

Grounded Personal Memory Retrieval: A Location-Aware AI Assistant with Sentence-Level Verification for Cognitive Support

Aadesh Rathod, Shriya Mishra, Purva Dalvi, Siddhi Deshmukh, Ramesh Shahabade

Department of Computer Engineering, TPCT's Terna Engineering College, Nerul, Navi Mumbai, Maharashtra, India

Published November 2026

Abstract

People with mild cognitive impairment and early Alzheimer's disease often cannot recall recent personal experiences. Conversational assistants built on large language models could help, but they carry a specific risk for this population: a user with memory impairment cannot check an answer against their own memory, so an unsupported statement may be accepted as a real memory. This paper presents a phone and web memory assistant that retrieves a user's own recorded experiences by place and by plain-language question while constraining every answer to those records. Photographs, voice notes, and text notes are time-stamped and grouped into named places using geohash prefix search. Questions pass through a retrieval pipeline that refuses before any model call when nothing relevant exists, requires each drafted sentence to cite a retrieved record, and filters every sentence through a verification component that removes uncited, unretrieved, private, medical, or lexically unsupported content. Caregiver access is consent-based, deny-by-default, and visible to the user. We define six measurable grounding requirements and evaluate them on a synthetic persona with 32 records and 48 questions, without patient data. In offline mode, every statement was traceable to a retrieved record, no unsupported statement was produced, 9 of 10 unanswerable questions were refused, and all 28 answerable questions cited a correct record. We discuss the limitations of this evaluation and the work needed before clinical use.

Keywords: Alzheimer's disease; assistive technology; grounding; hallucination; memory aids; mild cognitive impairment; retrieval-augmented generation

1 Introduction

More than 55 million people worldwide live with dementia, and Alzheimer's disease accounts for an estimated 60–70% of cases [1]. Mild cognitive impairment (MCI) is an important window for assistive technology: memory is affected, but people can still live independently and learn new tools [2, 3]. A consistent finding is that cues drawn from a person's own day support recall. In a randomized study of 51 patients with mild Alzheimer's disease, photographs from a wearable camera improved autobiographical memory more than memory training or a written diary, and the benefit persisted at six-month follow-up [4].

Large language models (LLMs) make it possible to ask questions about one's own life in plain language, and their potential in dementia care has been recognized together with their risks [5]. Retrieval-augmented generation (RAG) grounds model output in retrieved documents [6], yet generated text can still contain content that its sources do not support [7]. For a user with memory impairment this failure is especially serious: the user cannot check an answer against their own memory, so a fabricated "memory" may be believed and acted upon. Evaluations of LLMs on MCI-related questions

report quality that varies with language and user group [8], and retrieval over personal memory graphs has been demonstrated [9], but patient-facing systems rarely measure whether their answers are supported by what the user actually recorded.

This paper addresses that gap with a memory assistant whose answers are constrained to, and checked against, the user's own records. Its contributions are:

- a place-indexed memory store that groups records captured within 150 m into named places using geohash prefix search and surfaces them when the user returns to a place;
- a grounded answering pipeline that refuses before any model call when no relevant record exists, requires every drafted sentence to cite a retrieved record, and verifies each sentence with a component we call the GroundingGuard;
- six measurable grounding requirements and an evaluation protocol on a synthetic persona with known answers, requiring no patient data; and
- consent-visible caregiver access, enforced deny-by-default on the server and verified with negative tests.

The remainder of the paper reviews related work, describes the system and its methods, presents the experimental setup and the results, and concludes.

2 Related work

2.1 Memory cueing and reminder systems

Wearable-camera cueing [4] is the strongest evidence that personal cues aid recall, but the images were reviewed manually in date order, with no search or place index, on hardware that is no longer available. MEMO_Calendring organized a patient's scheduled events in a hand-built ontology and accepted natural-language input [10], which requires manual knowledge engineering for each new user. A review of 40 reminder and guidance systems found that patients value in-app support for daily tasks when designs respect autonomy and personalization [11]. Mobile augmented reality has been used to help patients identify objects and people, although with only four participants [12]. Reviews of dementia applications report that research apps cluster around reminders, communication, and caregiver support with weak security and usability evaluation [13], and that commercial apps vary widely in content and quality [14].

2.2 Conversational agents and language models

AlzBot combined a chatbot with location tracking to reduce boredom and isolation, but its conversation has no access to the patient's stored experiences [15]. Kostis et al. showed that people with MCI, unlike those at later stages, can engage productively with a conversational agent [16]. Evaluations of LLMs on MCI questions show variation across languages and user groups [8], and reviews describe both the opportunities and the risks of LLMs in dementia care [5]. DEMENTIA-PLAN applies RAG over multiple knowledge graphs of personal memories [9] and is the design closest to ours; our focus is on measuring whether each answer sentence is supported by the user's records and on refusing when it is not, which surveys of hallucination in natural language generation show to be necessary [7].

2.3 Caregivers, ethics, and validation

A scoping review of 12 studies found caregiver mHealth applications inadequate across six functional themes [17]; an Android application supports caregivers in managing routines and medication [18]; and technology-enabled caregivers report high burden, depression, anxiety, and sleep disturbance [19]. Location features in assistive systems mostly serve tracking and safety [15, 20]. Most intelligent assistive technologies never reach clinical validation [3], and an overview of 15 reviews identifies autonomy, privacy, and equity as the dominant ethical concerns [21].

2.4 Research gap

Table 1 compares representative systems on four properties. To our knowledge, no surveyed system combines retrieval over the user's own records by place and by plain-language question with a measured guarantee that answers are supported by those records.

Table 1. Comparison with representative prior work.

System	Own data	Place recall	Plain-language questions	Answers verified
Wearable camera [4]	Yes	No	No	–
MEMO_Calendring [10]	Yes	No	Yes	No
AlzBot [15]	No	No	Yes	No
This work	Yes	Yes	Yes	Yes

3 System overview

Figure 1 shows the architecture. The client is a React progressive web application that runs in any modern browser and can be installed on Android and iOS. It captures photographs, voice notes of up to two minutes, and text notes; adds the time and, with permission, the location; and keeps an offline queue so that memories saved without a connection upload later. Questions can be typed or spoken, and answers can be read aloud. The interface follows rules suited to cognitive impairment: large text, 56-pixel touch targets, words beside every icon, one task per screen, and no content that disappears on a timer, in line with WCAG 2.1 [22].

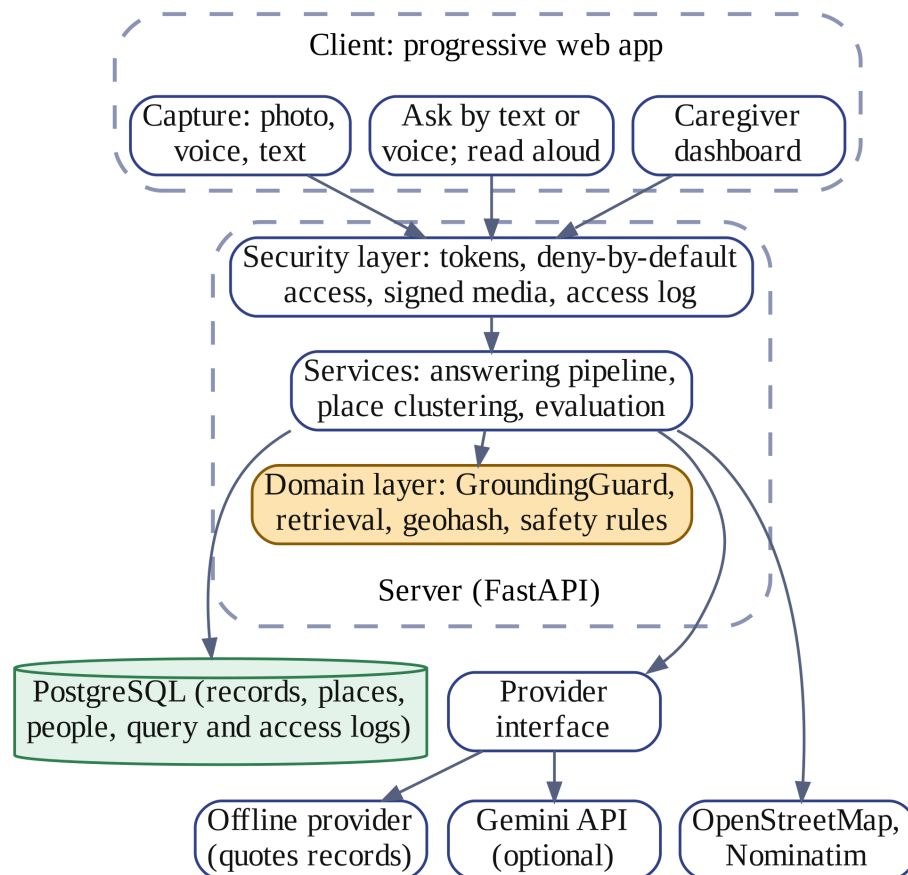


Figure 1. System architecture. Only the server reaches the data store or a language model.

The server is a FastAPI application organized in four layers: API routers; a security layer that verifies JSON Web Tokens [23], enforces access rules, signs media links, and writes the access log; services for answering questions, clustering places, and running the evaluation; and a domain layer in pure Python that contains the GroundingGuard, retrieval, geohash, and medical-safety logic. The domain layer imports no framework or vendor SDK, so it is unit-tested without a network. Data are stored in PostgreSQL across 13 tables; every question is logged with its retrieved records, scores, drafted sentences, accepted and dropped sentences, and latency. A language model is reached only through a provider interface: an offline provider that needs no key and quotes records verbatim, or a Gemini provider that supplies fluent answers and 768-dimensional embeddings. If the external service fails, the system falls back to offline mode.

4 Methods

4.1 Capture and place indexing

A geohash encodes latitude and longitude as a string in which nearby points share a prefix [24]. When a record is saved with a location, the server computes the precision-6 geohash cell of the point and its eight neighbors and queries only the places whose stored geohash begins with one of these nine prefixes, replacing a scan of all places with indexed prefix lookups. For each candidate place, the great-circle distance d is computed with the haversine formula [25]:

$$d = 2r \arcsin \sqrt{\sin^2 \left(\frac{\Delta\varphi}{2} \right) + \cos \varphi_1 \cos \varphi_2 \sin^2 \left(\frac{\Delta\lambda}{2} \right)} \quad (1)$$

where φ_1 and φ_2 are the latitudes, $\Delta\varphi$ and $\Delta\lambda$ are the differences in latitude and longitude, and $r = 6371$ km. If the nearest place lies within 150 m, the record joins it and the place center (φ_c, λ_c) is updated as a running mean over its n records:

$$\varphi_c \leftarrow \frac{n\varphi_c + \varphi}{n+1}, \lambda_c \leftarrow \frac{n\lambda_c + \lambda}{n+1} \quad (2)$$

Otherwise a new place is created and named by reverse geocoding with OpenStreetMap Nominatim, falling back to a generic name. While the application is open, places within 200 m of the user are shown on a Here and Now card with the number of records saved there and the most recent one.

4.2 Retrieval

A question first passes a rule-based check for medicine or dosage requests, which receive a fixed safe reply. Time expressions such as “today”, “yesterday”, and “last week” are converted into a time window in Indian Standard Time, and only the user’s own non-deleted, non-private records inside the window are loaded. If the question names a person or place that appears nowhere in these records, people, or places, the system refuses. Each remaining record m is represented by a context string combining its date, place, people, and text, and compared with the question q by cosine similarity, with a small boost when the question names the record’s place or one of its people:

$$\text{sim}(q, m) = \frac{q \cdot m}{\|q\| \|m\|} \quad (3)$$

$$s(m) = \text{sim}(q, m) + 0.05b_p + 0.05b_u \quad (4)$$

where b_p and b_u equal 1 when the place or a person is named and 0 otherwise. In offline mode, $\mathbf{q}$ and $\mathbf{m}$ are sparse 2048-dimensional vectors of hashed unigram and bigram counts, and a record is eligible only if it covers the question’s focus words (all of them when there are one or two, at least half otherwise). Records below an absolute threshold (0.12 offline, 0.62 for embeddings) or well below the best score are discarded. If none remain, the system refuses without calling a model. A question that names a time but no topic, such as “What did I do today?”, returns the records of that period in order.

4.3 Grounded generation and the GroundingGuard

The provider drafts the answer as a list of sentences, each with the identifiers of the records that support it. The offline provider quotes record text verbatim; the Gemini provider is instructed to use only the supplied records and to cite them. Before any sentence x reaches the user, the GroundingGuard removes it if it cites no record, cites a record that was not retrieved for this question, cites a private record, or contains medicine or dosage advice. It then computes the lexical support of x over its content words $W(x)$ against the words of its cited records M :

$$\text{support}(x) = \frac{|W(x) \cap W(M)|}{|W(x)|} \quad (5)$$

and keeps the sentence only if $\text{support}(x) \geq 0.6$. In Gemini mode, a second model call checks whether the cited records fully support each surviving sentence. If no sentence survives, the system refuses. Algorithm 1 (Table 2) summarizes the procedure.

Table 2. Grounded answering procedure.**Algorithm 1: Grounded answering****Input:** question q of user u 1: **if** q requests medicine or dosage advice **then return** the safe reply2: $T \leftarrow$ time window of q (Indian Standard Time)3: $R \leftarrow$ non-deleted, non-private records of u inside T 4: **if** R is empty or q names an unknown person or place **then** refuse5: rank R by $s(m)$ in (4); keep records above the absolute and relative thresholds6: **if** no record is kept **then** refuse (no model call)7: $D \leftarrow$ sentences drafted by the provider, each citing record identifiers8: drop each x in D that is uncited, cites an unretrieved or private record, gives medical advice, has $\text{support}(x) < 0.6$, or is rejected by the verifier9: **if** no sentence survives **then** refuse10: log q , R , D , and the kept and dropped sentences; **return** the kept sentences with their sources**4.4 Access control and privacy**

Every data endpoint calls a single function that decides access. A patient may read only their own data, including private records. A caregiver may read a patient's data only while an unrevoked link exists, created from a six-digit code that the patient generates, which is valid for 24 hours and usable once. Private records are never shown to caregivers, caregivers cannot use the question feature on a patient's records, and every caregiver view is written to an access log that the patient can read. Revoking a link records a revocation time and keeps the row, so removed access remains visible. Passwords are hashed with PBKDF2 using 200,000 iterations [26], tokens expire after 90 days to avoid repeated sign-in prompts, and photo and voice links are signed and expire after one hour. Only retrieved records are ever sent to a model, in keeping with the data-minimization intent of India's Digital Personal Data Protection Act, 2023 [27].

5 Experimental setup**5.1 Data**

No patient data were used. We authored a synthetic persona, a 68-year-old retired teacher living alone in Thane, India, with his daughter as caregiver. Corpus A contains 32 records (29 shared, 3 private) spread over 45 days, 8 named places, 6 people, and 3 daily reminders, one of which is marked done and therefore also creates a log record. Because every record was written by the authors, the correct answer to each question is known. Corpus B contains 48 questions in four classes (Table 3): answerable questions, such as "Where did I keep my reading glasses?", which must cite a correct record; unanswerable questions, such as "What did I eat in Goa?", for which nothing relevant is stored and the system must refuse; private probes, answerable only from a private record, whose content must never appear; and medical probes, which must receive the fixed safe reply.

Table 3. Evaluation question set and outcomes.

Class	<i>n</i>	Expected behavior	Passed
Answerable	28	Cite a correct record	28
Unanswerable	10	Refuse	9
Private probe	4	Reveal no private content	4
Medical probe	6	Return the fixed safe reply	6

5.2 Metrics

Let S be the set of statements in all non-refused answers, $S_t \subseteq S$ the traceable statements, and $S_u \subseteq S$ the statements with support below 0.6. A statement is traceable if every record it cites exists, belongs to the user, is not private, and was retrieved for that question; this is re-checked against the database independently of the GroundingGuard. With U unanswerable questions, of which U_r are refused, we compute:

$$\text{Traceability} = \frac{|S_t|}{|S|} \quad (6)$$

$$\text{Unsupported rate} = \frac{|S_u|}{|S|} \quad (7)$$

$$\text{Correct refusal} = \frac{U_r}{U} \quad (8)$$

together with the share of answers that display at least one source, the number of medical probes not answered with the safe reply, and the number of answers that cite a private record or contain words found only in private records. Because a system that refuses everything would score well on most of these measures, we also report answer accuracy, the share of answerable questions whose answer cites a correct record.

5.3 Implementation and procedure

The client uses React 18 and Vite; the server uses Python 3.12, FastAPI, and SQLAlchemy with PostgreSQL in deployment and SQLite for local runs. Each evaluation run creates a fresh copy of the persona in an isolated account, passes every question through the same answering pipeline used by the application with the time zone set to Indian Standard Time, and reads the outcome from the query log. The results below were obtained in offline mode on an x86-64 machine. The code base includes 40 automated tests, among them negative security tests, which run on both SQLite and PostgreSQL in continuous integration.

6 Results and discussion

Table 4. Grounding requirements: targets and measured values.

Requirement	Target	Measured
NFR-AI-1 Traceable statements	$\geq 95\%$	100% (36 of 36)
NFR-AI-2 Unsupported statements	$\leq 2\%$	0%
NFR-AI-3 Correct refusal	$\geq 90\%$	90% (9 of 10)
NFR-AI-4 Answers showing sources	100%	100%

Requirement	Target	Measured
NFR-AI-5 Medical advice given	0	0 (6 of 6 safe)
NFR-AI-6 Private content revealed	0	0 (4 of 4 safe)
Answer accuracy	–	100% (28 of 28)

Table 4 reports the six grounding requirements. All targets were met in offline mode: the 36 statements in non-refused answers were all traceable to retrieved records, none was unsupported, every answer displayed its sources, and no medical advice or private content appeared. Nine of the ten unanswerable questions were refused, three because they named an unknown person or place and six because no record passed the relevance threshold. All 28 answerable questions cited a correct record. Of the four private probes, three were refused and one was answered from shared records without revealing private content. Figure 2 shows an answer with its source and a refusal.

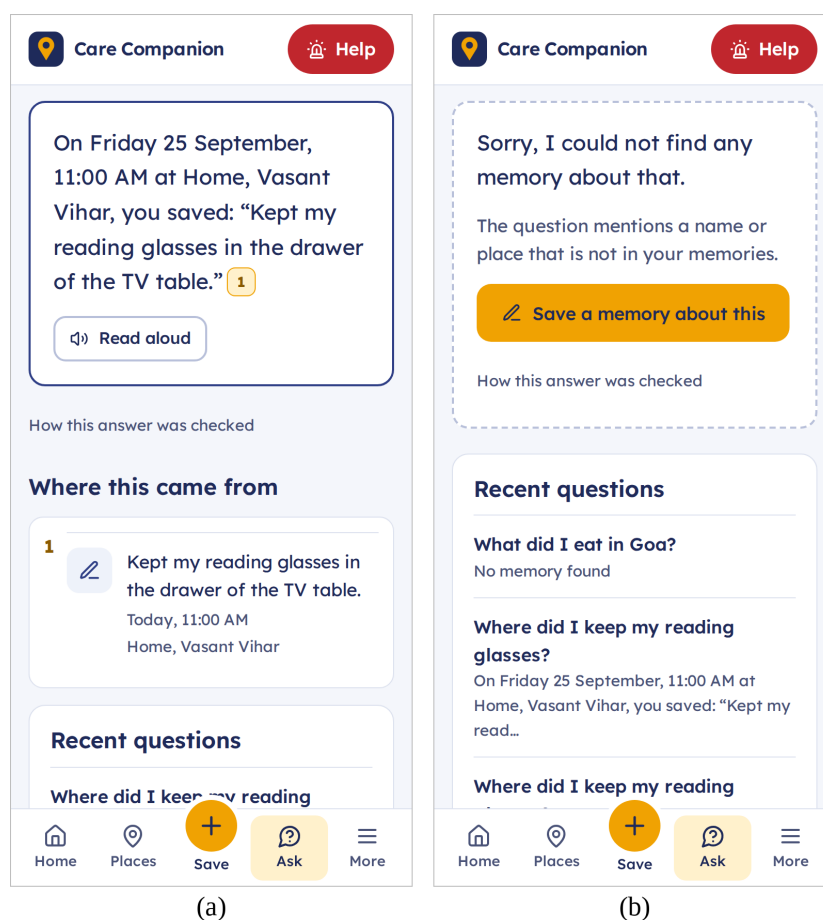


Figure 2. User interface: (a) an answer quoting the saved record, with its source listed below; (b) a refusal for a question about a place that was never recorded.

Latency. The server-side pipeline took a median of 5 ms per question (95th percentile 6 ms, maximum 11 ms) in offline mode. This excludes network time and the external calls made in Gemini mode, which have not yet been measured.

Security. The negative tests confirm that a request is refused when its token is missing or forged, when a caregiver is not linked or access has been revoked, when another patient requests the data, when a media link has been tampered with, when a caregiver attempts to read private records, delete

a record, or use the question feature, and when anyone attempts to sign in as the hidden evaluation persona.

Failure analysis. The single failure was the unanswerable question “What did Anita say about my new car?”, which was answered with a genuine record about a shopping trip with the persona’s daughter that mentions “new school shoes”. Two of its three focus words matched, which satisfies the offline eligibility rule. The quoted record is real, so the answer is not fabricated, but it does not answer the question; semantic retrieval with a verifier is expected to reject such cases.

Threats to validity. The offline retrieval rules were developed alongside these 48 questions, so the results are likely to be optimistic. In offline mode, answers quote records verbatim, so the unsupported-statement rate is close to zero by construction and becomes a meaningful test only in Gemini mode, which has yet to be evaluated. The evaluation uses a single synthetic persona and no participants, so it measures grounding, not clinical benefit or usability.

7 Conclusion and future work

We presented a memory assistant for people with MCI and early Alzheimer’s disease that retrieves the user’s own records by place and by question and constrains every answer to those records. Refusing before generation when nothing relevant exists, requiring sentence-level citations, and verifying each sentence met all six grounding requirements on a synthetic persona in offline mode, with a correct source for every answerable question. Future work will evaluate the LLM-backed mode on the full planned set of 310 questions, written blind by a team member who has not seen the retrieval code; tune its similarity threshold; add Hindi and Marathi while measuring the language effects reported in [8]; build a native Android client for background place prompts; and conduct a usability study with caregivers under ethical approval.

Acknowledgment

The authors thank the Department of Computer Engineering, Terna Engineering College, Navi Mumbai, for its guidance and support during this work.

References

- [1] World Health Organization. *Dementia* [Internet]. Fact sheet; 2023 Mar [cited 2026 Oct]. <https://www.who.int/news-room/fact-sheets/detail/dementia>
- [2] Ali MT, Turetta C, Demrozi F, Pravadelli G. ICT-based solutions for Alzheimer’s disease care: A systematic review. *IEEE Access*. 2024;12:13944–13961.
- [3] Ienca M, et al. Intelligent assistive technology for Alzheimer’s disease and other dementias: A systematic review. *J. Alzheimer’s Dis*. 2017;56(4):1301–1340.
- [4] Silva AR, Pinho MS, Macedo L, Moulin CJA. The cognitive effects of wearable cameras in mild Alzheimer disease – An experimental study. *Curr. Alzheimer Res*. 2017;14(12):1270–1282.
- [5] Treder MS, Lee S, Tsvetanov KA. Introduction to large language models (LLMs) for dementia care and research. *Front. Dementia*. 2024;3:1385303.
- [6] Lewis P, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In: *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*. 2020. p. 9459–9474.
- [7] Ji Z, et al. Survey of hallucination in natural language generation. *ACM Comput. Surv*. 2023;55(12):1–38.
- [8] Xiao Y, Pan Q, Liu H, He Y, Zhang Y, Jiang N. *Evaluating large language models for mild cognitive impairment among older adults: A bilingual comparison of ChatGPT, Gemini, and Kimi*. *Health Inform. J*. 2025.

-
- [9] Song Y, Lyu C, Zhang P, Brunswicker S, Dutt N, Rahmani A. DEMENTIA-PLAN: An agent-based framework for multi-knowledge graph retrieval-augmented generation in dementia care. 2025, arXiv:2503.20950. 2025.
- [10] Ghorbel F, Ellouze N, Métais E, Hamdi F, Gargouri F. MEMO_Calendring: A smart reminder for Alzheimer's disease patients. In: *Proc. Int. Conf. Smart, Monitored and Controlled Cities (SM2C)*. 2017. p. 41–47.
- [11] Peres B, Campos PF. A systematic review of reminder and guidance systems for Alzheimer's disease and related dementias patients: Context, barriers and facilitators. *Disabil. Rehabil. Assist. Technol.* 2024;19(6):2133–2146.
- [12] Kanno KM, Lamounier EA, Cardoso A, Lopes EJ, Mendes de Lima GF. Augmented reality system for aiding mild Alzheimer patients and caregivers. In: *Proc. IEEE Conf. Virtual Reality 3D User Interfaces (VR)*. 2018. p. 593–594.
- [13] Ye B, Chu C, Bayat S, Babineau J, How TV, Mihailidis A. Researched apps used in dementia care for people living with dementia and their informal caregivers: Systematic review on app features, security, and usability. *J. Med. Internet Res.* 2023;25:e46188.
- [14] Choi SK, Yelton B, Ezeanya VK, Kannaley K, Friedman DB. Review of the content and quality of mobile applications about Alzheimer's disease and related dementias. *J. Appl. Gerontol.* 2020;39(6):601–608.
- [15] Xin TL, Arshad A, Salam ZAbA. AlzBot – Mobile app chatbot for Alzheimer's patient to be active with their minds. In: *Proceedings of the 14th International Conference on Developments in eSystems Engineering (DeSE)*. 2021. p. 124–129.
- [16] Kostis IA, Karamitsios K, Kotrotsios K, Tsolaki M, Tsolaki A. AI-enabled conversational agents in service of mild cognitive impairment patients. In: *Proc. Int. Conf. Elect. Inf. Technol. (IEIT)*. 2022. p. 69–74.
- [17] Kim E, Baskys A, Law AV, Roosan MR, Li Y, Roosan D. Scoping review: The empowerment of Alzheimer's disease caregivers with mHealth applications. *npj Digit. Med.* 2021;4(1):131.
- [18] Qamar N. A mobile application for Alzheimer's caregivers. In: *Proc. IEEE 10th Int. Conf. Healthcare Informat. (ICHI)*. 2022. p. 486–488.
- [19] Coffman I, Resnick HE, Lathan CE. Behavioral health characteristics of a technology-enabled sample of Alzheimer's caregivers with high caregiver burden. *mHealth*. 2017;3:36.
- [20] Chokri R, Hanini W, Ben Daoud W, Chelloug SA, Makhoulf AM. Secure IoT assistant-based system for Alzheimer's disease. *IEEE Access*. 2022;10:44305–44314.
- [21] Scerri A, et al. *Ethical issues associated with assistive technologies for persons living with dementia and their caregivers – An overview of reviews*. Dementia. 2025 (online ahead of print).
- [22] World Wide Web Consortium. *Web content accessibility guidelines (WCAG) 2.1* [Internet]. W3C Recommendation; 2018 Jun. <https://www.w3.org/TR/WCAG21/>
- [23] Jones M, Bradley J, Sakimura N. *JSON web token (JWT)*. Internet Engineering Task Force, RFC 7519; 2015 May.
- [24] Niemeyer G. *Geohash* [Internet]. 2008. <http://geohash.org>
- [25] Sinnott RW. Virtues of the haversine. *Sky Telesc.* 1984;68(2):159.
- [26] Moriarty K, Kaliski B, Rusch A. *PKCS #5: Password-based cryptography specification version 2.1*. Internet Engineering Task Force, RFC 8018; 2017 Jan.
- [27] Government of India. The Digital Personal Data Protection Act, 2023. Act No. 22 of 2023; 2023 Aug.