbilt University, USA

Online caregiver forums are a critical support infrastructure for Alzheimer's disease and related dementias (ADRD), where people seek both emotional reassurance and practical guidance. Prior CSCW and HCI work has examined peer and clinician support in these communities and, separately, explored support-oriented chatbots and AI-assisted helping; however, we still lack systematic evidence on how AI responses compare with peer and clinician roles when answering the same caregiver questions in situ. We address this gap through a systematic, side-by-side comparison of responses to 85 real caregiver questions from *ALZConnected* across four responder groups: *peer caregivers*, *physicians*, *ChatGPT*, and a retrieval-augmented *CareGPT* grounded in peer discussions. We evaluate responses using 28 measures spanning linguistic form (psycholinguistic cues, structure, and style), emotional support, informational support, and topic scope, including LLM-based scoring for selected measures that we audit and calibrate against double-coded human ratings. Across analyses, we observe a consistent division of communicative labor: peers provided time-anchored narratives and community

Authors' Contact Information: [Congning Ni](#), Vanderbilt University, Nashville, TN, USA and Vanderbilt University Medical Center, Nashville, TN, USA, congning.ni@vanderbilt.edu; [Jinkyung Katie Park](#), Clemson University, Clemson, SC, USA, jinkyup@clemson.edu; [Yang Li](#), Vanderbilt University, Nashville, TN, USA, yang.li.2@vanderbilt.edu; [Sarvech Qadir](#), Vanderbilt University, Nashville, TN, USA, sarvech.qadir@vanderbilt.edu; [Sean S. Huang](#), Vanderbilt University Medical Center, Nashville, TN, USA, sean.huang@vumc.org; [James S. Powers](#), Vanderbilt University Medical Center, Nashville, TN, USA, james.powers@vumc.org; [Julia A. Hiner](#), University of Texas Health Science Center at Houston, Houston, TX, USA, julia.a.hiner@uth.tmc.edu; [Vania Leung](#), University of Illinois Chicago, Chicago, IL, USA, vleung3@uic.edu; [Patricia Commiskey](#), Vanderbilt University Medical Center, Nashville, TN, USA, patricia.commiskey@gmail.com; [Lijun Song](#), Vanderbilt University, Nashville, TN, USA, lijun.song@vanderbilt.edu; [Qingxia Chen](#), Vanderbilt University Medical Center, Nashville, TN, USA, cindy.chen@vumc.org; [Pamela J. Wisniewski](#), International Computer Science Institute, Berkeley, CA, USA, pwisniewski@icsi.berkeley.edu; [Bradley A. Malin](#), Vanderbilt University Medical Center, Nashville, TN, USA and Vanderbilt University, Nashville, TN, USA, b.malin@vumc.org; [Zhijun Yin](#), Vanderbilt University Medical Center, Nashville, TN, USA and Vanderbilt University, Nashville, TN, USA, zhijun.yin@vumc.org.



This work is licensed under a Creative Commons Attribution 4.0 International License.

© 2026 Copyright held by the owner/author(s).

ACM 2573-0142/2026/10-ARTCSCW132

<https://doi.org/10.1145/3816980>

grounding; clinicians provided concise, advice-forward, clinically bounded guidance; and LLMs produced a formal, verbose, highly structured “consultation” voice. Notably, CareGPT shows minimal measurable separation from ChatGPT, indicating that retrieval-plus-prompt-injection alone is insufficient to shift the model’s communicative stance. Quantitatively, LLM outputs score higher than peers on politeness and emotion-support markers, and are comparable to clinicians on informativeness and physician-alignment measures; yet, they are less narrative and less contextually situated. Topic modeling further shows that LLM responses concentrate narrowly on advice-oriented themes, while peers cover a broader range (e.g., emotional processing, legal/financial concerns, and daily cares). These insights show that AI systems can complement, but not replace, human contributions to ADRD caregiver support. We advocate for hybrid systems that combine AI’s structured delivery with the distinct contributions of human voices to strengthen emotional and informational caregiving support infrastructures, while future work should examine whether similar patterns hold in other caregiving domains.

CCS Concepts: • **Human-centered computing** → **Collaborative and social computing**; • **Computing methodologies** → *Natural language processing*; • **Social and professional topics** → *Health informatics*.

Additional Key Words and Phrases: Caregiving, Alzheimer’s disease, AI chatbots, empathy, NLP, online support, health communication

ACM Reference Format:

Congning Ni, Jinkyung Katie Park, Yang Li, Sarvech Qadir, Sean S. Huang, James S. Powers, Julia A. Hiner, Vania Leung, Patricia Commiskey, Lijun Song, Qingxia Chen, Pamela J. Wisniewski, Bradley A. Malin, and Zhijun Yin. 2026. One Question, Four Voices: How Advice for Alzheimer’s Caregiving Differs Between Caregivers, Clinicians, and Large Language Models. *Proc. ACM Hum.-Comput. Interact.* 10, 6, Article CSCW132 (October 2026), 35 pages. <https://doi.org/10.1145/3816980>

1 Introduction

Alzheimer’s caregiving is a profoundly demanding endeavor that requires a tremendous amount of care coordination, often undertaken by informal caregivers such as family members of those living with Alzheimer’s disease or related dementias (ADRD). In the United States alone, an estimated 11 million informal caregivers provide unpaid care to someone with ADRD at home [46]. Caregivers routinely face emotional stress, logistical complexity, and informational uncertainty [118], prompting many to seek connection and guidance through online peer support communities. CSCW researchers have articulated that such online communities operate as critical *care infrastructure* [59], sustaining day-to-day coordination, sense-making, and mutual support when formal resources are limited [3, 55, 105]. We contextualize our research in such a community, one called *ALZConnected*, where ADRD caregivers exchange advice, share lived experiences, and offer emotional validation to one another. Consistent with social support theory [112], these interactions often combine emotional support (e.g., empathy and solidarity) [39] with informational support (e.g., experiential knowledge and practical suggestions) [88]. Increasingly, such caregiving support infrastructures are also becoming sites where emerging AI systems participate alongside peers and professionals, reshaping how guidance and care are produced and encountered [84, 96, 128, 130].

Growing interest in AI-generated support as a scalable, on-demand supplement within caregiving communities stems from structural limitations in peer support, as peer-generated responses can vary in clinical precision and consistency [82], and many caregivers are themselves overwhelmed, leading to unanswered or delayed posts [133]. In contrast, professional responders, such as physicians, are trusted sources of medical expertise, but their limited availability and formal communication norms often constrain their presence in everyday online caregiving spaces and often limit deeper engagement even within clinical encounters [32]. From a CSCW perspective, these human and institutional limitations lead to the need for *care infrastructures* in which support work is distributed across family members, peer communities, and professionals [21, 57]. Emerging LLM-based AI

systems (e.g., ChatGPT) now add an easy-to-access source of on-demand guidance and emotionally supportive replies in health-related contexts, yet prior evidence highlights inconsistent reliability and calls for the need for careful evaluation of AI-mediated healthcare conversations [1, 23, 127]. This shift raises a core CSCW issue in hybrid care infrastructures: Caregivers increasingly receive support from peers, clinicians, and AI systems whose responses may differ in language, emotional framing, and informational content. Understanding these communicative differences is essential for designing socio-technical systems that deliver timely, relevant, and efficient support for everyday caregiving work [21, 57, 61, 102, 138].

Prior work has begun to examine how different support sources compare in the kinds of help they provide, but evidence remains fragmented across separate lines of work. Studies of online health communities characterize peer responses as rich in lived experience, storytelling, and emotional affirmation, yet also uneven in timeliness and informational rigor [28, 91]. Complementary work on clinician involvement highlights how professional expertise can improve clinical precision, while also showing that participation is constrained by limited availability and clinical communication norms [70]. More recently, evaluations of LLM-mediated health support compare model-generated replies with human guidance and document a trade-off between scalable, well-structured responses and risks of factual errors, missing context, or unclear provenance [5, 49]. Building on these themes, recent CSCW/HCI work further motivates the need to scrutinize how conversational agents function within caregiving networks and how evaluation practices should reflect real-world support needs [61, 138]. At the same time, it is increasingly common to augment general-purpose LLMs with retrieval-augmented generation (RAG) to ground responses in domain-relevant evidence, motivating systems such as our designed *CareGPT*. However, we still lack a *side-by-side* characterization of how peers, clinicians, and LLM-based systems (with or without RAG) differ when responding to the *same* caregiver needs, and how such differences should inform the design of hybrid care infrastructures [59] in which multiple support voices co-exist.

To fill this gap, we conduct a systematic comparison of how four support categories respond to real caregiver questions posted on *ALZConnected*: (1) online peer caregivers, (2) board-certified physicians, (3) ChatGPT, and (4) CareGPT, implemented by enhancing GPT-4 with RAG over online caregiving discussions. We select these four support categories to reflect a hybrid *care infrastructure* in which support is produced through multiple roles and tools: peer experience, clinical expertise, and AI systems that are increasingly available alongside them [21, 57]. Rather than ranking responders, we examine how AI responses align with or diverge from human communication when caregivers may encounter different sources for the *same* high-stakes question. Guided by care infrastructures [59], we examine how different responders contribute to caregiving communication using three complementary dimensions: linguistic characteristics, social support functions, and topical focus. In our evaluation, we chose support dimensions that aligned with social support theory, motivating our focus on *emotional* and *informational* support as core support functions in online scenarios [112]. We include linguistic features to capture how responses are delivered (e.g., structure, readability, affective framing) and topic patterns to characterize how caregiving needs are addressed across responders [76]. Therefore, we organize our research around three research questions:

- **RQ1:** How do responses from peer caregivers, physicians, and LLM-based systems (ChatGPT, CareGPT) differ in linguistic characteristics (e.g., structure, style, and psycholinguistic cues)?
- **RQ2:** How do these responder groups differ in emotional and informational support (e.g., reassurance/validation and advice/resources) when responding to caregiver questions?
- **RQ3:** How do linguistic and support-related patterns vary across **caregiving topics**, and how do these variations differ by responder group?

To investigate these questions, we analyze 85 caregiver-posted questions from *ALZConnected*¹, each paired with four responses, with one per responder group. This question-controlled design enables systematic comparison of how peers, clinicians, and LLM-based systems respond to the same caregiving needs, treating each response as a situated act of care within a broader caregiving infrastructure. We operationalize linguistic structure (e.g., verbosity, readability), psycholinguistic markers (e.g., emotional tone, cognition & perception), emotional and informational support cues (e.g., empathy, information evaluation), and topical focus (via BERTopic modeling), as observable traces of how care is enacted, coordinated, and made legible across different support roles. We develop NLP tools and LLM-based scoring pipelines that we validate through human evaluation metrics to ensure reliability and contextual sensitivity, aligning quantitative measures with CSCW theories of care, support work, and sociotechnical participation. Our contributions to CSCW and human-AI interaction research are three-fold:

- We empirically show that peer caregivers, clinicians, and LLM-based systems enact caregiving support through systematically different communicative roles, revealing that caregiving support is organized through distinct but complementary forms of relational, informational, and accountable labor within a hybrid care infrastructure, rather than convergence toward a single “best” form of support.
- We demonstrate how emotional and informational support are differently balanced across responder groups, with peers grounding care in lived, narrative experience, clinicians providing bounded and clinically oriented guidance, and AI systems offering structurally consistent, polite, and advice-forward responses. This extends social support theory by showing how AI operates as a distinct infrastructural participant within caregiving support ecosystems.
- We show that these differences are not uniform, but vary meaningfully across caregiving topics, indicating that caregiving support is situationally enacted and that the value and limitations of AI, peer, and clinician responses depend on the nature of the caregiving concern.

Overall, these findings advance CSCW discussions of care infrastructures by characterizing how support work is redistributed across human and AI actors through distinct relational roles and trust cues, and inform the design of hybrid caregiving systems that coordinate complementary roles rather than treating AI as a replacement for human care in online ADRD caregiving support communities.

While we focus on ADRD caregiving, this infrastructure-oriented approach can be extended to other caregiving and health support settings where people increasingly navigate hybrid human-AI support ecosystems.

2 Related Work

2.1 Care Infrastructures and Support Functions in CSCW Research

CSCW research has long conceptualized care and support not as isolated, one-off interactions, but as ongoing cooperative work sustained through coordination, participation, and day-to-day maintenance across people and technologies [21, 60]. Building on this tradition, recent CSCW scholarship has articulated an *infrastructural care* lens, which emphasizes how caregiving practices are assembled and strained through heterogeneous support sources such as family members, peer caregivers, clinicians, and sociotechnical systems that collectively enable care under conditions of uncertainty and constraint [10, 33, 60, 131]. In long-term Alzheimer’s caregiving contexts, where caregiving needs evolve over time and caregivers often live under high workload, engaging with and interpreting support from multiple sources becomes part of the work of sustaining

¹<https://alzconnected.org/discussions/tagged/Asking-for-Help>

care itself [3, 78, 129]. From this perspective, a key challenge for designing effective hybrid care infrastructures lies not only in the availability of support, but in understanding how different support sources communicate and frame support or assistance across caregiving challenges. Without a clear empirical understanding of how these responder groups differ in their communicative patterns, it remains difficult to design socio-technical systems that efficiently and appropriately support caregiving needs.

To describe *what kinds* of support are exchanged within these infrastructures, communication research distinguishes multiple support types. In text-based online forums, emotional and informational support are among the most visible and widely investigated forms [64]. CSCW and health communication research further suggest that informational support is closely tied to lay sensemaking and negotiation over what counts as actionable knowledge [50]. In ADRD caregiving, support needs often combine practical guidance with emotional strain and self-care tensions, reinforcing why emotional and informational support are central analytical dimensions [117]. At the same time, when support is mediated by technology, interactional cues (e.g., assertiveness, hedging, or adaptation to expertise) can shape how guidance is interpreted; similar HCI work in clinician-AI interaction shows communication style should be treated as part of system behavior rather than merely “packaging” [16–18]. In this paper, we use care infrastructure [59] to motivate why it is meaningful to compare peers, clinicians, and AI *together* as co-present sources of support. We further draw on social support theory [112] to focus our comparison on emotional and informational support, and examine these dimensions through linguistic characteristics and topic patterns in responses.

2.2 Support from Online Peer Caregivers and Clinicians

A large body of CSCW and health-community research shows that caregivers rely on online forums to share experiences, seek advice, and cope with long-term caregiving demands [56, 68, 78]. These communities commonly provide emotional support through encouragement and validation [45], and informational support through practical strategies and shared resources [126]. People living with ADRD often get worsen conditions by gradually losing memory and cognitive ability [26], thus demanding far more care than other older adults. Therefore, caregiving discussions often reflect prolonged emotional strain, uncertainty about progression, and ongoing logistical coordination [118]. Qualitative studies further document recurring tensions around ethics, coping, and family conflict in caregiver discourse [43, 123], including work that highlights how “informal” caregiving conversations surface complex responsibilities and strain [38]. Large-scale analyses also suggest measurable benefits of peer interaction, such as improved affect following peer responses [90]. At the same time, caregivers often seek support because they are under prolonged stress and isolation [58], which includes explicit requests for comfort and validation [89].

Online caregiving infrastructures face informational challenges. Caregivers often need reliable, context-relevant information when decisions can shape care trajectories [77], and studies document how caregivers search for and evaluate information under uncertainty [110]. Clinicians are authoritative sources, but their responses can be brief or sometimes difficult to interpret for lay audiences [44], and sustained participation in everyday online support settings is often constrained by time, institutional boundaries, and professional communication norms [32]. A study on clinician participation shows how professional input can complement peer experience, while still being difficult to sustain in practice [70]. Peer caregivers provide experiential knowledge that can feel more relatable to caregiving experience, but advice quality can vary in evidence and generalizability [15], including uneven clinical precision and consistency [82].

Care support, however, is not only about exchanging facts and knowledge; it also depends on how people respond to each other. CSCW research emphasizes the affective component of caregiving

communication in chronic illness contexts [9], including conversational practices that accomplish empathy through concrete interactional moves [106]. More broadly, scholarship on socio-emotional labor describes the effort required to manage one's own emotions while responding with care [69]. In online communities, this can appear through storytelling and responsive mirroring [66], and repeated peer support practices that sustain participation over time [104]. Such support can matter for relationship quality and outcomes for support seekers [125]. Conversely, messages that feel blunt or automated can reduce perceived warmth and credibility [35]. Similarly, while clinicians provide valued professional guidance, their expressions of empathy are often framed through clinical expertise rather than shared lived experience, which can differ from the emotionally grounded support offered by peer caregivers. These studies motivate a question-controlled comparison of how different support categories respond to the same caregiving needs in terms of linguistic delivery, topical patterns, and support types [117].

2.3 LLMs in Caregiving Support

Recent advances in LLMs have renewed interest in AI as a participant in health-related communication. In some settings, evaluators rate LLM answers as helpful and empathetic, and may even prefer them over clinician-written answers on perceived quality [5]. Related work finds that LLMs can produce emotion-aware and socially appropriate responses across different health question formats [119, 136], including evaluations of ChatGPT-style models for assessment and response generation in health contexts [132]. Researchers have also explored LLM-based systems that aim to better model supportive dialogue, including empathy-oriented tuning [22], alongside broader evaluations of LLM capabilities for health communication [20]. In caregiving-adjacent scenarios, interaction style also matters: caregivers and lay users may read coherence, politeness, and clarity as cues of support [137].

At the same time, research cautions that AI support can feel generic, overly polished, or simply mismatched to sensitive situations [27, 94]. Studies of AI-mediated emotional support raise concerns about authenticity and “performed” empathy [109], including analyses of how systems are designed to *create* or signal empathy in interaction [111]. Related findings suggest that users may accept chatbot support for everyday stress while still preferring human support in harder moments [135]. In caregiving networks, limitations often surface when nuanced, context-sensitive support is most needed [138]. This has motivated hybrid designs that try to *augment* human supporters rather than replace them: for example, Sharma et al. introduce a system that gives just-in-time feedback to peer supporters to help them craft better replies while keeping humans in the loop [115]. More broadly, HCI studies show that communication cues are part of system behavior, which is work on agent assertiveness [18], explanations adapted to expertise levels [17], and human-centered guidance for personalized agents [16] all point to how framing, tone, and explanation choices can shape user judgment and downstream action.

A further concern is reliability in high-stakes advice settings. LLMs can generate confident but incorrect statements [2], which has prompted calls to prioritize safety and trustworthy evaluation over fluency [4]. Related work discusses practical evaluation concerns and failure modes for AI-mediated healthcare conversations [1], as well as tensions around trustworthiness and normative fit when AI enters support communities [23]. Researchers have explored RAG and other verification-oriented mechanisms to improve grounding and transparency [37]. Work on how people judge AI outputs emphasizes interpretability and perceived reliability cues [54], and studies of algorithm-mediated services show that users develop “folk theories” and strategies under uncertainty [75]. Finally, adjacent HCI work on emotion-focused technologies cautions against overclaiming what AI can do in affective domains [74], and argues for treating emotion regulation and care as situated practices rather than generic text outputs [73]. Together, these concerns motivate evaluating LLM

responses as one voice in a broader care infrastructure, not as a standalone replacement to human support providers. From a CSCW perspective, this framing foregrounds caregiving support as ongoing, distributed work accomplished through heterogeneous actors and technologies. Building on this view, our study examines how AI systems, peer caregivers, and clinicians differentially enact care within a shared support infrastructure when responding to the same caregiver requests. We next describe our study design and evaluative approach that operationalizes and compares this infrastructural caregiving perspective.

3 Methodology

Guided by *care infrastructure* (§2.1), we designed a question-controlled comparative study to examine how different responder groups provide support within an online caregiving setting. We compare peer caregivers, clinicians, and LLM-based systems, evaluating responses along linguistic characteristics, emotional and informational support, and topic-specific patterns.

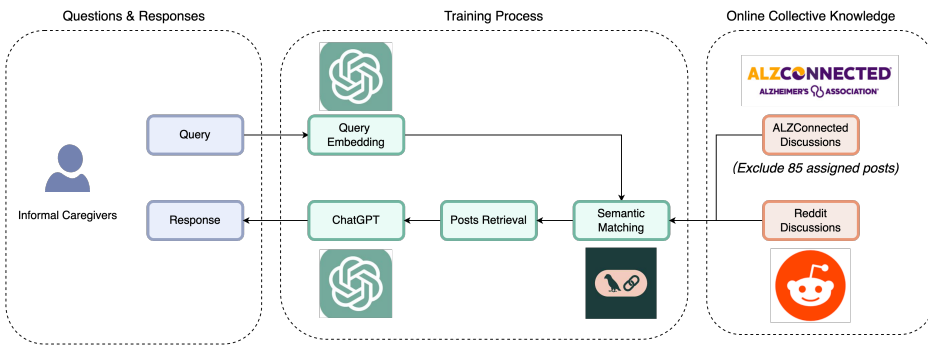


Fig. 1. The design of CareGPT, including query embedding, semantic matching, relevant post retrieval, and response generation.

3.1 Implementation of CareGPT

We developed CareGPT as a LangChain-based RAG system tailored to dementia caregiving. Fig. 1 illustrates the pipeline of CareGPT. Unlike vanilla ChatGPT, CareGPT grounds its responses in semantically retrieved content from real online peer discussions, enabling the model to draw on community-situated experiential knowledge that circulates within caregiving forums.

The system indexes a curated knowledge base constructed from 50,000+ posts (**excluding** the 85 authentic caregiver questions for our study) and 300,000+ comments drawn from ALZConnected and Reddit caregiving forums. These discussions were embedded using OpenAI's text-embedding-ada-002-v2 model and stored in a FAISS vector database for semantic search. A caregiver's query is embedded, and the top- k ($k = 4$) semantically similar posts, together with their associated comments, are retrieved. The retrieved content is anonymized using presidio [83] before being injected into GPT-4 via structured prompts. To minimize topical drift from long documents, posts are embedded independently rather than concatenated into chunks, thereby preventing unrelated comments from being conflated during retrieval.

Since the quality of RAG systems depends on both retrieval and generation, we manually assessed the retrieval performance. To ensure broad topical coverage, we sampled a stratified subset of caregiver questions. Two annotators independently labeled each retrieved post along two binary dimensions: (1) whether the post was topically **relevant** to the caregiver's query, and (2) whether

it showed observable **influence** on the generated CareGPT response. Annotator agreement was then summarized using percent agreement and Cohen's κ .

3.2 Study Design and Dataset

We conducted a mixed-methods, question-controlled comparative study to examine how different support categories respond to the *same* caregiver needs (i.e., based on requests posted on the *ALZConnected* forum) within an online caregiving infrastructure. Each caregiver question was paired with one response from four distinct responder groups, allowing us to compare how support is delivered across linguistic characteristics, emotional and informational support functions, and topic-specific patterns.

We curated 85 anonymized caregiver questions from the “*Asking for Help*” board of the *ALZConnected* forum², posted between April 7 and October 16, 2023. These posts are publicly accessible, and for this study we collected only textual content and removed usernames and other obvious identifiers prior to analysis. Within this time window, we included *all* advice-seeking posts that (a) were written in English by informal caregivers, (b) were first-person requests for help, and (c) had at least one peer reply, yielding 85 eligible questions. This final 85 questions, therefore, reflects the complete set of posts that met our inclusion criteria while remaining feasible for collecting matched physician and AI responses for each question.

Each question was matched with four responses from the following responder groups:

- *Online peer caregivers*, authentic replies posted by community members on *ALZConnected*;
- *Physicians*, answers authored by four board-certified physicians recruited to contribute expert responses for this study;
- *ChatGPT-4*, responses generated by OpenAI's GPT-4 model;
- *CareGPT*, responses from the GPT LangChain-based RAG system as described in §3.1.

We selected GPT-4 due to its prominence and extensive evaluation in health and well-being research at the time of this study [127]. To keep communicative intent comparable across non-peer responders, we used a single standardized instruction for both physicians and AI models: “*Envision yourself in your office with the caregiver and provide an answer that is both empathetic and accurate.*” This prompt establishes a shared consultation scenario without prescribing specific wording; we interpret stylistic differences (e.g., degree of formality) as properties of this controlled setup rather than fixed traits of any responder in all contexts. Peer caregiver responses were drawn directly from the forum and anonymized prior to analysis.

Because caregiver threads sometimes contained multiple comments, we treated the set of replies to a given question as the peer response for that item by first computing each metric at the individual-comment level and then averaging these values across comments, yielding one peer-level score per question. For physicians, ChatGPT, and CareGPT, each question received exactly one response (one physician answer and one first-pass generation from each AI model), so metrics were computed directly on these single responses. Figure 2 shows one example caregiver question and the corresponding responses from the four responder groups.

The study protocol was reviewed and approved by our Institutional Review Board (IRB). Four board-certified physicians with clinical experience in geriatric medicine, older adult care, and dementia-related care were recruited to contribute physician responses to the 85 caregiver questions under the approved study protocol. We recognize that caregiving discussions involve sensitive emotional content and that community members may not anticipate research reuse in all contexts. To reduce risk, we (1) removed all personally identifiable information (e.g., usernames, contact details), (2) analyzed data only in aggregate and report results at the level of responder groups rather

²<https://alzconnected.org/discussions>

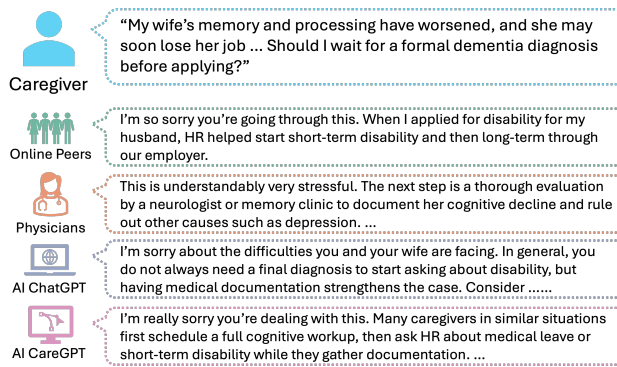


Fig. 2. Example caregiver question (top) and corresponding responses from online peer caregivers, a physician, ChatGPT-4, and CareGPT (bottom). The figure mirrors the conversational structure of our dataset, in which one caregiver question is paired with four responder groups. Text has been lightly paraphrased and truncated for readability and to preserve anonymity.

than individuals, (3) paraphrased and truncated any illustrative excerpts to avoid search-based re-identification, and (4) do not release raw post-level text in any public artifact. All data were stored and analyzed on secure institutional infrastructure.

We evaluated each response using 28 predefined features spanning five dimensions: linguistic structure, psycholinguistic markers, emotional support, informational content, and topic relevance. Quantitative comparisons using Analysis of Variance (ANOVA) were conducted across responder groups. To control for Type I error from multiple comparisons, we applied a Bonferroni correction with a conservative threshold of $\alpha = 0.05/28 = 0.0018$, treating results with $p < 0.001$ as statistically significant. A post hoc power analysis using G*Power [36] confirmed that our dataset of 340 responses achieved a statistical power of 0.981 for detecting medium-sized effects in one-way ANOVA, indicating high sufficiency of the response volume for our planned comparisons.

3.3 Evaluation Metrics

To support a question-controlled comparison across responder groups, we operationalize each response using a set of measurable communication properties aligned with our three research questions: linguistic characteristics (RQ1), emotional and informational support (RQ2), and topic-specific patterns (RQ3). We organize these measures into six categories: **Psycholinguistic Characteristic**, **Linguistic Structure**, **Linguistic Style**, **Emotional Support**, **Informational Support**, and **Topic Modeling**. Our psycholinguistic and linguistic measures build on foundational work in computational psycholinguistics and language use [100, 121] and on text comprehension and readability research [81]. For emotional and informational support, we draw on social support theory and related operationalizations in online health communication, where these two support types are widely studied and measurable in text [11, 30, 90, 112].

The final set of 28 features was chosen from a larger pool of candidate metrics derived from prior work and available NLP tools. We retained measures that (a) mapped cleanly onto our three research questions, (b) avoided redundancy (e.g., highly correlated variants of the same construct), and (c) behaved stably in pilot analyses on a small subset of the data. For instance, among several available readability indices, we focused on Dale-Chall and SMOG as representative measures of comprehension difficulty.

Table 1. Feature categories, metrics, and their theoretical meaning in response communication.

Category	Metric Name	Measurement Explanation
Psycholinguistic	Emotional Tone	Share of affective words (positive or negative), reflecting emotional valence.
	Cognition & Perception	Use of reasoning, insight, or certainty terms, indicating cognitive processing.
	Social Concerns	References to family, friends, or personal roles, showing social orientation.
	Health Concerns	Mentions of illness or health-related terms, reflecting biomedical focus.
	Temporal References	Use of past/present/future verbs, showing temporal framing.
	Informal Language	Presence of slang, netspeak, or fillers, signaling casual tone.
Linguistic Structure	Verbosity	Total word and sentence counts, indicating response length.
	Readability	Estimated ease of comprehension using standard formulas.
	Complexity	Syntactic depth from dependency parsing, reflecting structural intricacy.
	Lexical Diversity	Vocabulary range and uniqueness, indicating linguistic variation.
	Repetition	Rate of repeated words, suggesting redundancy.
Linguistic Style	Formality	Degree of formal grammar and tone.
	Analytic/Narrative Style	Dominance of analytic vs. narrative framing.
Emotional Support	Politeness	LLM-derived score for courteous or softening phrases.
	Emotion Evaluation	LLM-estimated empathy, encouragement, and validation.
	Emotional Diversity	Variety of expressed emotions, measured by entropy.
	Emotional Composition	Distribution of emotion types (e.g., joy, sadness).
Informational Support	Physician Semantic Proximity	Semantic similarity between the response and physician-authored responses, indicating informational alignment.
	Relative Score	Similarity to the original question, indicating relevance.
	Uncertainty Cues	Presence of hedging or cautious language.
	Information Evaluation	LLM-based score for factual or instructional content.
	Clinical Concept Overlap	Count of medical concepts (e.g., diseases, treatments).
Topic Modeling	Dominant Topic	Main topic assigned by unsupervised modeling.

Table 1 summarizes the complete set of metrics used in this study. For each feature, we specify its category, name, and measurement rationale. These metrics were introduced as part of our standard evaluation pipeline and integrated into the methodological design of **RQ1–RQ3**.

3.3.1 Linguistic Feature Extraction (RQ1).

Psycholinguistic. Psycholinguistic features capture affective, cognitive, and social orientations expressed in everyday language. In caregiver-support settings, these cues help characterize *how* different responders (peers, physicians, and LLM-based systems) frame their guidance and relate to caregivers, beyond the literal content of what is said [100, 121]. For example, language indicating reasoning and certainty can shape how advice is interpreted, and affective framing (e.g., encouragement or validation) can influence how support is received [48]. This is particularly relevant when comparing AI-generated responses with human replies in sensitive contexts such as dementia caregiving.

We operationalized these constructs using **LIWC 2022** [13], a lexicon-based tool that maps word use to psychologically meaningful dimensions. From each response, we extracted six LIWC-based

metrics: (1) *Emotional Tone*, reflecting affect-laden wording; (2) *Cognition & Perception*, capturing terms related to insight, causality, and certainty (e.g., *because*, *think*, *understand*) [99]; (3) *Social Concerns*, measuring references to relationships and social context (e.g., family), which are common in peer support [91]; (4) *Health Concerns*, capturing illness- and clinical-related terms [97]; (5) *Temporal References*, tracking past/present/future orientation as a proxy for narrative grounding and planning [95]; and (6) *Informal Language*, including netspeak, fillers, and mild swearing, which are often associated with conversational style and approachability in online discourse [120]. Feature values were normalized by total word count to enable comparison across responses of different lengths.

Linguistic Structure. Linguistic-structure features characterize how a response is organized and how difficult it may be to read. These properties matter in caregiving forums because caregivers often seek clear, usable guidance under stress; structure that is too sparse can reduce interpretability, while overly complex writing can increase cognitive load.

We measured five structural dimensions. *Verbosity* reflects the amount of detail provided and can shape perceived thoroughness in health-related exchanges [71]. *Readability* captures comprehension difficulty; simpler writing is often associated with higher engagement and trust in health contexts [34]. *Syntactic complexity* reflects how layered sentence constructions are, which can differ between professional and informal communication [67]. *Lexical diversity* measures vocabulary variation and linguistic richness [79, 108]. *Repetition* captures redundancy or emphasis strategies [47].

Operationally, *Verbosity* was measured by total word count and sentence count using *textstat* (v0.7.3). *Readability* was assessed with Dale-Chall and SMOG via *py-readability-metrics* (v1.2.0) and *textstat*. *Syntactic complexity* was computed as average dependency parse tree depth using *spaCy* (v3.7.2; *en_core_web_lg*). *Lexical diversity* was quantified using Type-Token Ratio (TTR) [108] and MTLT [79], computed with *lexical-diversity* (v0.1.6). Finally, *Repetition* was defined as the proportion of repeated non-stopword tokens within a response, computed from normalized counts of recurring words.

Linguistic Style. Style features capture how a message is expressed (e.g., formal vs. conversational, narrative vs. analytic), which can shape perceptions of authority, relatability, and engagement in caregiver support settings. Prior work suggests that formality often signals expertise and distance, while more conversational language can support social closeness and approachability [98]. Likewise, narrative language grounded in lived experience can convey empathy and shared understanding, whereas analytic language tends to emphasize structured reasoning [65, 101].

We operationalized two style dimensions: *Formality* and *Analytic/Narrative Style*. *Formality* was measured using a RoBERTa-based classifier (*s-nlp/roberta-base-formality-ranker*) implemented with HuggingFace transformers (v4.38.1), producing a continuous score where higher values indicate more formal responses. This aligns with evidence that LLM outputs can skew formal unless prompts explicitly elicit a more conversational tone [29, 40]. *Analytic/Narrative Style* was measured using the Categorical-Dynamic Index (CDI), which distinguishes categorical, structured language from dynamic, experiential discourse [101]. CDI was computed from LIWC 2022 outputs as:

$$\text{CDI} = \text{article}_{\text{scaled}} + \text{prep}_{\text{scaled}} - \text{ppron}_{\text{scaled}} - \text{ipron}_{\text{scaled}} - \text{auxverb}_{\text{scaled}} - \text{conj}_{\text{scaled}} - \text{adverb}_{\text{scaled}} - \text{negate}_{\text{scaled}} \quad (1)$$

Each term is a normalized LIWC category count scaled by total word count. Higher CDI indicates a more analytic, categorical orientation, whereas lower CDI indicates a more narrative, experiential style.

3.3.2 Emotional and Informational Support Metrics (RQ2). Beyond general linguistic characteristics, we evaluated two caregiver-relevant support functions, *emotional supportiveness* and *informational value*, to compare how different responder groups address caregivers’ affective needs and practical questions (RQ2). This focus aligns with social support theory, where emotional and informational support are commonly treated as distinct (but often co-occurring) forms of support in health-related exchanges [6, 124]. To enable consistent scoring at scale without extensive manual annotation, we used LLMs as automated judges for selected dimensions [122]. Specifically, we used LLaMA 3.3 for LLM-as-judge scoring, separating generator and evaluator roles to reduce shared-model bias. We then validated these LLM-based judgments with a dedicated human study (§3.3.4), confirming substantial alignment with expert coder ratings.

Emotional Support. Emotional support includes encouragement, validation, and expressions of concern, and is a central component of caregiver support exchanges [124]. In caregiving forums, perceived emotional quality can shape whether caregivers feel understood and whether they continue to engage with support sources [51]. Because emotional support can appear both in overall stance and in finer-grained affective signals, we operationalized four complementary aspects: (1) *Politeness*, (2) *Emotion Evaluation*, (3) *Emotional Diversity*, and (4) *Emotional Composition*.

Politeness was scored using a prompt-based classifier built with LLaMA 3.3. Each response was passed to the model with a schema-constrained prompt that required a structured JSON output containing: (a) a politeness score in $[0, 1]$, (b) a self-rated confidence level (1-5), and (c) salient cue words supporting the score. Using the same LLaMA-based setup, *Emotion Evaluation* captured whether a response explicitly conveyed empathy-adjacent behaviors (e.g., encouragement or validation). We treated these outputs as coarse-grained indicators of emotional support and evaluated their quality via human validation (§3.3.4).

To characterize affective content more finely, we used emotion classification under the GoEmotions taxonomy [31], grounded in Plutchik’s primary emotion theory [103]. Using a pre-trained BERT-based model from Google, we extracted a 28-dimensional probability vector over discrete emotions (e.g., *admiration*, *joy*, *sadness*, *remorse*, *confusion*). We refer to this vector as *Emotional Composition*, representing the intensity profile of affective categories in each response.

We then computed *Emotional Diversity* (“emodiversity”) as the entropy of the normalized GoEmotions vector $\vec{p} = (p_1, \dots, p_{28})$. Prior work links higher emotional diversity with better emotional regulation and well-being [42, 107]. We calculated:

$$\text{EmoDiv}_{28} = - \sum_{i=1}^{28} p_i \cdot \log_2(p_i) \quad (2)$$

Higher values indicate a broader spread of affective categories rather than reliance on a single dominant tone. Together, these measures capture both stance-level supportiveness and the affective texture of responses across responder groups.

Informational Support. Informational support refers to how much a response provides relevant, usable guidance (e.g., explanations, clarifications, or concrete suggestions) for a caregiver’s question [63]. In health-related settings, information that is clear and sensitive to context is especially important when caregivers are making decisions under stress [80]. To compare informational support across responders, we operationalized five constructs: (1) *Physician Semantic Proximity*, (2) *Relative Score*, (3) *Uncertainty Cues*, (4) *Information Evaluation*, and (5) *Clinical Concept Overlap*.

Physician Semantic Proximity assessed how closely a non-physician response (peer or AI) aligned with the physician response for the same question. We computed cosine similarity between the two responses, following the general idea of using expert alignment as a reference point when

evaluating informal health content [92]. Higher similarity indicates greater convergence in semantic space, though not necessarily better interpretability.

Relative Score measured response relevance to the caregiver's original question. We generated sentence embeddings using all-MiniLM-L6-v2 (sentence-transformers v2.7.0) and computed cosine similarity in [0, 1]. Higher values indicate stronger alignment with the caregiver's stated intent, consistent with relevance-based notions of answer quality in community Q&A [52].

Information Evaluation was scored by prompting LLaMA 3.3 to rate the extent to which a response provided concrete information, actionable advice, or practical suggestions. The model returned (a) a score in [0, 1], (b) a confidence level (1-5), and (c) a short rationale. As with emotional scoring, we validated these LLM-judge outputs against human annotations (§3.3.4).

Uncertainty Cues captured whether responses signaled caution or conditionality (e.g., *might*, *could*, *seems like*). We extracted modal/hedging markers via part-of-speech tagging in spaCy (v3.7.2) and counted hedging frequency per response, following pragmatic work on hedging as a strategy of epistemic humility and face-saving [53, 72].

Finally, *Clinical Concept Overlap* approximated clinical specificity by identifying biomedical entities using scispaCy (en_core_sci_md, v0.5.2) [87]. We extracted concepts related to diagnoses, medications, symptoms, and procedures, and used the total number of medically relevant concepts per response as a proxy for clinically grounded language.

Together, these five measures provide a multidimensional view of informational support, spanning relevance, actionable content, uncertainty signaling, and clinical specificity, enabling systematic comparison across responder groups for **RQ2**.

3.3.3 Topic Modeling and Contextual Analysis (RQ3). In addition to linguistic and support features, we examined how caregiver concerns are distributed across thematic domains using unsupervised topic modeling. This content-aware analysis addresses **RQ3**, which investigates whether the nature and quality of responses vary depending on the topics of the original caregiver question. Prior studies in online health support communities have shown that topics such as emotional burnout, medication decisions, or financial stress often elicit distinct patterns of emotional expression and informational delivery [8, 24, 93]. By accounting for this variation, we aim to contextualize the linguistic and support differences observed across responder groups, including online peers, physicians, ChatGPT, and CareGPT.

We applied BERTopic (v0.16.1), a transformer-based topic modeling method that combines class-based TF-IDF with semantic clustering over document embeddings [41]. This approach builds upon the foundations of Latent Dirichlet Allocation (LDA) [12], but improves cluster quality by incorporating modern contextual embeddings. We used DistilBERT representations (sentence-transformers/distilbert-base-nli-stsb-mean-tokens) and configured the model with a minimum topic size of 15 to balance coherence and diversity. Each response, regardless of whether it was written by a peer, physician, or AI model, was assigned a primary topic based on the highest-probability label output by the model. For online peer caregivers, we treated each individual comment as a separate document when fitting the topic model; for physicians, ChatGPT, and CareGPT, each answer to a caregiver question was modeled as a single document. We then used BERTopic's `visualize_topics_per_class` function to compute the distribution of dominant topics across responder classes. This enabled us to identify whether different responder groups focused on distinct caregiving domains.

3.3.4 Human Validation of LLM-based Metric Scores. To make clear that our LLM-as-judge scores are not used as unexamined proxies, we conducted a human validation study for the three judge-scored dimensions: *Politeness*, *Emotional Evaluation*, and *Informational Evaluation*. We sampled a stratified subset of caregiver questions (covering diverse topics and intents) that remained feasible

for manual coding. For each sampled question, we included the corresponding responses and asked three graduate-level coders, with prior experience analyzing online caregiver and health-support communication, to rate each dimension on a 5-point Likert scale (1=*very low*, 5=*very high*). We averaged ratings to form a per-item human consensus and linearly rescaled scores to $[0, 1]$ to match the range of the automatic metrics.

To support consistent interpretation, coders followed a written rating rubric that defined each dimension (politeness, emotional evaluation, informational evaluation) and its anchors on the 5-point scale; they also completed a brief calibration round on held-out examples to align their use of the rubric before rating the study subset. The rubric was developed by the research team for this study, informed by prior literature on politeness, emotional support, and informational support, and tailored to the three evaluation dimensions used in our analysis. We assessed agreement using pairwise weighted Cohen's κ and pairwise Spearman rank correlation, and report these statistics to characterize reliability across coders. Finally, to evaluate alignment between human consensus and automatic scores, we computed Spearman's ρ between the human consensus and the corresponding LLM-based metric for each dimension, and inspected residuals to check for systematic bias.

3.4 Statistical Analysis

To examine differences in linguistic and support-related features across responder groups, we performed both statistical analyses and evidence-based interpretation. For each numeric feature relevant to **RQ1**, **RQ2**, and **RQ3**, we began with a one-way ANOVA, using *responder group* (Online Peer, Physician, ChatGPT, CareGPT) as the independent variable. This statistical test assessed whether the population means for each group significantly differed. When the ANOVA yielded a statistically significant result ($p < 0.001$), we conducted post-hoc pairwise comparisons using **Tukey's Honestly Significant Difference (HSD)** test to identify specific group differences. Tukey's test reports the *meandiff* between Group₂ and Group₁ ($\text{meandiff} = \text{Group}_2 - \text{Group}_1$), where a positive value indicates that Group₂ scored higher. In contrast, we computed **Cohen's d** using the reverse direction: $(\text{Group}_1 - \text{Group}_2)$ divided by the pooled standard deviation. Thus, *meandiff* and d values are always of opposite sign, ensuring clear directional interpretation, e.g., a negative d indicates that Group₂ outperformed Group₁.

Each ANOVA result includes the F -statistic, p -value, and the proportion of explained variance, reported as eta-squared (η^2). We interpret η^2 using conventional benchmarks [25]: 0.01 (small), 0.06 (medium), and 0.14 (large). Cohen's d effect sizes were similarly classified as 0.2 (small), 0.5 (medium), 0.8 (large), and above 1.2 (very large). We highlight findings that were both statistically significant and practically meaningful, defined as $\eta^2 > 0.14$ and/or $|d| > 1.2$. In addition to the statistical analysis, we also conducted an interpretation of the results, focusing on the thematic richness and emotional depth of responses. After statistical testing, we inspected short representative quotes from each responder group to help interpret statistically significant differences in style, emotional support, and information quality. This targeted qualitative reading was used solely to contextualize the quantitative findings and does *not* constitute a standalone qualitative analysis. All statistical tests were performed using *statsmodels* (v0.14.1) for ANOVA and *pingouin* (v0.5.4) for post-hoc analysis and effect size estimation. Visualization of group comparisons, including boxplots and bar plots with mean $\pm$ standard error of the mean (SEM), was performed using *matplotlib* (v3.8.2) and *seaborn* (v0.13.2).

4 Results

4.1 Validation of Retrieval and Evaluation Metrics

4.1.1 Validation of Retrieval Performance in CareGPT. To evaluate retrieval quality, we sampled 15 caregiver questions (15/85, $\approx 18\%$ of the corpus) and two annotators annotated all retrieved posts for both *relevance* and *influence* on the generated CareGPT response. Inter-annotator agreement was strong for both dimensions: relevance ($\kappa = 0.80$, 94.44% agreement) and influence ($\kappa = 0.91$, 97.21% agreement). On average, the majority of retrieved posts were judged positively, with a relevance rate of 83% and an influence rate of 82% (proportion of items labeled “True” across annotators). These findings indicate that CareGPT’s retrieved content was not only on-topic but also visibly shaped its generated responses.

4.1.2 Verification of LLM-based Metric Scoring. To assess whether LLM-as-judge metrics were qualified, we validated them against human annotations on the same 15-question subset. Human consensus ratings across politeness, emotional support, and informational support demonstrated strong inter-annotator reliability (mean $\rho = 0.81$, mean $\kappa_w = 0.78$). Comparisons with LLM-as-judge scores showed strong and statistically significant correlations: politeness ($\rho = 0.86$, $p < 0.001$), emotional support ($\rho = 0.81$, $p < 0.001$), and informational support ($\rho = 0.85$, $p < 0.001$). These alignments confirm that automated metric scores closely tracked human consensus across communicative dimensions, supporting the validity of using LLM-as-judge for large-scale evaluation.

4.2 Linguistic Differences Across responder groups (RQ1)

We examined the linguistic characteristics of responses across four responder groups using 17 linguistic features grouped into psycholinguistic, structural, and stylistic categories. Table 2 summarizes statistically significant differences between groups (ANOVA $p < .001$, $\eta^2 > 0.14$), and we report effect sizes for pairs exhibiting large practical differences (Cohen’s $|d| > 1.2$). Figure 3 visualizes the most differentiating features using boxplots. Beyond the numerical contrasts, we interpret these linguistic patterns as different ways of “doing care” (e.g., narrating lived experience vs. delivering structured guidance).

4.2.1 Psycholinguistic Complexity. Among psycholinguistic features, only **Temporal References** showed significant divergence. **ChatGPT** and **CareGPT** used fewer temporal cues than **Online Peers** ($p = 3.12 \times 10^{-15}$), with large effect sizes ($d = -1.34$ for CareGPT; $d = -1.21$ for ChatGPT). This suggests that peer caregivers more often anchor support in time-bound trajectories (e.g., what happened *recently*, what changed *last week*), while LLMs more often adopt a generalized, decontextualized voice. For example, a peer may situate support with time-anchored framing such as “*Last week, I visited my mom... I don’t want to be the only one...*”, while AI more often uses timeless phrasing such as “*Over time, it becomes more difficult...*”. Socially, temporal anchoring can signal “we are in this together, right now,” while a timeless voice can sound broadly applicable but less situated in a shared trajectory. Other psycholinguistic dimensions, such as **Emotional Tone** and **Cognition & Perception**, were not statistically significant by our criteria ($\eta^2 < 0.10$), and we observed no meaningful differences across responder groups. Notably, physician responses did not significantly differ from peers or AI on any psycholinguistic feature.

4.2.2 Linguistic Structure. Structural variation was most pronounced. **AI responses were markedly more verbose** than human replies, both in sentence count and word count. **ChatGPT** generated the longest replies, averaging nearly 12 more sentences ($d = 2.77$) and 180 additional words ($d = 2.01$) than physicians; these differences were highly significant ($p = 6.87 \times 10^{-58}$). This verbosity can reduce the need for caregivers to ask multiple follow-ups, but it may also increase the burden of

reading and triaging information, i.e., shifting labor from *writing work* to *interpretive work*. It also reflects a socio-technical[102] tradeoff in CSCW: shifting labor from humans to infrastructure [19]. This contrast is visible in typical openings: ChatGPT often provides templated scaffolding such as “*I’m really sorry you’re going through this... Here are a few steps you can consider...*”, whereas physicians more often offer compressed, workflow-oriented guidance such as “*I would recommend an evaluation... Please discuss treatment options and safety planning...*”. In this sense, AI verbosity reshapes caregiving work by reflecting AI’s lack of practical constraints on response length, while potentially increasing caregivers’ interpretive burden and the emotional labor required to process, evaluate, and act on extensive guidance under stress.

Table 2. Statistically significant differences in linguistic features across responder groups (RQ1), based on ANOVA ($p < .001$, $\eta^2 > 0.14$) and post-hoc Tukey’s test. Only group pairs with very large effect sizes ($|d| > 1.2$) are reported. Positive Cohen’s d indicates Group₁ scored higher than Group₂.

Category	Metric	Group ₁	Group ₂	MeanDiff	Cohen’s d
Psycholinguistic	Temporal References	$F = 26.16, \eta^2 = 0.19, p = 3.12 \times 10^{-15}$			
		CareGPT	Online Peer	0.77	-1.34
		ChatGPT	Online Peer	0.71	-1.21
Linguistic Structure	Verbosity (sentence count)	$F = 136.68, \eta^2 = 0.55, p = 6.87 \times 10^{-58}$			
		CareGPT	Online Peer	-8.16	1.63
		CareGPT	Physician	-10.96	2.44
		ChatGPT	Online Peer	-9.15	1.89
		ChatGPT	Physician	-11.95	2.77
	Verbosity (word count)	$F = 95.68, \eta^2 = 0.46, p = 8.85 \times 10^{-45}$			
		CareGPT	Online Peer	-147.51	1.64
		CareGPT	Physician	-170.67	1.84
		ChatGPT	Online Peer	-156.58	1.82
		ChatGPT	Physician	-179.74	2.01
	Readability (Dale-Chall)	$F = 103.76, \eta^2 = 0.48, p = 1.48 \times 10^{-47}$			
		CareGPT	Online Peer	2.37	-3.20
		ChatGPT	Online Peer	2.15	-2.87
		Online Peer	Physician	-2.03	1.70
	Complexity (Depth)	$F = 21.05, \eta^2 = 0.16, p = 1.62 \times 10^{-12}$			
		CareGPT	Online Peer	-0.35	1.20
	Lexical Diversity (TTR)	$F = 82.70, \eta^2 = 0.43, p = 4.34 \times 10^{-40}$			
		CareGPT	Online Peer	0.06	-1.68
		CareGPT	Physician	0.03	-1.56
		ChatGPT	Online Peer	0.06	-1.70
		ChatGPT	Physician	0.03	-1.60
	Repetition (ratio)	$F = 97.47, \eta^2 = 0.47, p = 2.10 \times 10^{-45}$			
		CareGPT	Online Peer	-0.14	1.85
		CareGPT	Physician	-0.15	1.69
		ChatGPT	Online Peer	-0.15	2.06
		ChatGPT	Physician	-0.16	1.86
Linguistic Style	Formality (RoBERTa)	$F = 44.69, \eta^2 = 0.29, p = 2.50 \times 10^{-24}$			
		CareGPT	Online Peer	-0.20	2.50
		ChatGPT	Online Peer	-0.17	1.74
	Analytic/Narrative Style (CDI)	$F = 46.77, \eta^2 = 0.30, p = 2.78 \times 10^{-25}$			
		CareGPT	Online Peer	-0.60	1.76
		ChatGPT	Online Peer	-0.59	1.70

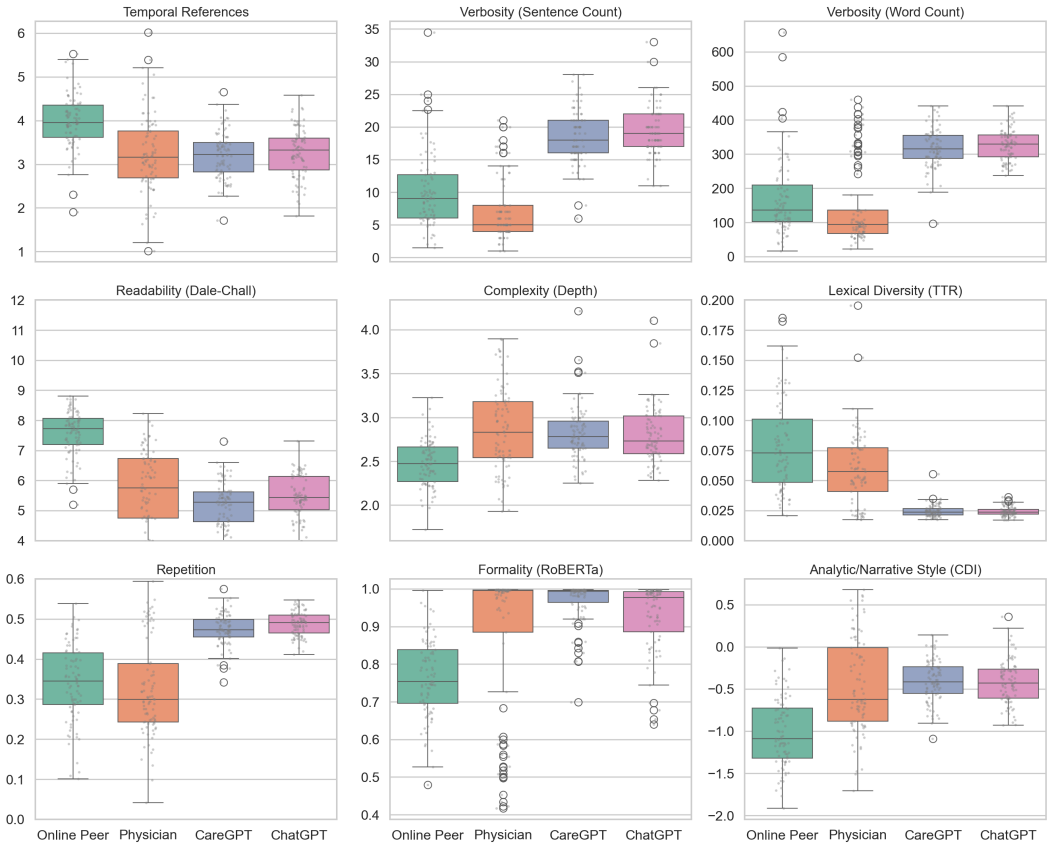


Fig. 3. Distribution of key linguistic style features across responder groups, including verbosity, formality, lexical diversity, syntactic structure, and use of temporal references.

Readability, assessed using the Dale-Chall readability index, also varied notably. AI-generated responses were significantly easier to read than those of **online peer caregivers**, with **CareGPT** and **ChatGPT** both about two points lower (CareGPT: $d = -3.20$; ChatGPT: $d = -2.87$). Because the Dale-Chall index penalizes unfamiliar vocabulary, this suggests that peer-authored responses more often use colloquialisms, shorthand, or community-specific phrasing that may feel socially resonant even when scored as “difficult.” In contrast, AI and physician responses tend to adopt clearer, more standardized phrasing, which may enhance accessibility at the cost of lived specificity. These differences indicate that caregivers may communicate using shared, in-group community-specific language that AI systems and clinicians may not fully recognize (e.g., “DW/DH,” “POA/DPOA,” “MCF”), yet mismatches in vernacular across caregivers could potentially lead to misunderstandings when advice is framed in unfamiliar terms.

Syntactic complexity, operationalized by average dependency parse tree depth, was modestly higher for **CareGPT** compared to **online peer caregivers** ($p = 1.62 \times 10^{-12}$; $d = 1.20$). This indicates that CareGPT pack conditions, rationale, and recommendations into layered sentences (e.g., “Therefore, it’s crucial to safeguard...” and “If possible, consider hiring...”), which can read as “professional,” but does not necessarily translate into greater clarity or engagement like online peers did (e.g., “DH doesn’t have a diagnosis...”).

Lexical diversity, measured via the Type-Token Ratio (TTR), showed a significant drop in AI responses relative to human-written ones. Both **CareGPT** and **ChatGPT** exhibited notably lower TTR values compared to peers and physicians (all $d < -1.56$; $p = 4.34 \times 10^{-40}$), consistent with repeated advice templates within a single reply (e.g., recurring stems such as “consider,” “important,” “help,” “support,” and “don’t hesitate” across successive bullet points), whereas human responses more often vary wording through narrative detail and situational specifics.

Repetition, as a distinct metric, also exhibited clear differences. **ChatGPT** was significantly more repetitive than both **online peer caregivers** ($d = 2.06$) and **physicians** ($d = 1.86$), with similar trends observed in **CareGPT** (all $d > 1.69$; $p = 2.10 \times 10^{-45}$). In practice, this often appears as repeated “safety” and “care coordination” framing within the same response, e.g., “...can ensure... ensure... ensure...”. Together, these results highlight that while LLM-generated outputs are longer and structurally polished, they tend to be less lexically diverse and more repetitive than human responses, even as they achieve greater surface readability.

4.2.3 Linguistic Style. Formality and discourse tone differed markedly by response type. **ChatGPT** and **CareGPT** responses were significantly more formal than those from peers (CareGPT vs. Peer: $d = 2.50$), closely aligning with physicians’ tone. This suggests that despite prompt chaining or retrieval, LLMs often revert to professional discourse defaults. This alignment may also reflect the study design, as both AI systems and physicians were given comparable instructions that encouraged empathetic, consultation-style responses, whereas peer caregivers’ replies were produced organically without prompts or guidance.

Discourse framing, measured by the Analytic/Narrative Style (CDI), revealed that peers favored narrative storytelling while AI models leaned toward analytic exposition ($F = 46.77$, $\eta^2 = 0.295$); CareGPT and ChatGPT differed from peers by over $d = 1.69$, emphasizing structured explanation over dynamic storytelling. This contrast is visible in how peers foreground personal trajectories, e.g., “Door locks up high work. I went through this... it was the only way I could sleep...”, while AI foregrounds generalized guidance, e.g., “Here are a few suggestions: 1) ... 2) ... 3) ...”. This difference likely reflects that AI systems do not have personal caregiving experience to draw on, and therefore default to analytic, generalized explanations rather than the lived, time-anchored narratives that peers use to make sense of and communicate care.

Notably, **CareGPT showed no clear stylistic improvement over ChatGPT** in these style metrics, despite narrative prompt enhancements. Overall, RQ1 results show that each responder group enacts care through a different linguistic persona: peers signal relational closeness through narrative and time-bound lived experience, while AI and physicians emphasize legible, analytic, and formal guidance, shaping how empathy, authority, and “work” are distributed in caregiver support interactions. Viewed through a care infrastructure lens, these differences point to a potential division of caregiving labor in which peers contribute experiential grounding, while AI systems and clinicians provide standardized guidance that supports coordination but leaves interpretation and emotional processing to caregivers.

4.3 Emotional and Informational Support Patterns Across responder groups (RQ2)

We next assess how each responder group fulfills two core support functions in this forum: providing **emotional support** and **informational support**. This analysis focuses on six features with statistically significant between-group differences, (ANOVA $p < .001$, $\eta^2 > 0.14$; Table 3): three emotional support metrics (*Politeness*, *Emotion Evaluation*, *Emotional Composition (Caring)*) and three informational support metrics (*Physician Semantic Proximity*, *Relative Score*, *Information Evaluation*). Figure 4 visualizes feature distributions across responder groups, and Figure 5 summarizes secondary emotional tones. Beyond the numerical contrasts, we interpret these patterns as different

Table 3. Statistically significant differences in emotional and informational support features across responder groups (RQ2), based on ANOVA ($p < .001$, $\eta^2 > 0.14$) and post-hoc Tukey's test. Only group pairs with very large effect sizes ($|d| > 1.2$) are reported. Positive Cohen's d indicates Group₁ scored higher than Group₂.

Category	Metric	Group ₁	Group ₂	MeanDiff	Cohen's d
Emotional Support	Politeness	$F = 76.08, \eta^2 = 0.41, p = 1.41 \times 10^{-37}$			
		CareGPT	Online Peer	-0.10	1.53
	Emotion Evaluation	$F = 43.41, \eta^2 = 0.28, p = 9.81 \times 10^{-24}$			
		CareGPT	Online Peer	-0.25	1.35
	Emotional Composition (<i>Caring</i>)	$F = 56.99, \eta^2 = 0.34, p = 8.31 \times 10^{-30}$			
		CareGPT	Online Peer	-0.47	1.74
Informational Support	Physician Semantic Proximity	$F = 70.13, \eta^2 = 0.36, p = 6.11 \times 10^{-25}$			
		CareGPT	Online Peer	-0.19	1.52
	Relative Score	$F = 79.87, \eta^2 = 0.42, p = 5.04 \times 10^{-39}$			
		CareGPT	Online Peer	-0.20	2.08
	Information Evaluation	$F = 84.80, \eta^2 = 0.43, p = 7.22 \times 10^{-41}$			
		CareGPT	Online Peer	-0.21	1.51
	Information Evaluation	ChatGPT	Online Peer	-0.21	1.51
		Online Peer	Physician	0.18	-1.23
		Online Peer	Physician	0.18	-1.23

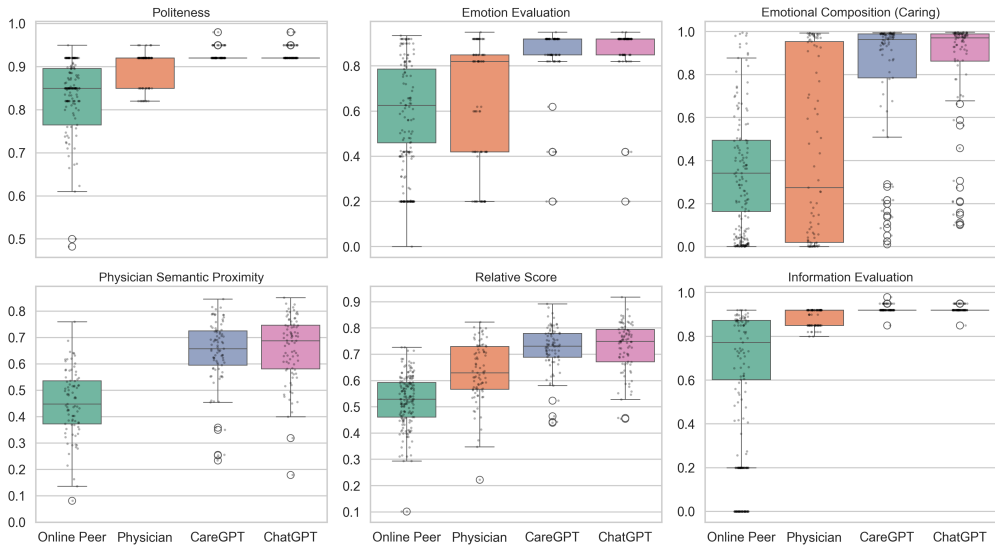


Fig. 4. Boxplots illustrating key linguistic and support-related features across four responder groups: Online Peer, Physician, CareGPT, and ChatGPT. Top row shows **Politeness**, **Emotion Evaluation**, and **Emotional Composition (*Caring*)**, while the bottom row shows **Physician Semantic Proximity**, **Relative Score**, and **Information Evaluation**.

ways of enacting care and trust: standardized, risk-averse empathy and structured guidance (AI) versus relational, context-rich support grounded in lived experience (peers), with physicians often occupying a constrained but authoritative middle.

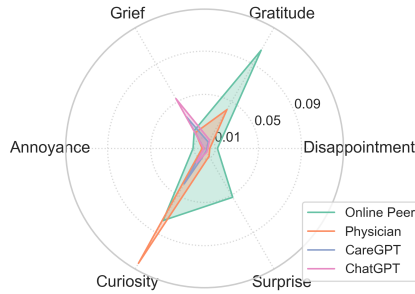


Fig. 5. Radar plot showing the distribution of additional moderate-effect emotional tone dimensions across responder groups.

4.3.1 Emotional Support Behaviors. Across emotional support metrics, AI responses clustered near the top of the scale. Both **ChatGPT** and **CareGPT** scored near the ceiling on *Politeness* (means ≈ 0.93), significantly higher than peers (ChatGPT vs. Peer $d = 1.56$; CareGPT vs. Peer $d = 1.53$; $p = 1.41 \times 10^{-37}$). This difference also appeared in judge confidence: LLM-based politeness confidence scores were 5.0 for ChatGPT and CareGPT, compared to 4.72 for peers and 4.98 for physicians. In practice, AI messages repeatedly opened with templated deference and validation, e.g., *ChatGPT/CareGPT*: “I’m sorry you’re going through this...”, a reliable “polite” surface style that our judges consistently recognized as supportive evidence. These findings raise an open question about whether caregivers value or meaningfully benefit from AI’s highly consistent displays of politeness and empathy, particularly when such expressions are templated and detached from shared lived experience.

A similar pattern emerged in broader *Emotion Evaluation*. AI again scored higher (ChatGPT mean ≈ 0.88 , CareGPT mean ≈ 0.85) than peers (mean ≈ 0.60 ; $p = 9.81 \times 10^{-24}$), and AI confidence remained very high (ChatGPT 4.95; CareGPT 4.92). However, the *type* of empathy differed. AI support was often generalized and reusable, e.g., *ChatGPT*: “I’m sorry to hear of the challenges you are facing... your concern is understandable.”, whereas peer caregivers more often affirmed through situated, relational language that connects to concrete experiences, e.g., *Online Peer*: “I am so sorry for your situation. I have a dear friend whose husband has FTD...” or *Online Peer*: “I am sorry. That sounds really hard...”. Physicians, in contrast, rarely used overt emotional wording, but occasionally blended gentle validation with clinical framing, e.g., *Physician*: “...cognitive decline in a loved one is stressful...” before transitioning into next steps.

The *Emotional Composition (Caring)* metric further amplified this split. ChatGPT and CareGPT scored substantially higher than humans (Peers mean = 0.33, Physicians mean = 0.45, CareGPT ≈ 0.81 with *MeanDiff* = 0.47 vs. peers, $d = 1.74$; ChatGPT *MeanDiff* = 0.51, $d = 2.06$; $p = 8.31 \times 10^{-30}$). This suggests that AI responses consistently trigger “caring” markers according to GoEmotions, even when the message remains generic. Meanwhile, Figure 5 shows that peers (and sometimes physicians) expressed more *varied* emotional tones at moderate effect sizes ($p < 0.001$, $\eta^2 > 0.06$), including gratitude and curiosity, interactional signals that often sustain community relationships. For instance, peers frequently acknowledged others’ efforts (*Online Peer*: “thank you for this thread...”) or asked follow-up questions to invite continued dialogue, whereas AI responses more often relied on polished sympathy statements that can drift toward clinical-sounding grief (e.g., AI: “it’s unfortunate that...”), potentially mismatching the poster’s context.

Overall, RQ2 shows a distinction between *performative* and *relational* empathy: AI reliably produces emotionally “acceptable” and high-scoring support cues, while peers more often enact empathy through contextualized, sometimes messy narratives and reciprocal engagement. Physicians tend to offer limited but sincere validation constrained by a clinical, advice-forward style. Designing emotionally intelligent AI for caregiving communities thus requires more than amplifying politeness templates; it requires calibrating tone to context and supporting the kinds of relational signals that peers use to sustain trust.

4.3.2 Informational Support Patterns. Informational support exhibited clearer alignment between AI and physicians than emotional expression. Across *Physician Semantic Proximity*, *Relative Score*, and *Information Evaluation*, **ChatGPT** and **CareGPT** closely tracked physician responses and exceeded peers (Table 3; Figure 4).

First, *Physician Semantic Proximity* indicates how closely a response resembles physician-style informational content. Both AI systems scored near physicians (means ≈ 0.64 - 0.67), while peers were lower (mean ≈ 0.44 ; $p = 6.11 \times 10^{-25}$; AI vs. Peer $d > 1.5$). This gap reflects how AI and physicians converge on standardized, referral- and resource-oriented guidance, while peers more often embed information inside narrative trajectories. For example, physician- and AI-style responses often take a referral-and-resource form (illustrative phrasing: “I would discuss this with your neurologist/PCP; they can review medications and refer you to local caregiver support and respite resources.”), whereas peer caregivers more often fold similar information into lived timelines (illustrative phrasing: “When my husband started doing this, his doctor adjusted meds, and I also called the Alzheimer’s Association helpline—having a plan for respite made a big difference.”).

Second, *Relative Score* (response-to-question semantic alignment) highlights how tightly a responder stays anchored to the original caregiver question. ChatGPT and CareGPT achieved the highest alignment (means = 0.73 and 0.72), with physicians close behind (mean = 0.70), and peers lower (mean = 0.64 ; $p = 5.04 \times 10^{-39}$; AI vs. Peer $d > 2.0$). In practice, AI and physicians frequently produced stepwise, on-topic answers (e.g., AI: “Here are a few steps you can consider...”; Physician: “...recommend an evaluation... discuss treatment options and safety planning...”), whereas peers often broadened the frame by sharing what worked in their own cases, which can lower semantic similarity while increasing experiential usefulness.

Third, *Information Evaluation* (scored by LLaMA 3.3) captured perceived actionability and informativeness. CareGPT and ChatGPT scored near the upper bound (means > 0.92), significantly higher than peers (mean ≈ 0.69 ; CareGPT vs. Peer $d = 1.51$; ChatGPT vs. Peer $d = 1.51$; $p = 7.22 \times 10^{-41}$), while physicians also scored highly (mean = 0.93). Substantively, this pattern reflects how physicians and LLMs tend to present help in a standardized, stepwise format (e.g., concrete recommendations, options, and next steps), whereas peer caregivers more often prioritize contextualizing the situation—sharing what happened in their own case, expressing solidarity, and offering advice that is embedded in lived constraints rather than distilled into an “action plan.” Confidence ratings reinforced this difference in stance: ChatGPT, CareGPT, and physicians all had mean confidence = 5.0 (std = 0), while peer confidence was lower and more variable (mean = 4.52 , std = 0.78), consistent with peers positioning their guidance as contingent and experience-based rather than authoritative.

These aggregate differences align with how each group “packages” information. AI responses frequently recommended named institutional resources and standardized care pathways (e.g., AI: “...contact the Alzheimer’s Association...” or AI: “...reach out to your primary care doctor or neurologist...”), reflecting a formal guidance persona. CareGPT also foregrounded caregiving logistics common in forum data (e.g., CareGPT: “...adult day care or respite care...”). Physicians similarly emphasized diagnostics, medication, referrals, and coordination, often concisely and with clear

authority. Peers contributed community-specific practical knowledge, frequently referencing legal/coordination realities such as POA and care placement, and embedded such information in lived trajectories (e.g., *Online Peer*: “...Do you have financial POA?...” or stories about navigating symptoms and family dynamics). Even when these narratives drift from the immediate question (lowering semantic alignment), they can surface constraints and “what it actually looks like” details that structured responses omit.

Overall, RQ2 results suggest a division of labor in informational support: AI and physicians deliver consistently organized, high-density guidance that supports cognitive trust (clarity, coverage, actionability), while peers contribute contextualized, community-grounded knowledge that often supports relational trust (shared reality, local constraints, “what worked for me”). Rather than treating “more AI empathy” as the primary design goal, these findings motivate hybrid designs that orchestrate strengths: structured guidance and resources from AI/clinicians paired with peer-sourced experiential nuance that helps caregivers interpret and act on that guidance in real life.

4.4 Topic-Specific Performance of AI vs. Human Responders (RQ3)

4.4.1 *Topic Modeling.* To explore the topics of responses across different responder groups, we applied the BERTopic algorithm to the entire corpus of replies and questions. This unsupervised topic modeling approach yielded 19 coherent topics, determined by BERTopic’s hierarchical clustering algorithm, with each characterized by a distribution of representative words and most representative documents. Figure 6 presents a heatmap of the average topic probability per response type. These values represent the average share of each topic within all responses of a given type, providing a high-level view of content alignment with responder roles.

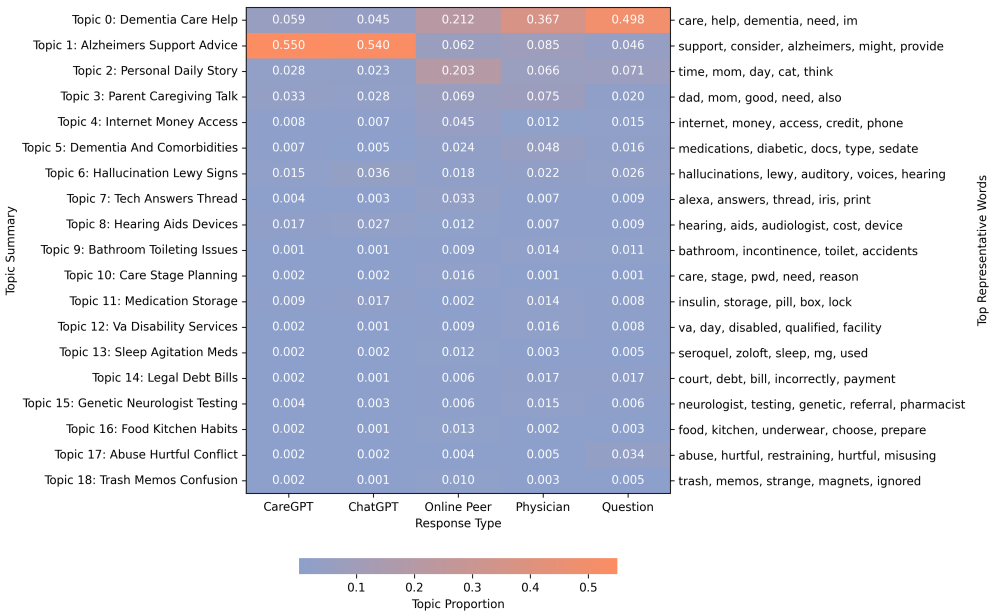


Fig. 6. Average topic proportions across responder groups. Each cell presents the mean topic probability assigned to responses of each type, based on BERTopic outputs. Topics are derived from clustering response content across the full dataset, labeled with brief summaries and representative keywords.

Several patterns are evident from the heatmap. First, Topic 0 (*"Dementia Care Help"*) dominates the question distribution ($mean = 0.498$), consistent with the forum's overall focus on practical caregiving concerns. Physicians also show the highest average proportion for Topic 0 ($mean = 0.367$), reflecting their alignment with general care inquiries. In contrast, both CareGPT and ChatGPT concentrate on Topic 1 (*"Alzheimer's Support Advice"*), a domain focused on direct, prescriptive recommendations. Their strong emphasis on this topic ($mean > 0.54$) suggests a tendency to interpret questions as advice-seeking prompts and to respond with structured, guidance-oriented content. Online Peer responses, on the other hand, are less centered on these general or directive themes and show markedly higher engagement with Topic 2 (*"Personal Daily Story"*, $mean = 0.203$), and also Topic 3 (*"Parent Caregiving Talk"*), and Topic 4 (*"Internet Money Access"*). These patterns point to a peer preference for narrative sharing and personal situational reflections rather than clinical or instructional replies. These findings align with our RQ2 results, which showed that peer caregivers tend to provide emotionally nuanced, community-based support, whereas AI systems emphasize structured politeness and high-confidence factual suggestions. The topic-level distributions further reinforce this divide in support styles.

4.4.2 Topic Proportions. To move beyond average topic probabilities and capture which topics most define each response, we leveraged BERTopic's `topics_per_class` function. This approach identifies the most representative topic for each response document and counts their distribution across responder groups. The resulting bar chart (Figure 7) shows how many responses of each type were primarily associated with each of the top 5 topics (excluding the generic Topic 0). The plot illustrates striking differences in topical scope and specialization across responder groups. **Online Peer responses demonstrate the greatest topical breadth**, with substantial representation across all five shown topics. Notably, over 70 responses were best matched with Topic 2 (*"Personal Daily Story"*), highlighting the peer community's emphasis on lived experience and empathetic sharing. Substantial numbers of peer responses also centered on Topic 3 (*"Parent Caregiving Talk"*), Topic 4 (*"Internet Money Access"*), and Topic 5 (*"Dementia and Comorbidities"*), suggesting a diversity of concerns addressed informally and practically. By contrast, CareGPT responses were almost entirely aligned with Topic 1 (*"Alzheimer's Support Advice"*), indicating a narrow but consistent pattern of structured caregiving guidance. ChatGPT responses were limited to Topic 5, focusing on treatment-specific content, such as medication or symptom management. No other topics appeared as the most frequent or defining themes in either LLM's output, suggesting that current AI systems may follow strongly scripted or training-biased topic patterns rather than adapting to a variety of caregiver concerns.

Physician responses were more topically balanced, appearing across Topics 1–4, but notably absent from Topic 5. This gap may be meaningful; physicians may hesitate to discuss chronic illness comorbidities like diabetes in dementia-focused forums, sticking instead to domain-relevant or directly observable issues such as caregiving advice and financial burden. Lastly, questions posted in the forum showed limited topic diversity, but interestingly had primary alignment with Topic 1 (caregiving advice), as well as Topics 2 (personal story), 4 (finance-related issues), and 5 (treatment queries). This confirms that the original questions span both emotional and practical needs, and reaffirms the need for diverse responder groups to cover this range. Overall, these findings show that LLM-generated responses are topically narrow but predictable, while peer responses are wide-ranging and adaptive to many caregiver concerns. Physicians occupy a middle ground, staying largely within their domain of expertise.

5 Discussion

This work provides a systematic, side-by-side comparison of peer, clinician, and AI responses to the same ADRD caregiver questions. We contribute to CSCW discussions of AI participation in caregiving communities by characterizing role-differentiated care work across a hybrid care infrastructure: peers, clinicians, and AI each enact support through distinct linguistic forms, trust cues, and topical behaviors. Guided by prior work on care infrastructures, emotional and informational labor, and socio-technical trust calibration, we synthesize (i) overarching empirical patterns, (ii) mechanisms that plausibly produce these patterns, and (iii) implications for designing hybrid support arrangements where AI augments rather than replaces human care work.

5.1 Role Differences Across Peers, Clinicians, and AI

We read these results through distributed care infrastructure: differences in linguistic style, emotional expression, and informational framing point to how caregiving labor is allocated and coordinated across people and technologies, not simply to “better” or “worse” writing. Table 4 summarizes the patterns by making the tradeoffs explicit across responder groups, consistent with CSCW accounts of care infrastructures as relational and contingent systems that rely on articulation, sensemaking, and ongoing mutual adjustment rather than uniform service provision [59]. Within

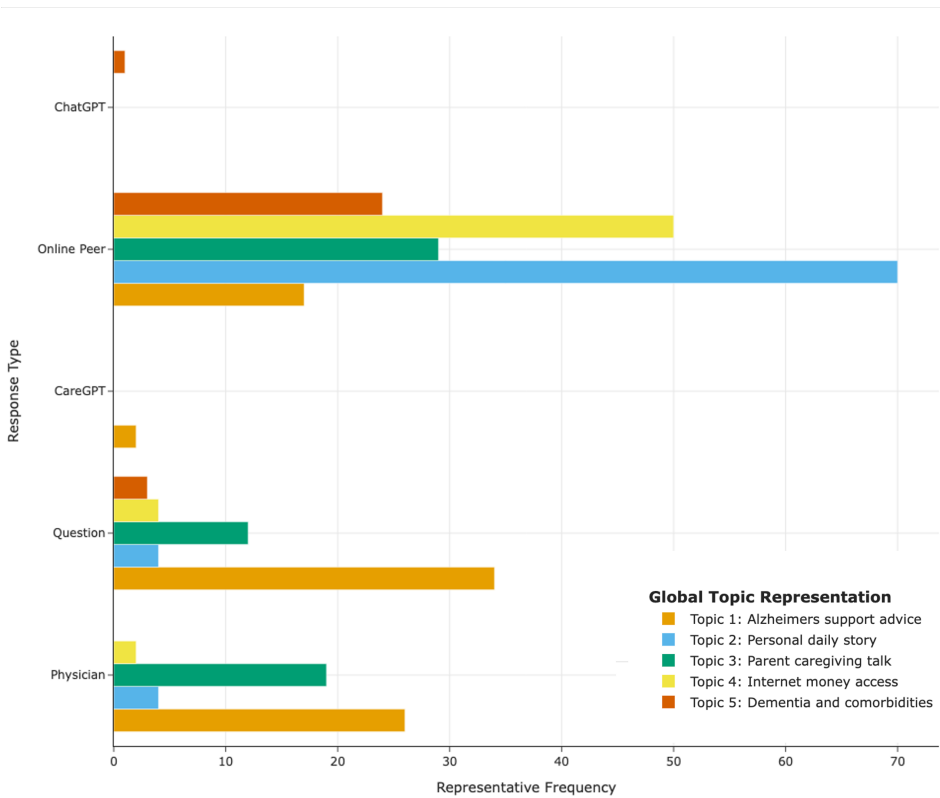


Fig. 7. Distribution of the top five most representative topics across responder groups, based on BERTopic’s `topics_per_class` function. Each bar represents the frequency of responses primarily associated with each of the selected topics (Topics 1–5) for a given responder group.

Table 4. Empirically observed strengths and limitations by responder group, illustrating how support work is distributed across a hybrid care infrastructure.

Responder group	Strengths (observed)	Limitations (observed)	Role in distributed care
Peer caregivers	Time-anchored narratives; community vernacular; broader topic range (e.g., emotional, daily logistics, legal and financial)	Less standardized structure and readability; lower physician-style proximity and information-evaluation	Lived context and relational legitimacy for sense-making
Physicians	Concise, clinically bounded guidance; high physician-style informational content	Less narrative and relational work; emotional support often brief and secondary	Accountable clinical guidance and care pathways
ChatGPT	Highly structured and readable; high information-evaluation and physician proximity; consistently high politeness	Low temporal and narrative grounding; repetitive and templated phrasing; advice-heavy topical focus	On-demand “consultation-style” organization of options and resources
CareGPT	Tracks ChatGPT on informational and emotional metrics; retrieval-conditioned on peer discussions	Minimal measurable separation from ChatGPT; continued formal, advice-forward stance	Reinforces LLM role as structured guidance (not peer-like lived experience)

this framing, AI participation is characterized by standardized, polite, and scalable support that is less tied to lived trajectories and community language. This consistency can lower interactional barriers and make resources easier to access, but it also shifts more interpretive and emotional work onto caregivers, echoing CSCW descriptions of sociotechnical tradeoffs where infrastructure absorbs some labor while generating new demands elsewhere [19]. The coexistence of peer narratives, clinically bounded physician guidance, and AI-mediated support also aligns with social support theory, which treats emotional and informational support as multiple, co-present forms rather than a single support function [112]. These findings therefore caution against treating AI as a standalone solution in affective or high-stakes contexts [73, 74], and motivate hybrid caregiving systems that support coordination across distributed actors and make the limits of AI-mediated empathy visible within the care infrastructure. These patterns indicate that care support is organized through complementary forms of relational and informational labor, rather than through convergence toward a single, idealized style of response. In this sense, linguistic differences matter not only as descriptive features of responses, but as cues through which caregivers may interpret relational stance, legitimacy, and the fit of a response to a given support need.

Across RQ1-RQ3, we observe a consistent division of communicative labor in how care work is enacted through language. First, **peers** disproportionately contribute *narrative, temporally anchored lived-experience support*. They often situate advice in personal trajectories (e.g., “my mom passed away in 2020,” “I took early retirement to care 24/7,” “some days are so much worse than others”), using time markers and first-person storytelling to signal solidarity and shared reality. This aligns with CSCW observations that online health communities value lived experience and being heard as core signals of support and belonging [7, 78, 118], and that such narrative participation can sustain trust and legitimacy within peer infrastructures [88]. Second, **clinicians** tend to provide *concise, advice-forward, clinically bounded guidance*, reflecting professional norms of authority

and clarity [32]. Their support often takes the form of referral and scoped actionability (e.g., “the Alzheimer’s Association has wonderful resources”; “an evaluation of decision-making capacity... will also help for legal decision making”). This style can increase perceived credibility, but it may offer less reciprocal vulnerability and less colloquial warmth than peers, reflecting tensions documented in CSCW health communities between expertise and affirmation [14, 104, 118]. Third, **LLMs (ChatGPT and CareGPT)** consistently enact a *formal, verbose, structured guidance persona*. In grief-support scenarios, the models often open with acknowledgement (e.g., “I’m very sorry for your loss”), then enumerate institutional pathways (e.g., “a 24/7 helpline,” “local chapters,” “bereavement services”), and close with generalized reassurance (e.g., “you are not alone”). This echoes findings that LLMs default to essay-like, highly structured support styles [114, 134]. In our setting, this yields consistently readable and “acceptable” surface cues, yet it can also read as templated or socially misaligned with community norms.

These three voices represent complementary but non-substitutable forms of care work: peers offer relational grounding through time-anchored narratives; clinicians offer bounded authority and clinically accountable pathways; AI offers scalable structure, coverage, and consistently on-topic scaffolding. The theoretical insight is therefore not merely that these responder groups differ, but that they enact distinct forms of caregiving labor and legitimacy within the same support ecology. Importantly, our results suggest that AI’s strongest observable contribution is not “becoming human-like,” but producing consistent, organized, action-oriented support that reduces search and synthesis burden, while human voices (peer and clinician) supply situated judgment, accountability, and community legitimacy.

Design Implications. These findings motivate design considerations for systems that account for the complementary roles of peers, clinicians, and AI, rather than treating AI as a replacement for human care work. Systems should maintain the legitimacy of peer narratives (including their willingness to surface non-clinical constraints such as POA, guardianship, or insurance risks that arise in “unstable” family environments), protect clinicians’ bounded authority and uncertainty signaling, and position AI as a transparent infrastructure layer that provides scaffolding (e.g., resource routing, checklists, summarization) without competing with humans for identity or interpersonal status [73]. Rather than optimizing for a single “best” response, CSCW designs can specify which participant is best suited to provide which type of support, and how these contributions are presented to caregivers.

5.2 Sources of Differences in Language, Trust Cues, and Topic Scope

We interpret these patterns through three related reasons that shape how each actor “does care”.

(1) Norms and constraints produce distinct authenticity signals. Peers operate within a community norm where credibility is often earned through lived experience, temporal anchoring, and reciprocal disclosure [7, 78, 118]. In our corpus, peers frequently mark legitimacy by narrating concrete trajectories and recent incidents (e.g., “over the last few years it has been getting worse,” “last week I came home and her car was left on overnight”), positioning advice as emerging from ongoing, situated responsibility rather than abstract expertise. Clinicians operate under professional constraints that prioritize clinical accuracy and appropriate scope [32], which can limit narrative participation even when emotional support is needed [14, 104, 118]. Their supportive language often appears as bounded empathy plus action-oriented triage (e.g., acknowledging that dementia diagnoses are hard, then shifting to self-care, primary-care follow-up, therapy, medication options, and crisis escalation when needed). In contrast, LLMs can generate empathy at scale, but that empathy often appears as stable templates and polished phrasing [114, 134]. This makes AI “reliably polite” yet potentially vulnerable to perceived inauthenticity, a form of artificial empathy that can feel ungrounded in shared experience [113].

(2) Trust has more than one dimension, and peers, clinicians, and AI tend to rely on different cues. Our findings are consistent with the distinction between emotional trust (“they understand me”) and cognitive/epistemic trust (“they know what they are talking about”) [62]. Peers often earn emotional trust through contextualized empathy and pragmatic, experience-linked warnings (e.g., noticing safety-relevant slips like leaving a car running, or anticipating family disagreement about symptom severity). Clinicians and AI more strongly signal cognitive trust via structured, advice-forward content and expert-like framing, including referrals, formal coping strategies, and clear boundary conditions for urgent escalation. However, pushing AI too far toward human-like affect can trigger discomfort akin to an “uncanny valley” of empathy [85], while under-expressive AI may be dismissed as cold. This tension motivates explicit trust-calibration features (e.g., transparency, sourcing, and appropriate disclaimers) that help caregivers interpret AI behavior [49, 116] and develop strategies under uncertainty [75].

(3) Topic scope reflects how each responder frames what the question is asking for. Peers treat caregiving as a holistic life condition; their topical breadth reflects the way caregiving problems spill across emotional, social, legal/financial, and medical domains [86]. This also defines “expertise boundary”: peers may explicitly caution against informal legal advice and redirect caregivers toward certified elder-law resources (e.g., emphasizing that guardianship is costly/contested and recommending CELA/NAELA directories) while still sharing forum-grounded heuristics. Clinicians remain more bounded by domain remit [32]. AI, especially under an office-framed instruction, tends to interpret posts as requests for prescriptive guidance, reinforcing a narrower, consultation-like posture [114, 134]. In other words, topic breadth is not only “what knowledge one has,” but *how one frames the interactional purpose* of the post.

These reasons indicate that differences across peers, clinicians, and AI are not merely stylistic. Rather, they reflect how different responder groups enact support roles under different norms, constraints, and expectations for what counts as credible or authentic care. This is why simply making AI “more empathetic” is unlikely to reproduce the relational credibility of peer infrastructures [113], and why hybrid designs must support ongoing trust calibration rather than assuming stable trust [75].

5.3 Design Implications and Interpreting CareGPT's Similarity to ChatGPT

Our results suggest a practical design direction: support systems should be designed not only to combine what peers, clinicians, and AI each do well, but also to preserve their distinct roles rather than treat them as interchangeable. Peers, clinicians, and AI contribute different kinds of value. Peers bring lived constraints and relational resonance. Clinicians bring bounded authority and accountability. AI provides scalable structure and broad informational support [62]. The design question is how to present these voices together in a way that preserves their distinct roles.

This also matters for trust. Interfaces need to help caregivers interpret support, rather than assuming trust is something a system can simply *earn*. Caregivers develop strategies and folk theories for judging advice under uncertainty [75]. Hybrid systems should therefore make AI status and limits clear through trust-calibration features such as transparency about data resources, links to authoritative references, and simple disclaimers that distinguish general guidance from clinician-appropriate decisions [49, 116]. At the same time, systems should avoid over-claiming emotional competence. Overly human-like empathy can feel uneasy or inauthentic when it is not grounded in shared experience [85, 113]. These supports help caregivers decide when to use AI as a first pass, when to defer to a clinician, or when to seek peer input that accounts for family dynamics and practical constraints. Prior work also shows that users often choose AI precisely because of its non-human traits [96]. This aligns with our finding that AI's strongest contribution

lies not in mimicking human empathy but in offering scalable availability and structured guidance that complements rather than replaces human care work.

This complementarity is visible when the same caregiver question receives different kinds of help. Clinicians tend to respond with concise, clinically bounded guidance, often pointing to evaluations, referrals, or safety steps, consistent with professional expectations around scope and responsibility [32]. Peer caregivers more often respond by describing what similar situations look like in their own lives over time, including how grief, exhaustion, and family tension accumulate alongside daily tasks. These experience-anchored narratives do more than convey information. They signal recognition, legitimacy, and solidarity in ways that support communities consistently value [7, 78, 118]. AI responses follow a different pattern. They typically organize support into structured, step-by-step suggestions paired with polite reassurance and resource pointers, producing replies that are easy to follow and consistently read as empathetic [114, 134]. These differences do not point to a single ideal response style, but show how caregiving support is built through multiple forms of help: AI surfaces options and common pathways, clinicians set medical boundaries and accountability, and peers translate guidance into lived realities while sustaining participation in the community [86].

Design Implications. First, in caregiving contexts, RAG should be treated as a grounding component rather than a guarantee of peer-like relational support. If the goal is to better align with community norms, the system may require instructions tailored to specific domains, example-based response shaping, or clearer role definitions to encourage it to consider specific contexts (e.g., time, relationships, and constraints) rather than providing generic advice [114, 134]. Second, a promising path is to use AI in a support role for human responders and to optimize it for what it does best: structured, instrumental support. For example, AI can draft structure and resource-oriented content, while peers and clinicians add lived specificity, authentic empathy, boundary-setting (e.g., when certified legal or clinical advice is required), and accountability. This preserves transparent non-human status while augmenting, rather than displacing, the emotional and informational labor embedded in caregiving communities [73].

CareGPT's similarity to ChatGPT is itself a design result. Across RQ1-RQ3, retrieval-plus prompt-injection alone did not measurably shift the model's communicative stance. Even when community discourse was available, the system often retained the same consultation-like, polished guidance role. This indicates that adding retrieved peer text is not sufficient to align model behavior with community practice. If the goal is to change how the model frames its role, stronger signals are needed, such as role framing and interactional expectations that shape what counts as an appropriate response [115]. RAG may improve topical grounding, but it does not by itself produce peer-like narrative positioning or relational credibility.

5.4 Limitations and Future Work

This study has several limitations that motivate future research. First, our data are drawn from a single online community; caregiving norms vary across platforms, cultures, and languages. Replication across other caregiving and health communities is needed. Second, our AI responses reflect models and configurations available at the time of study; rapid model evolution may change observed patterns, reinforcing the need for ongoing evaluation. Third, although we validated LLM-as-judge metrics against human ratings, automated judgments cannot fully capture caregivers' subjective experience of being supported. Future work should incorporate caregiver-facing evaluations (e.g., preference testing, perceived emotional safety, and trust calibration effects) and examine how interface scaffolds influence uptake. Finally, we focus on message-level comparisons rather than longitudinal trajectories. Future longitudinal and ethnographic work could theorize how hybrid care infrastructures evolve over time: when AI stabilizes practices, when it introduces new burdens

or dependencies, and how communities negotiate governance and accountability as non-human actors perform visible care work.

6 Conclusion

In this study, we compared how *peer caregivers*, *physicians*, *ChatGPT*, and *CareGPT* responded to the same set of Alzheimer's disease caregiver questions, offering the first controlled benchmark of human and AI voices in this sensitive context. This allows us to move beyond hypothetical debates to concrete evidence of what AI currently does well and where it falls short. We found that peer caregivers contribute empathy through lived experiences, physicians provide trusted accuracy, and AI offers content consistency. While AI models can emulate human knowledge and polite tone, they lack the personal nuance and contextual adaptability of peers and physicians. These findings contribute to CSCW theory by showing that peer caregivers, physicians, and AI systems occupy distinct but complementary roles within caregiving support infrastructures, and that these roles reflect different forms of relational, informational, and accountable support labor. The analytical pipeline is also transferable to other domains in future work. Overall, our findings suggest that human-AI collaboration in caregiving is best seen as integration rather than replacement.

These findings point to sociotechnical systems that strengthen, rather than undermine, the values of caregiver communities and domain experts by using AI where it can provide support credibly, while preserving space for peer empathy and clinician accountability.

Acknowledgments

Sean S. Huang and James S. Powers were partly supported by the HRSA Geriatric Workforce Enhancement Program, Award No. 6 U1QHP33085-05-07.

Portions of this manuscript were revised for clarity and presentation with assistance from OpenAI's ChatGPT (GPT-5.2; accessed January 2026). The tool was used only to rephrase and reorganize author-written text based on author-provided ideas and previous drafts. All analyses, interpretations, and conclusions remain those of the authors, who take full responsibility for the accuracy and integrity of the content.

References

- [1] Mahyar Abbasian, Elahe Khatibi, Iman Azimi, David Oniani, Zahra Shakeri Hossein Abad, Alexander Thieme, Ram Sriram, Zhongqi Yang, Yanshan Wang, Bryant Lin, et al. 2024. Foundation metrics for evaluating effectiveness of healthcare conversations powered by generative AI. *NPJ Digital Medicine* 7, 1 (2024), 82.
- [2] Fadi Aljamaan, Mohamad-Hani Temsah, Ibraheem Altamimi, Ayman Al-Eyadhy, Amr Jamal, Khalid Alhasan, Tamer A Mesallam, Mohamed Farahat, Khalid H Malki, et al. 2024. Reference hallucination score for medical artificial intelligence chatbots: development and usability study. *JMIR Medical Informatics* 12, 1 (2024), e54345.
- [3] Kimberley R Allison, Pandora Patterson, Daniel Guilbert, Melissa Noke, and Olga Husson. 2021. Logging on, reaching out, and getting by: A review of self-reported Psychosocial impacts of online peer support for people impacted by cancer. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–35.
- [4] Joshua Au Yeung, Zeljko Kraljevic, Akish Luintel, Alfred Balston, Esther Idowu, Richard J Dobson, and James T Teo. 2023. AI chatbots not yet ready for clinical use. *Frontiers in digital health* 5 (2023), 1161098.
- [5] John W Ayers, Adam Poliak, Mark Dredze, Eric C Leas, Zechariah Zhu, Jessica B Kelley, Dennis J Faix, Aaron M Goodman, Christopher A Longhurst, Michael Hogarth, et al. 2023. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA internal medicine* 183, 6 (2023), 589–596.
- [6] Robert S Baron, Carolyn E Cutrona, Daniel Hicklin, Daniel W Russell, and David M Lubaroff. 1990. Social support and immune function among spouses of cancer patients. *Journal of personality and social psychology* 59, 2 (1990), 344.
- [7] Carolina Figueras Bates. 2023. Empathy in online mental health communities. In *Stance. Inter Subjectivity and Identity in Discourse*. Peter Lang Berlin, 353–377.
- [8] Jennifer L Bevan, Kathryn E Rogers, Nicole F Andrews, and Lisa Sparks. 2012. Topic avoidance and negative health perceptions in the distant family caregiving context. *Journal of Family Communication* 12, 4 (2012), 300–314.

- [9] Karthik S Bhat, Amanda K Hall, Tiffany Kuo, and Neha Kumar. 2023. "We are half-doctors": Family Caregivers as Boundary Actors in Chronic Disease Management. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–29.
- [10] Karthik S. Bhat, Mohit Jain, and Neha Kumar. 2021. Infrastructuring Telehealth in (In)Formal Patient-Doctor Contexts. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–28. doi:10.1145/3476064
- [11] Prakhar Biyani, Cornelia Caragea, Prasenjit Mitra, and John Yen. 2014. Identifying emotional and informational support in online health communities. In *Proceedings of COLING 2014, the 25th international conference on computational linguistics: technical papers*. 827–836.
- [12] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3, Jan (2003), 993–1022.
- [13] Ryan L Boyd, Ashwini Ashokkumar, Sarah Seraj, and James W Pennebaker. 2022. The development and psychometric properties of LIWC-22. *Austin, TX: University of Texas at Austin* 10 (2022), 1–47.
- [14] Robin N Brewer, Sarita Schoenebeck, Kerry Lee, and Haripriya Suryadevara. 2021. Challenging passive social media use: older adults as caregivers online. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–20.
- [15] Matthew E Brock and Heartley B Huber. 2017. Are peer support arrangements an evidence-based practice? A systematic review. *The Journal of Special Education* 51, 3 (2017), 150–163.
- [16] Francisco Maria Calisto. 2024. Human-centered design of personalized intelligent agents in medical imaging diagnosis. *Unpublished*. <https://doi.org/10.13140/RG.2.28353.33126> (2024).
- [17] Francisco Maria Calisto, João Maria Abrantes, Carlos Santiago, Nuno J Nunes, and Jacinto C Nascimento. 2025. Personalized explanations for clinician-AI interaction in breast imaging diagnosis by adapting communication to expertise levels. *International Journal of Human-Computer Studies* 197 (2025), 103444.
- [18] Francisco Maria Galamba Ferrari Calisto, João Gabriel De Matos Fernandes, Margarida Morais, Carlos Santiago, João Maria Veigas Abrantes, Nuno Jardim Nunes, and Jacinto Nascimento. 2023. Assertiveness-based Agent Communication for a Personalized Medicine on Medical Imaging Diagnosis. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 13. doi:10.1145/3544548.3580682
- [19] Lindsey Cameron, Angele Christin, Michael Ann DeVito, Tawanna R. Dillahunt, Madeleine Elish, Mary Gray, Rida Qadri, Noopur Raval, Melissa Valentine, and Elizabeth Anne Watkins. 2021. "this seems to work": Designing technological systems with the algorithmic imaginations of those who labor. In *Extended abstracts of the 2021 CHI conference on human factors in computing systems*. 1–5.
- [20] Siyuan Chen, Mengyue Wu, Kenny Q Zhu, Kunyao Lan, Zhiling Zhang, and Lyuchun Cui. 2023. LLM-empowered chatbots for psychiatrist and patient simulation: application and evaluation. *arXiv preprint arXiv:2305.13614* (2023).
- [21] Yunan Chen, Victor Ngo, and Sun Young Park. 2013. Caring for caregivers: designing for integrality. In *Proceedings of the 2013 conference on Computer supported cooperative work*. 91–102.
- [22] Yirong Chen, Xiaofen Xing, Jingkai Lin, Huimin Zheng, Zhenyu Wang, Qi Liu, and Xiangmin Xu. 2023. SoulChat: Improving LLMs' empathy, listening, and comfort abilities through fine-tuning with multi-turn empathy conversations. *arXiv preprint arXiv:2311.00273* (2023).
- [23] Hyojin Chin, Hyeonho Song, Gumhee Baek, Mingi Shin, Chani Jung, Meeyoung Cha, Junghoi Choi, and Chiyoungh Cha. 2023. The potential of chatbots for emotional support and promoting mental well-being in different cultures: mixed methods study. *Journal of Medical Internet Research* 25 (2023), e51712.
- [24] Soyoung Choi and JooYoung Seo. 2020. An exploratory study of the research on caregiver depression: using bibliometrics and LDA topic modeling. *Issues in mental health nursing* 41, 7 (2020), 592–601.
- [25] Jacob Cohen. 2016. A power primer. (2016).
- [26] Connie S Cole, Ashley Dafoe, Caroline K Tietbohl, Sarah R Jordan, Amy G Huebschmann, Hillary D Lum, and Christine D Jones. 2023. Care challenges of home health patients living with dementia: a pathway forward with palliative care. *BMC palliative care* 22, 1 (2023), 122.
- [27] John H Connolly and Ernest A Edmonds. 2012. *CSCW and artificial intelligence*. Springer Science & Business Media.
- [28] Rosemary Daynes-Kearney and Stephen Gallagher. 2023. Online support groups for family caregivers: scoping review. *Journal of medical Internet research* 25 (2023), e46858.
- [29] Mariano de Rivero, Cristhiam Tirado, and Willy Ugarte. 2021. FormalStyler: GPT based Model for Formal Style Transfer based on Formality and Meaning Preservation.. In *KDIR*. 48–56.
- [30] Ulrike Deetjen and John A Powell. 2016. Informational and emotional elements in online support groups: a Bayesian approach to large-scale content analysis. *Journal of the American Medical Informatics Association* 23, 3 (2016), 508–513.
- [31] Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. *arXiv preprint arXiv:2005.00547* (2020).
- [32] Xianghua Ding, Xinning Gui, Xiaojuan Ma, Zhaofer Ding, and Yunan Chen. 2020. Getting the Healthcare We Want: The Use of Online" Ask the Doctor" Platforms in Practice. In *Proceedings of the 2020 CHI Conference on Human Factors*

in *Computing Systems*. 1–13.

- [33] Michaelanne Dye. 2021. Un Grano de Arena: Infrastructural Care, Social Media Platforms, and the Venezuelan Humanitarian Crisis. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW3 (2021), 1–28. doi:10.1145/3432946
- [34] Tatiana Ermakova, Annika Baumann, Benjamin Fabian, and Hanna Krasnova. 2014. Privacy policies and users' trust: does readability matter?. In *AMCIS*.
- [35] Camille Fauchon, I Faillenot, AM Perrin, C Borg, Vincent Pichot, Florian Chouchou, Luis Garcia-Larrea, and Roland Peyron. 2017. Does an observer's empathy influence my pain? Effect of perceived empathetic or unempathetic support on a pain test. *European Journal of Neuroscience* 46, 10 (2017), 2629–2637.
- [36] Franz Faul, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. 2007. G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods* 39, 2 (2007), 175–191.
- [37] Andrea Ferrario and Michele Loi. 2022. How explainability contributes to trust in AI. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*. 1457–1466.
- [38] Michelle L Foster, Chinenye Egwuonwu, Erin Vernon, Mohammad Alarifi, and M Courtney Hughes. 2025. Informal Caregivers Connecting on the Web: Content Analysis of Posts on Discussion Forums. *JMIR Formative Research* 9, 1 (2025), e64757.
- [39] Uğur Genç and Himanshu Verma. 2024. Situating Empathy in HCI/CSCW: A Scoping Review. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW2 (2024), 1–37.
- [40] Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences* 120, 30 (2023), e2305016120.
- [41] Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794* (2022).
- [42] Igor Grossmann, Harrison Oakes, and Henri C Santos. 2019. Wise reasoning benefits from emodiversity, irrespective of emotional intensity. *Journal of Experimental Psychology: General* 148, 5 (2019), 805.
- [43] Inbal Halevi Hochwald, Daniella Arieli, Zorian Radomyslsky, Yehuda Danon, and Rachel Nissanholtz-Gannot. 2022. Emotion work and feeling rules: Coping strategies of family caregivers of people with end stage dementia in Israel—A qualitative study. *Dementia* 21, 4 (2022), 1154–1172.
- [44] Heidi E Hamilton, Cynthia Gordon, Meaghan Nelson, and Marian Kerbleski. 2006. Physicians, nonphysician healthcare providers, and patients communicating in hepatitis C: an in-office sociolinguistic study. *Gastroenterology Nursing* 29, 5 (2006), 364–370.
- [45] Jing Han, Guifang Guo, and Li Hong. 2022. Impact of professionally facilitated peer support for family carers of people with dementia in a WeChat virtual community. *Journal of Telemedicine and Telecare* 28, 1 (2022), 68–76.
- [46] Wordh Ul Hasan, Kimia Tuz Zaman, Xin Wang, Juan Li, Bo Xie, and Cui Tao. 2024. Empowering Alzheimer's caregivers with conversational AI: A novel approach for enhanced communication and personalized support. *npj Biomedical Innovations* 1, 1 (2024), 1–10.
- [47] Scott A Hawkins, Stephen J Hoch, and Joan Meyers-Levy. 2001. Low-involvement learning: Repetition and coherence in familiarity and belief. *Journal of Consumer Psychology* 11, 1 (2001), 1–11.
- [48] Alice Y Hua, Jenna L Wells, Casey L Brown, and Robert W Levenson. 2021. Emotional and cognitive empathy in caregivers of people with neurodegenerative disease: Relationships with caregiver mental health. *Clinical Psychological Science* 9, 3 (2021), 449–466.
- [49] Sean S Huang, Qingyuan Song, Kimberly J Beiting, Maria C Duggan, Kristin Hines, Harvey Murff, Vania Leung, James Powers, TS Harvey, Bradley Malin, et al. 2024. Fact check: assessing the response of ChatGPT to Alzheimer's disease myths. *Journal of the American Medical Directors Association* 25, 10 (2024), 105178.
- [50] Jina Huh and Wanda Pratt. 2014. Weaving clinical expertise in online health communities. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1355–1364.
- [51] Xiaoyun Jia, Yan Pang, and Liangni Sally Liu. 2021. Online health information seeking behavior: a systematic review. In *Healthcare*, Vol. 9. MDPI, 1740.
- [52] Xiao-Ling Jin, Zhongyun Zhou, Matthew KO Lee, and Christy MK Cheung. 2013. Why users keep answering questions in online question answering communities: A theoretical and empirical investigation. *International Journal of Information Management* 33, 1 (2013), 93–104.
- [53] TANG Jingwei. 2013. Pragmatic functions of hedges and politeness principles. *International Journal of Applied Linguistics and English Literature* 2, 4 (2013), 155–160.
- [54] Douglas Johnson, Rachel Goodman, J Patrinely, Cosby Stone, Eli Zimmerman, Rebecca Donald, Sam Chang, Sean Berkowitz, Avni Finn, Eiman Jahangir, et al. 2023. Assessing the accuracy and reliability of AI-generated medical responses: an evaluation of the Chat-GPT model. *Research square* (2023), rs–3.
- [55] Jazette Johnson, Vitica Arnold, Anne Marie Piper, and Gillian R Hayes. 2022. "It's a lonely disease": Cultivating Online Spaces for Social Support among People Living with Dementia and Dementia Caregivers. *Proceedings of the*

ACM on Human-Computer Interaction 6, CSCW2 (2022), 1–27.

- [56] Jazette Johnson, Rebecca W Black, and Gillian R Hayes. 2020. Roles in the discussion: an analysis of social support in an online forum for people with dementia. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–30.
- [57] Kazi Sinthia Kabir, Johanna Cohoon, John R Lund, Jacqueline M Chen, AJ Metz, and Jason Wiese. 2025. Balancing Care Responsibilities with Remote Work. *Proceedings of the ACM on Human-Computer Interaction* 9, 1 (2025), 1–21.
- [58] Hyo Shin Kang, Woojae Myung, Duk L Na, Seong Yoon Kim, Jae-Hong Lee, Seol-Heui Han, Seong Hye Choi, SangYun Kim, Seonwoo Kim, and Doh Kwan Kim. 2014. Factors associated with caregiver burden in patients with Alzheimer's disease. *Psychiatry investigation* 11, 2 (2014), 152.
- [59] Naveena Karusala, Victoria G., Shirley Yan, and Richard Anderson. 2023. Unsettling Care Infrastructures: From the Individual to the Structural in a Digital Maternal and Child Health Intervention. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, 1–16. doi:10.1145/3544548.3581553
- [60] Naveena Karusala, Azra Ismail, Karthik S Bhat, Aakash Gautam, Sachin R Pendse, Neha Kumar, Richard Anderson, Madeline Balaam, Shaowen Bardzell, Nicola J Bidwell, et al. 2021. The future of care work: towards a radical politics of care in CSCW research and practice. In *Companion Publication of the 2021 Conference on Computer Supported Cooperative Work and Social Computing*. 338–342.
- [61] Aman Khullar, Nikhil Nalin, Abhishek Prasad, Ann John Mampilli, and Neha Kumar. 2025. Nurturing Capabilities: Unpacking the Gap in Human-Centered Evaluations of AI-Based Systems. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–18. doi:10.1145/3706598.3713278
- [62] Nakyung Kyung and Hyeokkoo Eric Kwon. 2022. Rationally trust, but emotionally? The roles of cognitive and affective trust in laypeople's acceptance of AI for preventive care operations. *Production and Operations Management* (2022), 10591478231225891.
- [63] C Lai and JC Chung. 2007. Caregivers' informational needs on dementia and dementia care. *Asian J Gerontol Geriatr* 2 (2007), 78–87.
- [64] Catherine Penny Hinson Langford, Juanita Bowsher, Joseph P Maloney, and Patricia P Lillis. 1997. Social support: a conceptual analysis. *Journal of advanced nursing* 25, 1 (1997), 95–100.
- [65] Virginia Langum and Kirk PH Sullivan. 2017. Writing academic English as a doctoral student in Sweden: Narrative perspectives. *Journal of Second Language Writing* 35 (2017), 20–25.
- [66] Eric B Larson and Xin Yao. 2005. Clinical empathy as emotional labor in the patient-physician relationship. *Jama* 293, 9 (2005), 1100–1106.
- [67] Tove Larsson and Henrik Kaatari. 2020. Syntactic complexity across registers: Investigating (in) formality in second-language writing. *Journal of English for Academic Purposes* 45 (2020), 100850.
- [68] Amanda Lazar and Emma E Dixon. 2019. Safe enough to share: Setting the dementia agenda online. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–23.
- [69] Carrie Leana, Jirs Meuris, and Cait Lambertson. 2018. More than a feeling: The role of empathetic care in promoting safety in health care. *ILR Review* 71, 2 (2018), 394–425.
- [70] Zhiying Li, Xiaonan Wu, Hongxun Jiang, and Xun Liang. 2025. Bridging Expertise: Doctor Recommendations for Cross-Disciplinary Collaborations in Online Medical Consultations. *Proceedings of the ACM on Human-Computer Interaction* 9, 2 (2025), 1–32.
- [71] Tung-Cheng Lin, Lih-Lian Hwang, and Yung-Jye Lai. 2017. Effects of argument quality, source credibility and self-reported diabetes knowledge on message attitudes: An experiment using diabetes related messages. *Health Information & Libraries Journal* 34, 3 (2017), 225–235.
- [72] Jin Liu. 2020. A Pragmatic Analysis of Hedges from the Perspective of Politeness Principle. *Theory and Practice in Language Studies* 10, 12 (2020), 1614–1620.
- [73] Jiaying" Lizzy" Liu, Shuer Zhuo, Xingyu Li, Andrew Dillon, Noura Howell, Angela DR Smith, and Yan Zhang. 2025. From Regulation to Support: Centering Humans in Technology-Mediated Emotion Intervention in Care Contexts. *Proceedings of the ACM on Human-Computer Interaction* 9, 7 (2025), 1–32. doi:10.1145/3757605
- [74] Yihe Liu, Anushk Mittal, Diyi Yang, and Amy Bruckman. 2022. Will AI console me when I lose my pet? Understanding perceptions of AI-mediated email writing. In *Proceedings of the 2022 CHI conference on human factors in computing systems*. 1–13.
- [75] Claudia López, Alexandra Davidoff, Francisca Luco, Mónica Humeres, and Teresa Correa. 2025. Users' experiences of algorithm-mediated public services: Folk theories, trust, and strategies in the Global South. *International Journal of Human-Computer Interaction* 41, 8 (2025), 5039–5056. doi:10.1080/10447318.2024.2356910
- [76] Lena Mamykina, Drashko Nakikj, and Noemie Elhadad. 2015. Collective sensemaking in online health forums. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 3217–3226.
- [77] Nia F Mason, Diane B Francis, and Loretta L Pecchioni. 2022. Health information seeking as a coping strategy to reduce Alzheimer's caregivers' stress. *Health Communication* 37, 2 (2022), 131–140.

- [78] Michael Massimi, Jackie L Bender, Holly O Witteman, and Osman H Ahmed. 2014. Life transitions and online health communities: reflecting on adoption, use, and disengagement. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. 1491–1501.
- [79] Philip M McCarthy. 2005. *An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD)*. Ph.D. Dissertation. The University of Memphis.
- [80] Clare McCormack, Mark W Wiggins, Thomas Loveday, and Marino Festa. 2014. Expert and competent non-expert visual cues during simulated diagnosis in intensive care. *Frontiers in Psychology* 5 (2014), 949.
- [81] Danielle S McNamara, Scott A Crossley, and Philip M McCarthy. 2010. Linguistic features of writing quality. *Written communication* 27, 1 (2010), 57–86.
- [82] Tamir Mendel, Oded Nov, and Batia Wiesenfeld. 2024. Advice from a doctor or AI? Understanding willingness to disclose information through remote patient monitoring to receive health advice. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW2 (2024), 1–34.
- [83] Omri Mendels, Coby Peled, Nava Vaisman Levy, Tomer Rosenthal, Limor Lahiani, et al. 2018. Microsoft Presidio: Context aware, pluggable and customizable PII anonymization service for text and images. *Microsoft* (2018).
- [84] Jingbo Meng, Minjin Rheu, Yue Zhang, Yue Dai, and Wei Peng. 2023. Mediated social support for distress reduction: AI Chatbots vs. Human. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–25.
- [85] Catrin Misselhorn. 2009. Empathy with inanimate objects and the uncanny valley. *Minds and Machines* 19 (2009), 345–359.
- [86] Drashko Nakikj and Lena Mamykina. 2017. A park or a highway: Overcoming tensions in designing for socio-emotional and informational needs in online health communities. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*. 1304–1319.
- [87] Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*. Association for Computational Linguistics, Florence, Italy, 319–327. doi:10.18653/v1/W19-5034 arXiv:arXiv:1902.07669
- [88] Congning Ni, Qingxia Chen, Lijun Song, Patricia Commiskey, Qingyuan Song, Bradley Malin, and Zhijun Yin. 2025. Catalysts of Conversation: Examining Interaction Dynamics Between Topic Initiators and Commentors in Alzheimer's Disease Online Communities. In *Proceedings of the ACM on Web Conference 2025*. 4913–4924.
- [89] Congning Ni, Bradley Malin, Lijun Song, Angela Jefferson, Patricia Commiskey, and Zhijun Yin. 2022. “Rough day... need a hug”: learning challenges and experiences of the Alzheimer's disease and related dementia caregivers on Reddit. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16. 711–722.
- [90] Congning Ni, Qingyuan Song, Qingxia Chen, Lijun Song, Patricia Commiskey, Lauren Stratton, Bradley Malin, Zhijun Yin, et al. 2024. Sentiment Dynamics Among Informal Caregivers in Web-Based Alzheimer Communities: Systematic Analysis of Emotional Support and Interaction Patterns. *JMIR aging* 7, 1 (2024), e60050.
- [91] Congning Ni, Qingyuan Song, Bradley Malin, Lijun Song, Patricia Commiskey, Lauren Stratton, and Zhijun Yin. 2023. Examining online behaviors of adult-child and spousal caregivers for people living with Alzheimer disease or related dementias: comparative study in an open online community. *Journal of Medical Internet Research* 25 (2023), e48193.
- [92] Ingemar Nordin. 2000. Expert and non-expert knowledge in medical practice. *Medicine, Health care and philosophy* 3 (2000), 295–302.
- [93] Marloes Oldenkamp, Mariët Hagedoorn, Rafael Wittek, Ronald Stolk, and Nynke Smidt. 2017. The impact of older person's frailty on the care-related quality of life of their informal caregiver over time: results from the TOPICS-MDS project. *Quality of Life Research* 26 (2017), 2705–2716.
- [94] Leif Oppermann, Alexander Boden, Britta Hofmann, Wolfgang Prinz, and Stefan Decker. 2019. Beyond HCI and CSCW: Challenges and useful practices towards a human-centred vision of AI and IA. In *Proceedings of the Halfway to the Future Symposium 2019*. 1–5.
- [95] Gregory Park, H Andrew Schwartz, Maarten Sap, Margaret L Kern, Evan Weingarten, Johannes C Eichstaedt, Jonah Berger, David J Stillwell, Michal Kosinski, Lyle H Ungar, et al. 2017. Living in the past, present, and future: Measuring temporal orientation with language. *Journal of personality* 85, 2 (2017), 270–280.
- [96] Jinkyung Katie Park, Vivek K Singh, and Pamela Wisniewski. 2025. Current Landscape and Future Directions for Mental Health Conversational Agents for Youth: Scoping Review. *JMIR Medical Informatics* 13, 1 (2025), e62758.
- [97] Umashanthi Pavalanathan and Munmun De Choudhury. 2015. Identity management and mental health discourse in social media. In *Proceedings of the 24th international conference on world wide web*. 315–321.
- [98] Ellie Pavlick and Joel Tetreault. 2016. An empirical analysis of formality in online communication. *Transactions of the association for computational linguistics* 4 (2016), 61–74.
- [99] James W Pennebaker. 2011. The secret life of pronouns. *New Scientist* 211, 2828 (2011), 42–45.
- [100] James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric properties of LIWC2015. (2015).

- [101] James W Pennebaker, Cindy K Chung, Joey Frazee, Gary M Lavergne, and David I Beaver. 2014. When small words foretell academic success: The case of college admissions essays. *PLoS one* 9, 12 (2014), e115844.
- [102] Melina Petsolari, Seray B. Ibrahim, and Petr Slovak. 2024. Socio-technical Imaginaries: Envisioning and Understanding AI Parenting Supports through Design Fiction. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 98, 27 pages. doi:10.1145/3613904.3642619
- [103] Robert Plutchik. 1960. The multifactor-analytic theory of emotion. *the Journal of Psychology* 50, 1 (1960), 153–171.
- [104] Anthony Poon, Vaidehi Hussain, Julia Loughman, Ariel C Avgar, Madeline Sterling, and Nicola Dell. 2021. Computer-mediated peer support needs of home care workers: emotional labor & the politics of professionalism. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–32.
- [105] Anthony Poon, Matthew Luebbe, Julia Loughman, Ann Lee, Lourdes Guerrero, Madeline Sterling, and Nicola Dell. 2023. Computer-mediated sharing circles for intersectional peer support with home care workers. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–35.
- [106] Christopher Pudlinski. 2005. Doing empathy and sympathy: Caring responses to troubles tellings on a peer support line. *Discourse studies* 7, 3 (2005), 267–288.
- [107] Jordi Quoidbach, June Gruber, Moira Mikolajczak, Aleksandr Kogan, Ilios Kotsou, and Michael I Norton. 2014. Emodiversity and the emotional ecosystem. *Journal of experimental psychology: General* 143, 6 (2014), 2057.
- [108] Brian Richards. 1987. Type/token ratios: What do they really tell us? *Journal of child language* 14, 2 (1987), 201–209.
- [109] Matan Rubin, Hadar Arnon, Jonathan D Huppert, Anat Perry, et al. 2024. Considering the role of human empathy in AI-driven therapy. *JMIR Mental Health* 11, 1 (2024), e56529.
- [110] Emma Scharett, Kapil Chail Madathil, Snehal Lopes, Hunter Rogers, Sruthy Agnisarman, Shraddha Narasimha, Aparna Ashok, and Cheryl Dye. 2017. An investigation of the information sought by caregivers of Alzheimer's patients on online peer support groups. *Cyberpsychology, behavior, and social networking* 20, 10 (2017), 640–657.
- [111] Corey Kai Nelson Schultz. 2023. Creating the 'virtual' witness: the limits of empathy. *Museum Management and Curatorship* 38, 1 (2023), 2–17.
- [112] John Scott, Peter J Carrington, Lijun Song, Joonmo Son, and Nan Lin. 2014. Social Support. In *The SAGE Handbook of Social Network Analysis*. SAGE Publications Ltd, 116–128.
- [113] Lennart Seitz. 2024. Artificial empathy in healthcare chatbots: Does it feel authentic? *Computers in Human Behavior: Artificial Humans* 2, 1 (2024), 100067.
- [114] Farhana Shahid, Maximilian Dittgen, Mor Naaman, and Aditya Vashistha. 2024. Examining Human-AI Collaboration for Co-Writing Constructive Comments Online. *arXiv preprint arXiv:2411.03295* (2024).
- [115] Ashish Sharma, Inna W Lin, Adam S Miner, David C Atkins, and Tim Althoff. 2023. Human-AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *Nature Machine Intelligence* 5, 1 (2023), 46–57.
- [116] Jocelyn Shen, Daniella DiPaola, Safinah Ali, Maarten Sap, Hae Won Park, Cynthia Breazeal, et al. 2024. Empathy Toward Artificial Intelligence Versus Human Experiences and the Role of Transparency in Mental Health and Social Support Chatbot Design: Comparative Study. *JMIR Mental Health* 11, 1 (2024), e62679.
- [117] Jiayue Melissa Shi, Keran Wang, Dong Whi Yoo, Ravi Karkar, and Koustuv Saha. 2025. Balancing Caregiving and Self-Care: Exploring Mental Health Needs of Alzheimer's and Dementia Caregivers. *Proceedings of the ACM on Human-Computer Interaction* 9, 7 (2025), 1–36. doi:10.1145/3757555
- [118] Diva Smriti, Lu Wang, and Jina Huh-Yoo. 2024. Emotion work in caregiving: the role of technology to support informal caregivers of persons living with dementia. *Proceedings of the ACM on human-computer interaction* 8, CSCW1 (2024), 1–34.
- [119] Vera Sorin, Dana Brin, Yiftach Barash, Eli Konen, Alexander Charney, Girish Nadkarni, and Eyal Klang. 2024. Large Language Models and Empathy: Systematic Review. *Journal of Medical Internet Research* 26 (2024), e52597.
- [120] Lauren Squires. 2010. Enregistering internet language. *Language in society* 39, 4 (2010), 457–492.
- [121] Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology* 29, 1 (2010), 24–54.
- [122] Aman Singh Thakur, Kartik Choudhary, Venkat Srinik Ramayapally, Sankaran Vaidyanathan, and Dieuwke Hupkes. 2024. Judging the judges: Evaluating alignment and vulnerabilities in llms-as-judges. *arXiv preprint arXiv:2406.12624* (2024).
- [123] Anneke Ullrich, Marianna Theochari, Corinna Bergelt, Gabriella Marx, Katharina Woellert, Carsten Bokemeyer, and Karin Oechsle. 2020. Ethical challenges in family caregivers of patients with advanced cancer—a qualitative study. *BMC Palliative Care* 19 (2020), 1–13.
- [124] Alan Vaux, Sharon Riedel, and Doreen Stewart. 1987. Modes of social support: The social support behaviors (SS-B) scale. *American Journal of Community Psychology* 15, 2 (1987), 209–232.

- [125] Lesley Verhofstadt, Inge Devoldre, Ann Buysse, Michael Stevens, Céline Hinnekens, William Ickes, and Mark Davis. 2016. The role of cognitive and affective empathy in spouses' support interactions: An observational study. *PloS one* 11, 2 (2016), e0149944.
- [126] Sarah J Wallace, Janaki Kothari, Anushki Jayasekera, Jessica Tointon, Toluwalase Baiyewun, and Kirstine Shrubsole. 2021. Do caregivers who connect online have better outcomes? A systematic review of online peer-support interventions for caregivers of people with stroke, dementia, traumatic brain injury, Parkinson's disease and multiple sclerosis. *Brain Impairment* 22, 3 (2021), 233–259.
- [127] Leyao Wang, Zhiyu Wan, Congning Ni, Qingyuan Song, Yang Li, Ellen Clayton, Bradley Malin, and Zhijun Yin. 2024. Applications and concerns of ChatGPT and other conversational large language models in health care: systematic review. *Journal of Medical Internet Research* 26 (2024), e22769.
- [128] Liuping Wang, Dakuo Wang, Feng Tian, Zhenhui Peng, Xiangmin Fan, Zhan Zhang, Mo Yu, Xiaojuan Ma, and Hongan Wang. 2021. Cass: Towards building a social-support chatbot for online health community. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–31.
- [129] Yi-Chia Wang, Robert Kraut, and John M Levine. 2012. To stay or leave? The relationship of emotional and informational support to commitment in online health support groups. In *Proceedings of the ACM 2012 conference on computer supported cooperative work*. 833–842.
- [130] Joel Wester, Henning Pohl, Simo Hosio, and Niels van Berkel. 2024. "This Chatbot Would Never...": Perceived Moral Agency of Mental Health Chatbots. *Proceedings of the ACM on human-computer Interaction* 8, CSCW1 (2024), 1–28.
- [131] Lauren Wilcox, Renee Shelby, Rajesh Veeraraghavan, and Oliver L. Haimson. 2023. Infrastructuring Care: How Trans and Non-Binary People Meet Health and Well-Being Needs through Technology. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, 1–? doi:10.1145/3544548.3581040
- [132] Yee Hui Yeo, Jamil S Samaan, Wee Han Ng, Peng-Sheng Ting, Hirsh Trivedi, Aarshi Vipani, Walid Ayoub, Ju Dong Yang, Omer Liran, Brennan Spiegel, et al. 2023. Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma. *Clinical and molecular hepatology* 29, 3 (2023), 721.
- [133] Zhijun Yin, Lauren Stratton, Qingyuan Song, Congning Ni, Lijun Song, Patricia Commiskey, Qingxia Chen, Monica Moreno, Sam Fazio, Bradley Malin, et al. 2024. Perceptions and utilization of online peer support among informal dementia caregivers: survey study. *JMIR aging* 7, 1 (2024), e55169.
- [134] Ziqi Yin, Hao Wang, Kaito Horio, Daisuke Kawahara, and Satoshi Sekine. 2024. Should we respect LLMs? A cross-lingual study on the influence of prompt politeness on LLM performance. In *Proceedings of the Second Workshop on Social Influence in Conversations (SICoN 2024)*. 9–35.
- [135] Jordyn Young, Laala M Jawara, Diep N Nguyen, Brian Daly, Jina Huh-Yoo, and Afsaneh Razi. 2024. The role of ai in peer support for young people: A study of preferences for human-and ai-generated responses. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–18.
- [136] Weixiang Zhao, Yanyan Zhao, Xin Lu, Shilong Wang, Yanpeng Tong, and Bing Qin. 2023. Is ChatGPT equipped with emotional dialogue capabilities? *arXiv preprint arXiv:2304.09582* (2023).
- [137] Zhao Zou, Omar Mubin, Fady Alnajjar, and Luqman Ali. 2024. A pilot study of measuring emotional response and perception of LLM-generated questionnaire and human-generated questionnaires. *Scientific reports* 14, 1 (2024), 2781.
- [138] Tamara Zubatiy, Niharika Mathur, Larry Heck, Kayci L Vickers, Agata Rozga, and Elizabeth D Mynatt. 2023. "I Don't Know How to Help with That:" Examining the Design and Function of a Caregiving Network by a Conversational Agent. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (2023), 1–28. doi:10.1145/3610170

Received 13 May 2025; revised 13 January 2026; accepted 9 April 2026