Human Interaction and Emerging Technologies (IHIET 2026), Vol. 236, 2026, 70–79
https://doi.org/10.54941/ahfe1008083

Assessment-Before-Intervention: A WHO iSupport-Grounded Conversational AI System for Dementia Family Caregiver Support

Chor-Kheng Lim[1], Hsien-Hui Tang[2], and Chia-Chi Chang[3]

[1]Department of Art and Design, Yuan Ze University, Taoyuan, Taiwan
[2]National Taiwan University of Science and Technology, Taipei, Taiwan
[3]College of Nursing, Taipei Medical University, Taipei, Taiwan

ABSTRACT

Family caregivers of people with dementia face daily behavioral challenges—aggression, wandering, agitation—that require timely, contextual guidance. Existing chatbot systems typically respond with direct advice, bypassing assessment of behavioral context and caregiver emotional state. This paper presents a LINE-deployed conversational AI system grounded in the WHO iSupport for Dementia framework, built around an Assessment-Before-Intervention dialogue mechanism. The system applies a four-step reasoning process derived from the iSupport ABC behavioral cycle, governed by five safety principles, to determine when to clarify before advising. A dual-layer response model ensures emotional acknowledgment is never omitted; a hybrid keyword-semantic Retrieval-Augmented Generation (RAG) architecture bridges the lay-to-clinical vocabulary gap. We evaluated the system through a formative review with four domain reviewers in dementia care, covering 13 BPSD scenarios (52 evaluations across five quality dimensions). Mean scores (4.78–4.81 on a 5-point Likert scale) are interpreted as preliminary perceived-appropriateness data rather than clinical effectiveness evidence. The principal contribution lies in the qualitative findings: five recurrent failure modes and a structural pattern of cultural misalignment between the international iSupport framework and Taiwanese caregiving realities. Findings offer practical implications for the design of AI-assisted care systems in dementia and other emotionally sensitive healthcare contexts.

Keywords: Conversational AI, Dementia caregiver support, Retrieval-augmented generation, Assessment-before-intervention, Empathy-centered design, Healthcare technology

INTRODUCTION

Dementia affects over 57 million people worldwide, with the majority of care provided by family members who lack formal training (WHO, 2023). These informal caregivers face daily behavioral challenges that generate substantial emotional strain. When they seek digital support, the structure of the response matters: a caregiver who messages "my father keeps wandering at night and I don't know what to do" is not only requesting management strategies—they may also be exhausted, frightened, or guilt-ridden. Effective responses must address both the behavioral symptom and the caregiver's emotional state, yet existing tools rarely do both.

Received June 5, 2026; Revised June 16, 2026; Accepted July 1, 2026; Available online August 25, 2026

Prior caregiver chatbots have largely framed this as information retrieval: match query to knowledge, return advice (Jiménez et al., 2022; Cheng and Ng, 2025). This direct advice-giving model presents two limitations. It assumes sufficient context from a single message—which brief, emotionally laden caregiver queries rarely satisfy—and it positions the caregiver as a passive information consumer rather than an emotionally present person who benefits from acknowledgment before advice. We address both limitations by operationalizing the WHO iSupport for Dementia framework (WHO, 2019) as structured dialogue logic, deployed as a LINE messaging bot. We report results from a formative multi-dimensional review with four domain reviewers in dementia care and discuss design implications for AI-assisted healthcare systems.

RELATED WORK

Conversational agents for dementia caregivers have been deployed in evidence-based clinical care (Ruggiano et al., 2024), psychoeducation (Cheng and Ng, 2025), and Retrieval-Augmented Generation contexts (Li et al., 2024). Espinoza et al. (2023) found generative AI produced more contextually responsive emotional replies than structured configurations, and Kim et al. (2025) mapped caregiver needs to chatbot design dimensions, emphasizing emotional-state adaptation. A consistent gap across these systems is the absence of formalized rules governing when to seek clarification versus when to advise; our work addresses this gap with domain-grounded reasoning derived from iSupport. Hybrid keyword-semantic retrieval has been shown to improve recall in biomedical NLP (Stuhlmann et al., 2025), while Ruggiano et al. (2020) documented systematic vocabulary mismatch between lay caregiver language and clinical terminology—motivating our two-layer retrieval design.

SYSTEM DESIGN

The system is a LINE messaging bot deployed on Google Apps Script, using Claude Haiku as the primary LLM with Gemini 2.5 Flash fallback, and the WHO iSupport Traditional Chinese handbook as the knowledge base. The Assessment-Before-Intervention mechanism is the primary dialogue-level contribution; the dual-layer response model and hybrid RAG architecture function as supporting infrastructure. Figure 1 summarizes how these components interact in a single message turn.

Dual-Layer Caregiver State Interpretation

Every caregiver message contains two layers: a *surface behavioral problem* (what the patient is doing) and an *underlying emotional state* (what the caregiver is experiencing). The iSupport framework makes this clinically explicit: the Consequence (C) in the behavioral cycle is the caregiver's emotional reaction, which becomes part of the triggering conditions for the next episode (WHO, 2019, Module 5). We operationalize this as a structured JSON output with five fields: **reply, empathy, clarify, steps,** and **reason**. The empathy field is always present, ensuring acknowledgment is never omitted regardless of response mode.

Assessment-Before-Intervention Dialogue Logic

The central contribution is a four-step reasoning process governing when the system seeks clarification before generating recommendations. In Step 1 (ABC Identification), the system identifies the behavioral symptom, known antecedents (triggers, timing, frequency), and the caregiver's emotional state from the current message and conversation history.

In Step 2, **five prioritized safety principles** derived from iSupport are applied: (**A**) **behavior as communication**—dementia behaviors reflect unmet needs, not deliberate acts; (**B**) **physical before psychological**—sudden behavioral changes prompt assessment of physical causes (infection, pain, medication) first; (**C**) **environment before person**—recommendations default to adjustable environmental factors rather than expecting behavioral change from a cognitively impaired person; (**D**) **dignity as system stability**—coercive caregiving advice (forced bathing, public correction) is avoided as it predictably escalates agitation; (**E**) **reattribution as prerequisite**—if caregiver language implies intentional interpretation ("he does this to spite me"), cognitive reattribution is introduced before practical advice.

In Step 3 (Information Sufficiency Assessment), if no safety principle triggers an immediate response and at least one contextual element is known, the system proceeds to advice; otherwise, it moves to Step 4 (Targeted Clarification). The clarifying question is derived from the specific ABC gap identified in Step 1, not a generic prompt. After three consecutive clarification rounds, a sufficiency fallback overrides: the system provides advice while acknowledging remaining uncertainty.

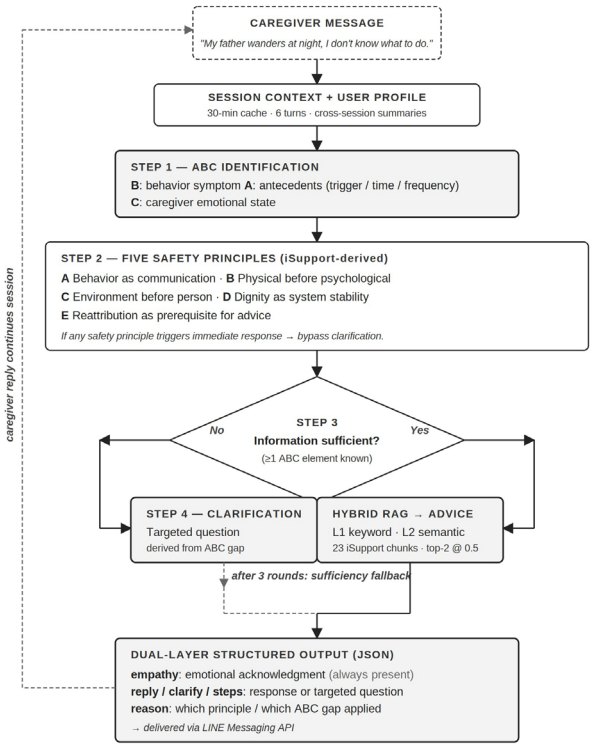

Figure 1: System architecture and Assessment-Before-Intervention dialogue flow.

Hybrid RAG Architecture

The knowledge base comprises 23 content chunks from the iSupport Traditional Chinese handbook, manually segmented by iSupport topic and caregiving scenario rather than by fixed token length to preserve clinical and caregiving coherence. Retrieval uses a two-layer sequential cascade designed to bridge the vocabulary gap documented between lay caregiver language and clinical terminology (Ruggiano et al., 2020). Layer 1 (keyword matching) uses approximately 100 iSupport-derived behavioral terms for direct chunk retrieval with zero API cost. Layer 2 (semantic search), triggered when keyword matching fails, computes cosine similarity over Gemini Embedding vectors (3,072 dimensions) and returns top-2 chunks above a 0.5 similarity threshold, set as an implementation heuristic to be formally optimized in subsequent retrieval evaluation. A context-aware rule concatenates short follow-up messages (<15 characters) with the prior turn before retrieval, preventing context-free embedding failures. Table 1 illustrates the vocabulary gap.

Table 1: Caregiver expressions, clinical referents, and retrieval layer.

Caregiver Expression (Chinese)	Clinical Term	Retrieval Layer
尿失禁、包尿布 (incontinence)	Urinary incontinence	Layer 1 — keyword
媽媽不讓我幫她換衣服 (won't let me help change)	Personal care resistance	Layer 2 — semantic
晚上一直起來走 (gets up walking at night)	Nocturnal wandering	Layer 2 — semantic
又發脾氣了 (lost his temper again)	Agitation / aggression	Layer 2 — semantic

IMPLEMENTATION

The system runs on Google Apps Script with the LINE Messaging API. A 5-question onboarding flow collects user profile data stored in Google Sheets and injected into each system prompt. A session cache (30 minutes, up to 6 turns) maintains conversation history; cross-session memory stores AI-generated summaries of up to 10 prior episodes. The system prompt was intentionally minimized to approximately 850 characters containing three core principles, after iterative development showed that longer, example-laden prompts caused the LLM to produce repetitive, templated responses.

ETHICAL CONSIDERATIONS

This study evaluated system responses through review by four domain reviewers in dementia care who consented to participate; no patient data were collected, and the test queries posed by reviewers were constructed from professional experience rather than drawn from real caregiver-patient interactions. All reviewer commentary was anonymized prior to analysis. The deployed LINE bot displays an opening disclaimer indicating that system responses are informational and do not substitute for professional medical advice or emergency care; the disclaimer is reissued whenever a

session pauses for more than 30 minutes. Test deployment with caregivers, planned as a subsequent evaluation phase, will be conducted under formal ethics review by the host institution prior to recruitment.

EVALUATION

Review Procedure

We conducted a formative multi-dimensional review with four domain reviewers in dementia care: a registered psychiatric nurse with seven years of hospital experience, a registered nurse with four to five years of experience in a dementia day-care service, a dementia care educator and researcher, and a final-year nursing student in a clinical practicum. Reviewers were recruited through the principal investigator's existing professional network in dementia care and consented to participate prior to interaction with the system.

Each reviewer independently interacted with the deployed system via LINE, posing test queries that mapped onto 13 BPSD scenarios drawn from the iSupport framework. To balance external validity with comparability, reviewers were provided with the 13 target scenarios but encouraged to phrase test queries in their own clinical voice rather than from a fixed script; this design choice prioritized realistic linguistic variation over inter-reviewer prompt identity. Following each interaction, reviewers independently rated the response on five dimensions using a 5-point Likert scale: completeness and accuracy (drawn from established AI health information quality literature), and actionability, relevance (personalization), and clarification guidance (derived from the project's design objectives, with clarification guidance directly targeting the Assessment-Before-Intervention contribution). The review yielded 52 evaluations (4 reviewers × 13 scenarios), each accompanied by open-ended commentary on strengths and improvement opportunities.

Open-ended comments were analyzed using an inductive thematic approach by the first author. Initial codes were derived from individual reviewer comments, then grouped into candidate themes; patterns that recurred across at least two reviewers or three or more scenarios were retained as the failure modes and cultural-misalignment categories reported in Section "Qualitative Findings". Given the formative scope, a second independent coder and formal inter-rater reliability were not used; the resulting categories should be read as design-relevant patterns surfaced from the review rather than as a quantitative content analysis.

Quantitative Results

Mean scores across the 52 evaluations were consistently favorable (Table 2), with overall means between 4.78 and 4.81 and narrow standard deviations (0.42–0.45). We interpret these scores as preliminary evidence of perceived response appropriateness in a small formative review rather than as evidence of clinical effectiveness; given N = 4 reviewers and a single regional network, the principal value of the quantitative results is in scenario-level patterning

(Table 3) and in their contrast with the qualitative findings reported in Section "Qualitative Findings".

Table 2: Review results across five quality dimensions (5-point Likert, N = 52).

Dimension	M	SD	Source
Completeness	4.78	0.42	Literature
Accuracy	4.80	0.45	Literature
Actionability	4.80	0.45	Design objective
Relevance (Personalization)	4.81	0.44	Design objective
Clarification Guidance	4.81	0.44	Design objective

A scenario-level breakdown (Table 3) reveals where the system performed least well. The lowest single-cell score was *Memory Decline* × *Actionability* (M = 4.25), with reviewers noting that visual-cue and written-reminder strategies—commonly recommended in iSupport itself—assume preserved reality testing that is absent in moderate-to-severe dementia. Three additional cells scored 4.50: Repetitive Behavior × Accuracy, Wandering × Accuracy, and Daily Eating × Relevance. Inter-rater variability was driven primarily by one reviewer (the nursing student), whose ratings averaged 4.15–4.46—markedly lower than the other three reviewers (4.85–5.00) and accompanied by the most detailed improvement-oriented commentary, indicating that the overall ratings should be read alongside the qualitative findings below.

Table 3: Scenario × dimension matrix of mean scores (cells below 4.75 highlighted).

Scenario	Complete	Accurate	Actionable	Relevant	Clarify
Behavioral Change	4.75	4.75	5.00	5.00	5.00
Memory Decline	4.75	4.75	**4.25**	4.75	**4.50**
Aggression	4.75	5.00	5.00	5.00	5.00
Depression / Anxiety	4.75	5.00	5.00	5.00	5.00
Sleep Disturbance	4.75	4.75	4.75	4.75	4.75
Delusion / Hallucination	4.75	4.75	4.75	4.75	4.75
Repetitive Behavior	4.75	**4.50**	4.75	4.75	4.75
Wandering	4.75	**4.50**	4.75	4.75	4.75
Judgment Decline	4.75	4.75	4.75	4.75	4.75
Daily Eating	4.75	4.75	4.75	**4.50**	4.75
Toileting	4.75	4.75	4.75	4.75	4.75
Care Service Needs	5.00	5.00	5.00	4.75	5.00
Caregiver Wellbeing	4.75	5.00	4.75	5.00	4.75

Dialogue Example

Table 4 presents an excerpt from one reviewer's interaction illustrating Assessment-Before-Intervention across multiple turns. The reviewer raised a sleep-disturbance concern; rather than advising immediately, the system

applied Step 1 ABC identification, found antecedents and frequency unknown, and issued targeted clarifying questions across two turns before producing structured recommendations. In Turn 4, when the caregiver introduced a Taiwanese-language term ("度辜", dùkku), the system asked for clarification of meaning before continuing—a culturally sensitive disambiguation behavior the reviewer specifically commended. This reviewer rated the interaction 5/5 across all dimensions.

Table 4: Excerpt illustrating multi-turn assessment and cultural-linguistic disambiguation.

Turn	Speaker	Message	System Logic
1	Caregiver	What can I do if grandfather doesn't sleep at night but keeps sleeping during the day?	ABC identified: B = day-night reversal; A and frequency unknown; C = fatigue implied.
1	System	Has this started recently or has it gone on for a while? When grandfather doesn't sleep, what does he do — lie awake, get up to move around, or walk in his room?	Step 4 targeted clarification: temporal antecedent + behavioral specificity.
2	Caregiver	He organizes the wardrobe at midnight.	(partial information received)
2	System	How long has this habit been going on? Also, does he sleep at fixed daytime hours — when does he start, and for how long?	Continued ABC assessment.
3	Caregiver	He usually sleeps after dinner, takes his sleeping pill, then wakes at 2–3 a.m. making noise.	(sufficient context)
3	System	[empathy] Being woken at 2–3 a.m. while still needing to function is exhausting. [steps] (1) increase daytime activity; (2) delay evening sleep onset; (3) restrict drawer access. [referral] consult physician on medication timing.	Advice mode. Principle B applied (medical-first).
4	Caregiver	In the morning he's not energetic, 一直度辜 (dùkku).	(Taiwanese expression)
4	System	Could you clarify what '度辜' means here — being dazed, slow to respond, or irritable?	Cultural-linguistic clarification.

Qualitative Findings: Failure Modes and Cultural Localization

While quantitative results indicate consistently favorable overall quality, the open-ended commentary surfaced five recurring failure patterns and a structural pattern of cultural misalignment between WHO iSupport and Taiwanese caregiving realities.

Five recurrent failure modes. (1) ***Cognitive Reality Gap***: iSupport-style cognitive cues (photographs, written reminders) assume preserved reality testing absent in moderate-to-severe dementia; two reviewers noted that recommendations require staging by disease severity. (2) ***Premature Advice***: when frequency or timing was already supplied in the initial message, the system asked again, indicating that Step 1 extraction needs sharpening. (3) ***Medical Red Flag Blindness***: in the daily-eating scenario, "holding rice in mouth without swallowing" could signal dysphagia, infection, or hypoglycemia—medical concerns that should override behavioral framing, but the system remained behaviorally framed. (4) ***Resource Referral Discontinuity***: in the wandering scenario, the system mentioned that day-care facilities require a formal diagnosis but did not continue guiding toward how to obtain one or alternative respite options, leaving a closed-loop referral. (5) ***Empathy Fatigue***: when users replied with minimal-engagement phrases ("還好"/"okay-ish"), the system sometimes responded with single-character acknowledgments ("嗯"/"mm"), which one reviewer read as disengagement—a content-level rather than structural failure of the empathy field.

Cultural misalignment. A second pattern points to structural gaps between the international iSupport framework and the lived realities of Chinese-speaking Taiwanese caregivers. In *Judgment Decline*, iSupport foregrounds social-disinhibition scenarios; the review surfaced *financial fraud* (telephone scams targeting older adults with cognitive impairment) and *life-safety incidents* (forgetting to turn off gas) as the more pressing concerns. *Toileting* has no stand-alone module in the iSupport handbook, yet reviewers identified it as one of the most dignity-eroding daily challenges. In *Behavioral Change*, iSupport emphasizes agitation and temper outbursts, while reviewers highlighted *compulsive consumption* (excessive shopping, hoarding of printed material) as frequent and underrepresented. In *Repetitive Behavior*, iSupport vocabulary emphasizes repetitive *speech*; reviewers noted that repetitive *physical actions* (packing a suitcase, organizing the same cabinet) cause comparable distress. These observations indicate that systems grounded in a single international framework may carry coverage gaps surfaced only through domain review in the deployment culture.

Additional Observations

Two additional observations were noted during development. First, in informal comparison of two LLMs for the empathy field, Gemini 2.5 Flash produced highly similar empathy strings across near-identical inputs while Claude Haiku produced more varied phrasings—suggesting response variability merits consideration in emotionally sensitive AI selection. Second, system logs showed that a substantial portion of initial caregiver messages used lay or colloquial expressions ("他都不睡", "一直罵人") that did not match iSupport clinical vocabulary directly and fell through to semantic retrieval, qualitatively supporting the lay-to-clinical vocabulary gap (Ruggiano et al., 2020) and the two-layer architecture's design rationale. Both observations warrant formal investigation in subsequent work.

DISCUSSION

Assessment as care, and the limits of single-framework grounding. The review supported the project's central design stance: in emotionally loaded support contexts, seeking clarification is itself a form of care. The Clarification Guidance dimension—added specifically to test this contribution—received among the highest mean ratings (M = 4.81), and the multi-turn dialogue (Table 4) shows the mechanism functioning as intended. At the same time, the cultural misalignment findings indicate the limits of grounding in a single international framework. For systems serving Chinese-speaking Taiwanese caregivers, the iSupport knowledge base should be augmented with locally observed concerns—financial fraud, gas-safety incidents, toileting, compulsive consumption—identified through domain review rather than transferred wholesale from international sources. This finding generalizes beyond dementia care: any LLM-based assistive system grounded in an internationally validated framework warrants culturally specific augmentation when deployed outside its original development context.

From failure modes to design iteration. The five recurrent failure patterns map cleanly to specific design improvements for the next iteration: severity-staged recommendations (Cognitive Reality Gap), tightened context extraction in Step 1 (Premature Advice), an explicit medical-red-flag detection layer that overrides behavioral framing whenever symptoms suggest medical urgency—for example, dysphagia-like eating problems, sudden confusion, fever, falls, medication-related changes, or risks of harm to self or others (Medical Red Flag Blindness)—a safety capability we now consider a design requirement rather than an optional refinement; referral-pathway continuation logic (Resource Referral Discontinuity); and content-level rules for empathy generation in low-engagement turns (Empathy Fatigue). The directness of this mapping suggests that multi-dimensional domain review with open-ended commentary is a productive evaluation method for formative-stage caregiver-support AI.

Limitations. This is a formative review (N = 4 reviewers) intended to surface design issues prior to broader deployment; it is not a substitute for a user study with family caregivers. Reviewers were drawn from a single regional clinical network and included one final-year nursing student whose perspective differed substantially from the more experienced clinicians (and proved valuable in surfacing failure modes). A single coder analyzed the qualitative data, so the failure modes and cultural-misalignment patterns should be read as design-relevant observations rather than validated categories. A randomized caregiver-facing study, multi-coder qualitative analysis, and formal log analysis of retrieval-layer usage are planned as the next evaluation phase.

CONCLUSION

This paper presented a WHO iSupport-grounded conversational AI system for dementia family caregivers that operationalizes the ABC behavioral model as structured dialogue logic. A formative multi-dimensional review with four domain reviewers across 13 BPSD scenarios yielded consistently

favorable quality ratings (M = 4.78–4.81)—best interpreted as preliminary perceived-appropriateness data given the small sample—alongside qualitative findings pointing to five recurrent failure modes and a structural pattern of cultural misalignment between the international iSupport framework and Taiwanese caregiving realities. The design approach offers a replicable model for AI-assisted care systems in dementia and other emotionally sensitive healthcare contexts, while underscoring that internationally validated frameworks may require culturally specific augmentation in non-Western caregiving deployments.

REFERENCES

Cheng, S.-T. and Ng, P. H. F. (2025). The PDC30 Chatbot—Development of a Psychoeducational Resource on Dementia Caregiving: Mixed Methods Acceptability Study. JMIR Aging, 8, e63715.

Espinoza, A. et al. (2023). Supporting Dementia Caregivers in Peru Through Chatbots: Generative AI vs Structured Conversations. Proceedings of BCS HCI 2023.

Jiménez, S. et al. (2022). Towards Conversational Agents to Support Informal Caregivers of People with Dementia: Challenges and Opportunities. Programming and Computer Software, 48(8), 606–613.

Kim, S. et al. (2025). Mapping Caregiver Needs to AI Chatbot Design: Strengths and Gaps in Mental Health Support for Alzheimer's and Dementia Caregivers. arXiv:2506.15047.

Li, J. et al. (2024). Empowering Alzheimer's Caregivers with Conversational AI: A Novel Approach for Enhanced Communication and Personalized Support. npj Biomedical Innovations, 1(4).

Ruggiano, N. et al. (2020). Free-Text Documentation of Dementia Symptoms in Home Healthcare: A Natural Language Processing Study. Research in Gerontological Nursing, 13(4), 176–185.

Ruggiano, N. et al. (2024). An Evidence-Based IT Program With Chatbot to Support Caregiving and Clinical Care for People With Dementia: The CareHeroes Development and Usability Pilot. JMIR Aging, 7, e57308.

Stuhlmann, A. et al. (2025). Efficient and Reproducible Biomedical Question Answering using Retrieval Augmented Generation. arXiv:2505.07917.

WHO (2019). iSupport for Dementia: Training and Support Manual for Carers of People with Dementia. World Health Organization, Geneva. (Taiwan Traditional Chinese edition, 2025.)

WHO (2023). Dementia. WHO Fact Sheet. World Health Organization.