Journal of
Dementia and Alzheimer's Disease

MDPI

Review

Large Language Models in Alzheimer's Care: Clinical Use Cases, Safety Challenges, and Implementation Pathways

Tursun Alkam *, Ebrahim Tarshizi and Andrew H. Van Benschoten

Master's Program of Applied Artificial Intelligence, University of San Diego, San Diego, CA 92110-2492, USA
* Correspondence: talkam@sandiego.edu

Abstract

Background: Alzheimer's disease (AD) care is increasingly communication-intensive, requiring sustained caregiver education, symptom monitoring, behavioral management, and coordination across fragmented clinical settings. Large language models (LLMs) can generate, summarize, and adapt natural language at scale, creating opportunities to support dementia care workflows, but they also introduce safety risks that may be amplified in cognitively vulnerable populations. Methods: We provide a narrative synthesis of emerging applications of LLMs across the AD care continuum, examine how general LLM safety risks may be amplified or modified in Alzheimer's care, and propose pragmatic implementation pathways for responsible clinical translation. Results: LLM use cases cluster into caregiver-facing support, patient-facing conversational agents (high caution), clinician workflow augmentation (highest near-term feasibility), clinical text intelligence for risk prediction (early-stage), and research/education support. Key safety threats include hallucinations, omission of critical information, over-reliance, bias, privacy leakage, and prompt-injection vulnerabilities. Conclusions: Implementation is most defensible using grounded architectures (e.g., retrieval-augmented generation), tiered deployment, human-in-the-loop verification, continuous monitoring, and security testing.

Keywords: Alzheimer's disease; dementia care; large language models; generative AI; caregiver support; retrieval-augmented generation; clinical decision support

1. Introduction

Alzheimer's disease (AD) care is increasingly shaped by long-term, communication-heavy interactions among patients, caregivers, and clinicians. Across the disease continuum, persistent challenges extend beyond diagnosis and pharmacologic decisions to day-to-day symptom interpretation, caregiver coaching, safety and functional planning, navigation of services, and coordination across fragmented settings. These demands generate substantial caregiver workload and large volumes of clinical and caregiver narratives, while workforce constraints can limit individualized education and support [1].

Large language models (LLMs) are general-purpose generative systems capable of producing fluent text, synthesizing long documents, answering free-text questions, and adapting explanations to different audiences [2]. This functionality has immediate appeal in Alzheimer's care because many unmet needs are language-intensive: converting complex plans into caregiver-friendly steps, accelerating chart review, summarizing longitudinal trajectories, and supporting consistent education between visits. However, LLMs can also generate incorrect or fabricated statements ("hallucinations"), and these errors may be difficult to detect because outputs can appear confident and clinically plausible [2,3].

Check for updates

Academic Editor: Junhui Wang

Received: 11 May 2026
Revised: 8 August 2026
Accepted: 14 September 2026
Published: 4 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

Factual reliability therefore remains a primary barrier to safe clinical use, particularly when LLMs generate or transform patient-care documentation [3].

Early dementia-focused evaluations illustrate both the promise and limits of general-purpose conversational models. In a study assessing ChatGPT-3.5s performance for dementia-related information needs, caregivers found the system more helpful for general, non-specialized questions than for clinically nuanced inquiries, reinforcing that caregiver-facing tools cannot be treated as substitutes for clinical guidance [4,5]. This gap has fueled increasing interest in Alzheimer's-specific architectures that emphasize grounded answers rather than unconstrained generation. For example, retrieval-augmented caregiver systems such as ADQueryAid are designed to provide more reliable support by enriching LLM responses with curated AD knowledge and structured retrieval [6]. Similarly, caregiving-specialized models such as the Caregiving Language Model (CaLM) represent a pragmatic approach to domain adaptation using caregiving-specific knowledge resources to improve reliability and accessibility for caregiver tools [7].

In parallel, investigators are exploring higher-complexity applications in clinical text intelligence and prediction, including multi-agent approaches that analyze longitudinal notes to forecast Alzheimer's risk trajectories. CARE-AD is one example of this emerging direction, proposing a multi-agent LLM framework for AD prediction using longitudinal clinical documentation [8]. While scientifically promising, such systems must be interpreted cautiously until they demonstrate external validation, calibration, and consistent performance across diverse healthcare settings and populations.

Beyond hallucinations, Alzheimer's deployment also raises distinct and underappreciated safety threats related to system integration. Prompt injection, maliciously crafted text that manipulates model behavior, has been shown to meaningfully alter medical advice outputs in commercial LLMs and is especially concerning for caregiver-facing tools that ingest external or user-provided content [9]. These risks reinforce the need to view LLM adoption in dementia care as a governed socio-technical intervention, not merely a software upgrade.

Accordingly, this narrative review synthesizes the emerging role of LLMs in Alzheimer's care through three integrated lenses: (1) clinical use cases across the AD continuum, with emphasis on caregiver-facing support, patient-facing conversational agents, clinician workflow augmentation, and text-based prediction; (2) general LLM safety risks and the ways in which cognitive vulnerability, caregiver mediation, fluctuating decisional capacity, and care transitions may amplify their consequences in Alzheimer's care; and (3) implementation pathways that prioritize grounded generation, conservative role design, and staged deployment with continuous monitoring. This approach aligns with broader governance and lifecycle risk-management guidance for generative AI in health care, including recommendations emphasizing ethics, accountability, transparency, and human oversight.

Rather than arguing that LLMs are ready to replace clinical expertise, we propose that the most defensible near-term value in Alzheimer's care lies in assistive infrastructure: tools that reduce documentation burden, improve clarity of caregiver instructions, and support navigation and continuity, while preserving human verification for care-critical decisions. Figure 1 provides a conceptual overview of LLM applications across the Alzheimer's care continuum, summarizing key clinical use cases, safety risks, and implementation pathways that structure the remainder of this review.

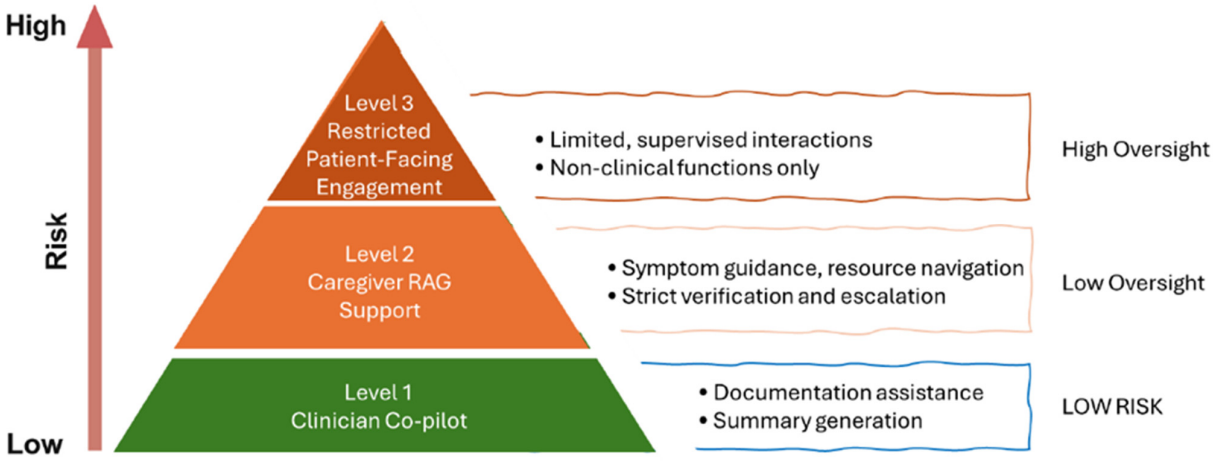


Figure 1. Tiered Deployment Model for Large LLMs in Alzheimer's Care. This figure introduces a structured, risk-stratified deployment framework specifically tailored to cognitively vulnerable populations. The model delineates three progressive levels of implementation: Level 1 (Clinician Co-Pilot), in which LLMs function strictly as documentation and summarization support tools under continuous professional oversight; Level 2 (Caregiver Retrieval-Augmented Generation [RAG] Support), where responses are constrained to source-grounded informational assistance for caregivers; and Level 3 (Restricted Patient-Facing Engagement), representing tightly bounded conversational interactions with predefined guardrails and escalation triggers. A visual gradient illustrates increasing autonomy and corresponding risk exposure across tiers, with oversight intensity increasing as applications move toward direct patient-facing use. The conceptual novelty lies in formalizing autonomy boundaries for dementia care, explicitly linking deployment context to vulnerability, supervision, and permissible system capabilities.

Literature Identification and Scope

For this narrative review, relevant literature was identified through targeted searches of PubMed/MEDLINE and Google Scholar, supplemented by screening the reference lists of relevant articles. Searches focused on combinations of terms related to Alzheimer's disease, dementia, large language models, generative artificial intelligence, caregiver support, clinical applications, safety, and implementation. We primarily considered English-language literature published from January 2023 through July 2026, while earlier publications were included when relevant to established dementia measures, foundational concepts, or reporting frameworks. Peer-reviewed studies were prioritized, together with guidance from recognized professional and public health organizations. Because this was a narrative rather than systematic review, study selection was purposive, and no formal risk-of-bias assessment or meta-analysis was performed.

2. What LLMs Are (And What They Are Not) in Clinical Care

LLMs are generative systems trained to predict and produce language, enabling them to generate coherent text, summarize long documents, answer questions, and adapt explanations to different audiences [2]. In clinical environments, their practical value is strongest for tasks that are fundamentally linguistic, including documentation, translation of medical text into patient-friendly language, and synthesis of narrative information across fragmented records [2]. At the same time, LLMs are not inherently reliable fact engines: they can produce fluent but incorrect statements, including invented details, unless outputs are appropriately grounded and verified [2,3]. This distinction is especially important in dementia contexts, where cognitive vulnerability or caregiver reliance may increase the consequences of persuasive but inaccurate outputs [10,11].

In other words, LLMs can be powerful clinical communication amplifiers, but they should not be treated as autonomous clinical decision-makers. Their safest clinical identity is a co-pilot that drafts and organizes information under human oversight, rather than an agent that independently diagnoses, triages, or prescribes.

2.1. What LLMs Can Do Well

2.1.1. Summarization and Rewriting at Different Literacy Levels

One of the most mature and clinically relevant capabilities of LLMs is transforming complex medical text into language that is easier to understand. This matters in Alzheimer's care because caregivers often become the operational "care team" after encounters and need clear instructions that are actionable at home. In a cross-sectional study of inpatient discharge summaries, an LLM was able to translate discharge documentation into patient-friendly language and formats that were significantly more readable and understandable than EHR versions, while the authors emphasized that real-world implementation requires improvements in accuracy, completeness, and safety, and should begin with physician review [12]. These findings support a practical use case for dementia care: using LLMs to draft caregiver-facing summaries (e.g., medication routines, safety recommendations, follow-up plans) that can then be verified and finalized by clinicians.

2.1.2. Extracting the "Clinical Story" from Notes

Clinical records are frequently fragmented, repetitive, and difficult to synthesize, especially for older adults with multimorbidity and frequent transitions. LLMs can help convert note-heavy records into problem-oriented narratives that summarize what happened, what changed, what matters now, and what the plan is; this may reduce review burden, although the effect on clinician cognitive workload requires direct evaluation [13,14]. A 2025 evaluation described the development and assessment of a clinical note summarization approach and highlighted the motivation for using LLMs to support physician review workflows, while reinforcing that these tools complement rather than replace clinical judgment [13]. In parallel, recent work has proposed methods to evaluate real-world multi-document EHR summaries, reflecting a broader shift toward structured assessment of summary quality and factual reliability rather than anecdotal impressions [14]. For Alzheimer's care, this form of "clinical story extraction" may be useful when care plans must be communicated across settings (clinic → ED → inpatient → post-acute → home) and caregiver narratives are essential for contextualizing symptoms.

2.1.3. Generating Caregiver Instructions and Checklists (With Guardrails)

LLMs can draft caregiver instructions, routines, and checklists that reduce ambiguity and improve adherence to care plans, especially when outputs are constrained to supportive, non-prescriptive domains (home safety steps, routine planning, symptom tracking logs, appointment question lists). Early caregiver-facing evaluations show real promise but also reveal the boundary of safe reliability. In a Scientific Reports evaluation, caregivers rated ChatGPT as more responsive and helpful for general, non-specialized dementia information needs than for clinically specialized questions, suggesting that LLMs may assist with education and planning but are not dependable as standalone clinical advisors [5]. This pattern is highly relevant to Alzheimer's care, where caregiver tools should prioritize grounded education, consistent tone, and escalation pathways rather than "open-ended medical advice."

2.1.4. Triage and Documentation Assistance (Co-Pilot Role)

The highest near-term feasibility for LLMs in medicine is clinician-facing workflow augmentation, particularly drafting clinical text that clinicians can verify [2]. In a com-

parative effectiveness study of emergency department discharge documentation, use of an on-site LLM assistant was associated with significantly reduced writing time and with LLM-assisted notes rated as more complete and clinically useful than manual notes [15]. This type of co-pilot workflow may be applicable to AD-related emergency and hospital-discharge settings, particularly where caregivers receive and operationalize discharge instructions; however, effectiveness and safety should not be assumed without dementia-specific evaluation [1,15]. In safe implementations, LLMs accelerate the first draft rather than replace clinician responsibility for the final clinical content.

2.2. Where LLMs Fail in Dangerous Ways

2.2.1. Hallucinations and Fabricated Clinical Facts

A widely recognized LLM safety failure mode is hallucination: generation of content that appears plausible but is unsupported or false [2,3,16]. In clinical workflows, hallucinations may appear as invented diagnoses, incorrect medications, fabricated timelines, or erroneous clinical recommendations. Evidence from patient-care documentation has emphasized that such inaccuracies remain a central barrier to safe clinical use, reinforcing that fact-checking and verification must be built into deployment strategies [3]. In Alzheimer's care, the consequences may be amplified when caregivers act on written instructions as if they were clinician-approved plans or when patients have reduced ability to question inaccuracies [10,11].

2.2.2. Confident Incorrectness and "Authority Tone"

LLMs can produce incorrect statements in fluent and confident language, which may create an "authority tone" and encourage over-trust [2,17,18]. In clinical settings, this can contribute to automation bias when users give model output undue weight relative to their own judgment or contradictory evidence [17,18]. In dementia contexts, the consequences may be greater when caregivers are seeking definitive guidance or when patients are more vulnerable to persuasive conversational framing [11]. The safest approach is therefore to constrain LLM roles to drafting and explanation while ensuring escalation rules and human oversight for care-critical decisions.

2.2.3. Inconsistent Performance Across Populations and Settings

LLM behavior is not uniform across patient populations, health literacy levels, language styles, and institutional contexts. This raises the risk that a system might provide safer, more complete, or more empathetic guidance for some groups than others, thereby reinforcing existing disparities in dementia care access and outcomes. A "toolbox" paper emphasizes that equity-related model failures are common and difficult to detect without intentional evaluation, and it proposes structured approaches to surface health equity harms and biases in LLMs used for health information needs [19]. Empirical work has also shown measurable bias patterns in health-oriented LLM outputs and decisions, underscoring that subgroup performance auditing cannot be optional if these tools are used at scale [20]. In Alzheimer's care, this is particularly important because caregiver support needs are tightly coupled to socioeconomic resources, language access, and health literacy.

2.2.4. Sensitivity to Prompt Injections and Unsafe Instructions in Downstream Workflows

As LLMs become embedded in tools that retrieve external content or process user-provided documents, they become vulnerable to prompt injections, maliciously crafted text that can override system instructions and induce unsafe outputs. A quality improvement study demonstrated that commercial LLMs can be highly susceptible to prompt-injection attacks that alter medical advice and validated client-side "man-in-the-middle" injection as a realistic attack vector [9]. This threat is especially relevant in Alzheimer's caregiver

support systems, which may ingest community resources, pasted text, portal messages, or third-party documents. Without security-by-design safeguards (source allow-lists, input sanitization, constrained tool permissions, and adversarial testing), prompt injection becomes a direct clinical safety risk rather than a purely technical concern.

To operationalize these strengths and failure modes, we propose a multidimensional evaluation framework spanning technical validity, clinical safety, usability and cognitive alignment, equity and fairness, privacy/consent/governance, implementation and monitoring, and patient-facing safety (Table 1).

Table 1. Multidimensional Evaluation Framework for LLMs in Alzheimer's Care.

Evaluation Domain	Core Questions	Primary Metrics/Outcomes	Alzheimer's-Specific Considerations	Recommended Study Designs & Reporting Standards
Technical Validity	Are outputs accurate, complete, reproducible, and calibrated?	Accuracy; hallucination and omission rates; factual verification; calibration for predictive models	Omissions (e.g., delirium or fall precautions) may be as harmful as fabrication; authoritative tone can amplify risk	Blinded expert adjudication; hallucination/omission audits; external validation; TRIPOD + AI [21]
Clinical Safety	Does the system avoid introducing new risks or delaying appropriate escalation?	Safety-critical error rate; escalation accuracy; unsafe-advice frequency; clinician override rate	Caregivers may operationalize outputs as care plans; failed escalation can carry amplified harm	Controlled before–after studies; pragmatic trials; DECIDE-AI [22]; CONSORT-AI [23]
Usability & Cognitive Alignment	Are outputs understandable, usable, and aligned with caregiver and patient cognitive capacity?	Readability; caregiver comprehension; System Usability Scale (SUS); time saved; satisfaction	Literacy variability, caregiver stress, and fluctuating capacity require simplification without loss of safety-critical detail	Mixed-method usability studies; caregiver-supervised feasibility trials; SPIRIT-AI [24]
Equity & Fairness	Does performance remain stable across demographic, linguistic, and literacy subgroups?	Subgroup error rates; disparities in unsafe outputs; calibration across strata; cultural-alignment assessment	Existing disparities in diagnosis and caregiver support may be worsened by uneven model performance	Stratified validation; fairness audits; predefined subgroup analyses [25,26]
Privacy, Consent & Governance	Are sensitive narratives protected and role boundaries clearly defined?	Data minimization; access control; audit logs; clarity of consent documentation	AD care includes sensitive home and caregiver narratives; patient capacity may fluctuate	Privacy impact assessment; governance review; role-based workflow testing [27–31]
Implementation & Monitoring	Does performance remain stable over time and across care settings?	Post-deployment hallucination drift; update-related failures; user feedback; sustained adoption	Transitions between ED, hospital, post-acute, and community care may introduce domain shift	Continuous monitoring; periodic audits; time-split and multi-site validation; lifecycle governance [32–39]
Patient-Facing Safety	Does direct interaction avoid confusion, dependency, or harmful advice?	Confusion/distress episodes; dependency indicators; unsafe directives; escalation compliance	Cognitive vulnerability may increase emotional dependency or misinterpretation of authoritative language	Short supervised feasibility trials; conservative scope limits; predefined escalation triggers [10,11,31]

Table 1. *Cont.*

Evaluation Domain	Core Questions	Primary Metrics/Outcomes	Alzheimer's-Specific Considerations	Recommended Study Designs & Reporting Standards
Technical Validity	Does the system produce factually accurate, complete, and reproducible outputs?	Accuracy; hallucination rate; omission rate; factual verification score; calibration (for predictive models)	Omissions (e.g., missed delirium or fall precautions) may be as harmful as fabricated content; overconfident tone increases risk in cognitively vulnerable users	Blinded expert adjudication; structured hallucination/omission audits; external validation; TRIPOD + AI for prediction models [37]
Clinical Safety	Does use of the system avoid introducing new safety risks or delaying appropriate escalation?	Safety-critical error rate; escalation accuracy; unsafe advice frequency; clinician override rate	Caregivers may operationalize outputs as care plans; escalation failures (e.g., failure to recommend urgent evaluation) carry amplified harm	Controlled before–after studies; pragmatic trials; DECIDE-AI early-stage evaluation [34]; CONSORT-AI for interventional trials [35]
Usability & Cognitive Alignment	Are outputs understandable, usable, and aligned with caregiver and patient cognitive capacity?	Readability scores; caregiver comprehension; System Usability Scale (SUS); time saved; satisfaction	Literacy variability, caregiver stress, fluctuating decisional capacity; simplified language must not sacrifice safety-critical detail	Mixed-method usability studies; caregiver-supervised feasibility trials; SPIRIT-AI for protocol transparency [36]
Equity & Fairness	Does performance remain stable across demographic, linguistic, and literacy subgroups?	Subgroup error rates; disparity in unsafe outputs; calibration across strata; qualitative cultural alignment assessment	Dementia diagnosis and caregiving burden already show disparities; uneven performance may exacerbate inequities	Stratified validation; fairness audits; predefined subgroup analyses [15,16]
Privacy, Consent & Governance	Are sensitive narratives protected, and are role boundaries clearly defined?	Data minimization adherence; access control integrity; audit logs; consent documentation clarity	AD care includes sensitive home narratives and caregiver stress disclosures; patient capacity may fluctuate	Privacy impact assessments; governance review; role-based workflow testing [12,17,18]
Implementation & Monitoring	Does system performance remain stable over time and across care settings?	Post-deployment hallucination drift; update-related failure modes; user feedback trends; sustained adoption	Workflow shifts (ED discharge, hospitalization, community care) may induce domain drift; updates can introduce new risks	Continuous monitoring with periodic audits; time-split and multi-site validation; lifecycle governance frameworks [31,32]
Patient-Facing Safety (if applicable)	Does direct interaction avoid confusion, dependency, or harmful advice?	Confusion/distress episodes; dependency indicators; unsafe directive frequency; escalation compliance	Cognitive vulnerability increases risk of emotional dependence or misinterpretation of authoritative tone	Short supervised feasibility trials; conservative scope limitation; predefined escalation triggers [18,40]

3. Clinical Use Cases Across the Alzheimer's Care Continuum

Alzheimer's care is uniquely "language intensive." Patients and caregivers must repeatedly interpret symptoms, communicate evolving needs, coordinate fragmented services, and translate clinical recommendations into day-to-day routines. Digital health technologies have been explored for cognitive impairment care, but the literature remains relatively recent, heterogeneous, and uneven in quality, with persistent gaps in scalability

and sustained engagement [41]. In this context, LLMs are attractive because they can deliver interactive dialogue, personalized explanations, and structured summaries, potentially reducing burden at multiple points across the AD trajectory. However, these systems must be framed as assistive infrastructure, not independent clinical authorities, given well-documented risks of misinformation and overconfidence in generated outputs, risks amplified in cognitively vulnerable populations [5,10]. As summarized in Supplementary Table S1, use cases cluster into caregiver-facing support, patient-facing agents (high caution), clinician workflow augmentation, clinical text intelligence, and research/education applications, which we discuss below.

3.1. Caregiver-Facing Support Systems

Caregiver-facing applications represent one of the most clinically plausible near-term pathways for LLMs in Alzheimer's care because they target high-frequency needs: learning how to respond to symptoms, planning daily routines, ensuring safety, and communicating effectively with clinicians and family. Digital caregiver support tools have shown potential to reduce burden and improve information access, but the landscape remains fragmented, suggesting room for more adaptive and scalable approaches [41]. Within this space, LLMs are particularly suited to education, question answering, and communication support, provided that responses are grounded in trusted resources and designed with clear safety boundaries.

A key proof-of-concept comes from an evaluation of ChatGPT for dementia-related information needs, where caregivers rated the system as generally responsive and helpful, especially for general or non-specialized questions, but noted weaker performance for more specialized, clinically nuanced inquiries [5]. This distinction is critical for Alzheimer's care: many caregiver needs involve practical "how-to" questions and emotional support, but clinical decision-making (medication changes, red flags, differential diagnosis) demands higher reliability than current general-purpose systems consistently provide.

Beyond open-ended question and answer (Q&A), caregivers frequently need help navigating services, planning daily care, and organizing stepwise routines such as safety checklists, behavior logs, and appointment preparation [1,41]. Retrieval-augmented generation (RAG) is especially relevant because it retrieves information from designated knowledge sources and uses that material to ground model responses, thereby reducing reliance on unconstrained generation [42,43]. ADQueryAid exemplifies this approach: it is a conversational system designed to support AD and related dementias (ADRD) caregivers by combining an LLM with ADRD knowledge resources through RAG [6]. These architectures are particularly aligned with caregiver guidance tasks because they prioritize source-grounded responses, although retrieval quality and source curation remain important determinants of reliability [42,43].

A third direction is specialization, rather than relying exclusively on large general-purpose chatbots, investigators have explored caregiving-specific models tailored to the caregiving domain. The Caregiving Language Model (CaLM) illustrates how a targeted caregiving knowledge base, retrieval module, and model adaptation can support caregiver-oriented tools while potentially reducing computational burden compared with relying on the largest foundation models alone [7]. Together, these developments suggest that caregiver-facing LLM systems may be most successful when they (1) focus on high-frequency informational needs, (2) incorporate grounding via RAG, and (3) adopt caregiver-centered conversational policies that emphasize empathy, clarity, and escalation when risk is high.

Finally, LLMs can support communication templates that help caregivers organize and convey critical information. Practical examples include symptom timelines, medication

reconciliation checklists, lists of questions for clinicians, and structured summaries for family coordination. Because caregiver–clinician communication is often fragmented across transitions, templated outputs can improve completeness and reduce misunderstandings, so long as caregivers remain the final editors and clinicians retain accountability for medical decisions [41].

3.2. Patient-Facing Conversational Agents (High Caution Zone)

Patient-facing conversational agents may appear appealing for companionship, cognitive engagement, and reminders; however, direct deployment for individuals with AD remains a high-caution domain. A systematic review of commercially available dementia-related chatbots highlighted variability in quality, features, and content, underscoring that many systems are not designed with the rigorous safeguards required for cognitively vulnerable users [10]. In AD, impaired memory, fluctuating insight, and dependence on caregivers may compound general AI risks by increasing the likelihood of confusion, distress, or inappropriate trust in confident-sounding outputs [10,11].

In the mild or early stages, conversational tools may have a role in structured companionship and social support, particularly when framed as engagement aids rather than sources of authoritative advice. Similarly, cognitive stimulation prompts represent a more defensible category when they are deliberately bounded (e.g., reminiscence-style conversation starters, guided storytelling, mood-supportive prompts). For example, a conversational system has been developed to support reminiscence therapy for people with Alzheimer's disease by enabling personalized dialogue grounded in the patient's life history and preferences, illustrating the feasibility of patient-tailored engagement experiences [40].

Patient-facing adherence reminders (medication timing, hydration prompts, appointments) may also be useful, but safest implementation typically involves caregiver mediation or supervision, especially as cognitive impairment progresses. The broader dementia chatbot literature suggests feasibility for supportive functions, yet it also highlights the need for careful adaptation to abilities and safety constraints [10]. Critically, patient-facing systems should remain assistive rather than directive: they must avoid medication changes, discourage self-management of urgent symptoms, and incorporate clear escalation instructions and caregiver involvement as default safeguards. Evidence from caregiver-facing evaluations further supports this caution, showing that even helpful general responses may be less reliable for clinically complex situations [5].

3.3. Clinician Workflow Augmentation (Highest Near-Term Feasibility)

Among all AD-related applications, clinician-facing workflow augmentation is arguably the most feasible near-term use case because outputs can be verified by professionals and integrated into existing accountability structures. Alzheimer's care frequently requires synthesis of fragmented histories across settings, reconciliation of complex medication regimens, and caregiver-friendly communication, all within a care environment characterized by substantial caregiver burden and navigation demands [1,41]. LLMs may support these tasks through accelerated chart review, structured summaries, and drafting of patient-facing materials, but dementia-specific effectiveness and safety should be evaluated directly.

For chart review acceleration, LLMs can generate problem-oriented narratives that highlight key diagnoses, recent events, and care plans from lengthy documentation, supporting continuity across clinics, hospitals, and emergency encounters. A particularly practical use case is drafting and improving the readability of discharge documentation. In a JAMA Network Open comparative effectiveness study, use of an on-site LLM assistant was associated with reduced writing time for emergency department discharge notes compared with manual note-taking, without compromising documentation quality [15].

Separately, a cross-sectional study showed that LLMs can translate discharge summaries into more patient-friendly language that is significantly more readable and understandable than typical EHR discharge documentation, while still requiring physician review for safety and completeness [12].

In AD care specifically, these functions matter because caregivers often become the operational "care team," and comprehension of discharge instructions and follow-up plans can directly influence safety outcomes. Patient-friendly summaries of diagnosis, prognosis, and next steps may improve caregiver understanding, especially when paired with grounded references to approved educational materials. However, clinician-facing LLM outputs must always be treated as drafts requiring verification, since even subtle inaccuracies or invented details can have downstream consequences. This verification principle becomes even more important when LLM tools offer visit note structuring or differential prompts, where plausible but unsupported suggestions could bias clinical reasoning or documentation [12,15].

3.4. Risk Prediction and Clinical Text Intelligence

Beyond communication and documentation, LLMs are increasingly being explored for clinical text intelligence, including extraction of predictive signals from unstructured narrative notes that may precede formal diagnostic coding. This direction is potentially relevant to prodromal and early symptomatic Alzheimer's disease, in which cognitive and functional changes can emerge gradually across a longitudinal disease continuum [44]. Longitudinal notes may contain caregiver concerns, functional changes, clinician impressions, and behavioral symptoms that are incompletely represented by structured variables [8]. However, evidence supporting LLM-based identification of prodromal AD remains preliminary, and these approaches should not be interpreted as validated tools for autonomous early diagnosis.

CARE-AD (Collaborative Analysis and Risk Evaluation for Alzheimer's Disease) exemplifies this emerging approach: it is a multi-agent LLM-based framework designed to forecast Alzheimer's disease onset by analyzing longitudinal electronic health record notes [8]. Multi-agent architectures are conceptually appealing because they can distribute subtasks (e.g., symptom extraction, risk reasoning, evidence aggregation) across specialized agents, potentially improving robustness compared with single-pass generation [8].

Nevertheless, prediction-focused LLM systems remain early-stage for real-world deployment. External validation across health systems, careful calibration, and subgroup performance auditing are essential because documentation style and access patterns vary widely. Even strong retrospective performance can fail when models encounter new populations, different note structures, or evolving clinical language. As a result, LLM-based prediction should currently be framed as an investigational tool that requires rigorous evaluation before it can influence patient-facing decisions [8].

3.5. Research and Education Use Cases

LLMs may also support Alzheimer's care indirectly through research and education workflows, including literature summarization, drafting of educational materials, and generation of structured outlines for training or caregiver resources [2]. In Alzheimer's contexts, these functions could reduce the time required to prepare practical materials such as safety-planning handouts, "what to expect" guides, and caregiver conversation scripts. However, LLM-based synthesis can omit limitations, overgeneralize evidence, or generate inaccurate references, so verification against primary sources remains necessary [2,45].

A particularly salient concern is reference accuracy in scientific writing. In a large analysis of LLM-generated references, hallucination and citation inaccuracies were substantial, reinforcing that LLM outputs should not be treated as reliable citation engines and must be checked against primary sources [45]. For patient-facing education handouts, the safest pathway is to use LLMs as drafting assistants that rewrite or simplify vetted content, preferably in a retrieval-grounded workflow and reviewed by clinicians or trained staff before distribution [6,12].

4. Safety Challenges of LLM Use in Alzheimer's Care: General Risks Amplified by Dementia-Specific Vulnerabilities

Many safety risks discussed in this review, including hallucinations, omissions, automation bias, bias, privacy leakage, and prompt injection, are general risks of clinical LLM use rather than risks unique to Alzheimer's disease. Their consequences may nevertheless be amplified in AD because cognitive impairment can limit a patient's ability to evaluate model outputs, caregivers frequently mediate day-to-day decisions, decisional capacity may fluctuate, and care often spans multiple settings [3,10,11]. As summarized in Supplementary Table S2, we therefore focus on how these general failure modes interact with dementia-specific vulnerabilities, caregiver-mediated workflows, and high-stakes care transitions.

4.1. Hallucinations, Omissions, and Factuality Failures in Care-Critical Content

A defining limitation of LLMs is their tendency to produce hallucinations, statements that sound plausible but are not supported by the source input (or are simply false). In clinical contexts, hallucinations may appear as fabricated diagnoses, invented medication changes, incorrect timelines, or inappropriate care instructions. This is particularly dangerous in Alzheimer's care because caregivers often rely on written information to implement safety plans and manage medications at home. It has been emphasized that factual inaccuracies in LLM-generated patient care documents remain a major barrier to safe clinical adoption, motivating the need for structured verification workflows rather than trust in fluent outputs [3].

Importantly, factuality risk is not limited to outright fabrication. A parallel problem is omission, missing critical details during summarization (e.g., delirium red flags, fall risk precautions, anticoagulation status, or recent medication adjustments). A recent clinical safety framework highlights hallucinations and omissions as distinct error modes that require deliberate evaluation strategies because both can cause harm even when text appears coherent and clinically "reasonable" [16].

These factuality failures can carry heightened consequences in AD care because caregiver instructions and discharge summaries often function as the practical care plan at home. If an LLM inserts an incorrect medication dose, omits aspiration precautions, or fails to include clear escalation guidance, serious harm may result even when the language appears polished and clinically appropriate [3,16].

These safety domains interact within a broader socio-technical system that extends beyond the model itself, encompassing users, workflows, governance structures, and external information sources (Figure 2).

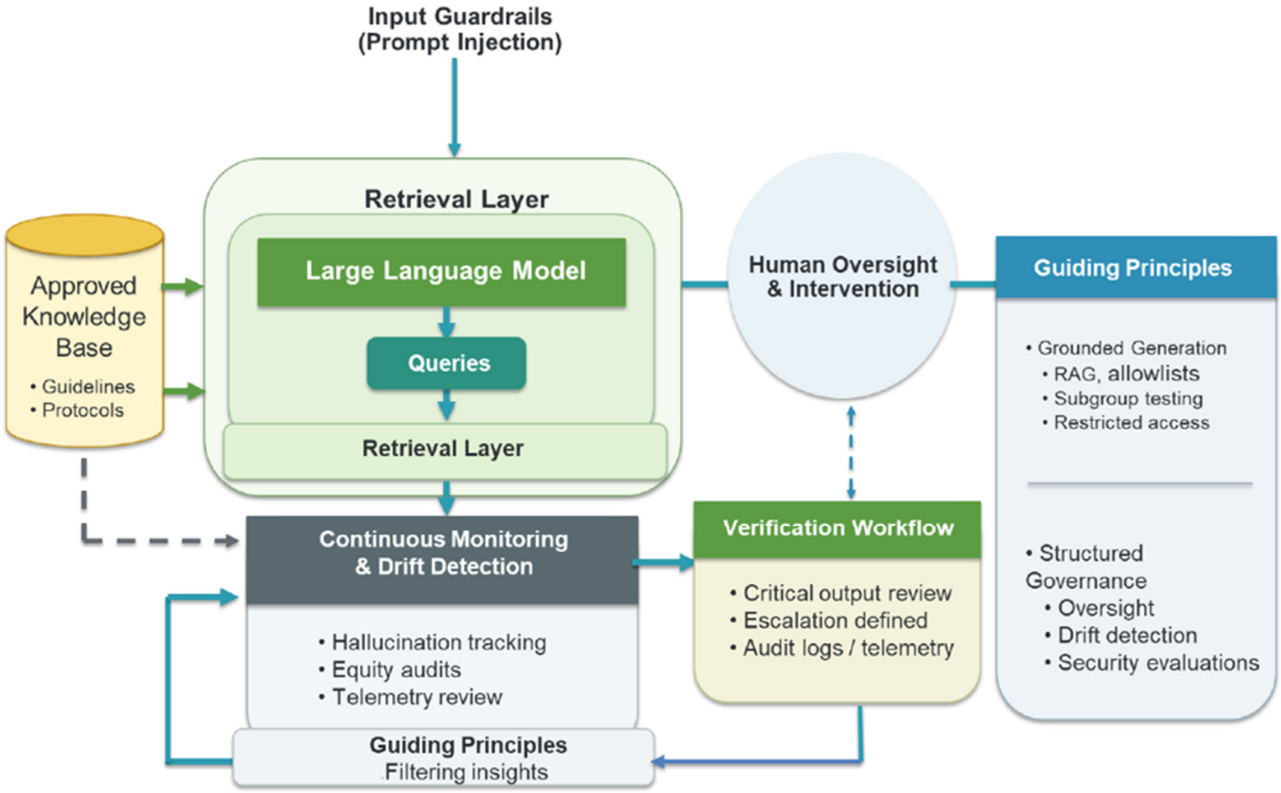


Figure 2. Socio-Technical Safeguard Architecture for LLM Implementation in Alzheimer's Care. This diagram proposes a layered safety architecture integrating model-level safeguards (prompt constraints, retrieval grounding, hallucination detection, audit logging) with human and institutional oversight (clinician review loops, caregiver mediation, governance policies, and escalation pathways). The framework emphasizes that safety is an emergent property of coordinated human–AI interaction rather than a function of model performance alone. Unlike generic AI governance models, this architecture is adapted to dementia contexts, where impaired insight, caregiver mediation, and fluctuating decisional capacity require structured supervision. The bidirectional feedback loops depicted in the model illustrate continuous monitoring and iterative recalibration, underscoring a systems-based approach to deployment.

4.2. Overconfidence, "Authority Tone," and Over-Reliance in Cognitively Vulnerable Users

LLMs frequently express uncertainty poorly, producing incorrect recommendations with confident, authoritative language. In general medicine, this creates a risk of automation bias, with humans trusting AI output over their own judgment or over contradictory evidence. In AD care, this risk is magnified because patients may have reduced capacity to appraise credibility, while caregivers may be overwhelmed and more likely to accept confident, step-by-step recommendations as safe "medical advice" [11].

Evidence from physician-facing contexts also supports caution: randomized evaluation has shown that LLM access can influence diagnostic reasoning and does not uniformly improve performance in a way that eliminates error risk, reinforcing that safe use requires structured workflows, verification, and clear role definition (assistive vs. decision-making) [17].

In patient-facing use, emotional engagement can further intensify reliance. Ethical analyses in older-adult assistive technology emphasize risks such as dependency, manipulation, and diminished autonomy, concerns that become particularly salient when conversational systems are used as "companions" for individuals with dementia [11].

4.3. Prompt Injection and Adversarial Manipulation in Caregiver and Clinical Tools

As health systems deploy LLMs in tools that interface with external content (web sources, PDFs, guidelines, portals, EHR extracts), they become vulnerable to prompt-injection, malicious or hidden instructions that hijack model behavior and can induce unsafe recommendations. A quality improvement study demonstrated that commercial LLMs can be susceptible to prompt-injection attacks capable of altering medical advice, and it validated client-side "man-in-the-middle" injection as a realistic attack pathway [9].

This matters in Alzheimer's care because caregiver-facing systems may retrieve resource information from community sites, support forums, or institutional portals, creating opportunities for hostile or misleading content to manipulate outputs. Prompt injection is therefore not a theoretical cybersecurity concern; it is a safety risk requiring adversarial testing, input sanitization, restricted tool permissions, and conservative autonomy (especially for patient/caregiver-facing deployment) [9].

4.4. Bias, Inequity, and Differential Performance Across Populations

LLM behavior can vary across demographic groups, language styles, and cultural contexts, creating the risk of inequitable care guidance. Alzheimer's and dementia care already contains well-known disparities in diagnosis timing, care access, and caregiver support; LLMs could unintentionally reinforce these patterns if their outputs differ systematically by patient identity cues, dialect, or socioeconomic framing [46,47].

Recent research and reviews have systematically examined demographic disparities and biases in LLMs, including how bias is measured and what mitigation strategies are feasible [25]. Additionally, a toolbox-style approach has been proposed to surface and mitigate health equity harms in LLMs, underscoring that "fairness" must be treated as an explicit design and evaluation requirement rather than an afterthought [26].

In AD care, equity concerns become practical and immediate: caregiver education, symptom interpretation, behavioral guidance, and navigation support must be consistently safe and culturally appropriate, particularly for families with limited health literacy or non-native English proficiency [26,47].

4.5. Privacy, Confidentiality, and Sensitive Caregiver Narratives

Alzheimer's care involves highly sensitive information: cognitive decline trajectories, functional impairment, home safety issues, caregiver strain, behavioral symptoms, and family conflict. LLMs used in messaging, coaching, or planning tools may inadvertently capture or expose details that patients would not have shared in clinical settings. Ethical and legal analyses in long-term and post-acute care emphasize that AI systems must be evaluated within the realities of caregiving environments, where privacy risks and downstream harms differ from typical outpatient documentation workflows [48].

For AD-focused LLM deployment, privacy safeguards are not optional add-ons; they are fundamental requirements. These include data minimization, strict access controls, clear retention policies, audit trails, and careful separation of caregiver notes from patient records when appropriate [11,48].

4.6. Consent, Capacity, Accountability, and Emotional Safety

A central challenge in Alzheimer's care is that decision-making capacity can fluctuate or decline, making consent and accountability complex. Ethical analyses of AI in elderly care emphasize recurring concerns around autonomy, safety, privacy, and responsible use, issues that intensify in dementia when patients may be unable to evaluate tool limitations or recognize misinformation [11].

Additionally, conversational systems may produce emotionally inappropriate responses, reinforce misunderstandings, or contribute to distress, particularly when used as companions or quasi-therapeutic supports. Broader ethical work on older adults and dementia-focused assistive technologies highlights risks such as dependency, replacement of human care, and insufficient stakeholder involvement in assessing harms and benefits [49].

The most defensible framing, therefore, is that LLMs should support communication and planning within a governed care relationship (caregiver + clinician oversight), rather than function as an independent authority in patient-facing contexts [3,11].

5. Implementation Pathways for LLMs in Alzheimer's Care (From Pilots to Safe Deployment)

AD is a high-stakes environment for language technologies because care is mediated through conversations, written instructions, caregiver involvement, and longitudinal narratives shaped by cognitive vulnerability and frequent care transitions [1]. Implementation success therefore depends less on fluent generation than on the surrounding sociotechnical system's ability to prevent unsafe outputs, enable verification, protect privacy, and sustain reliability in real-world settings. Global guidance for generative AI in health care emphasizes explicit governance, transparency, accountability, and continuous risk management, particularly when outputs may influence clinical decisions or patient and caregiver behavior. Supplementary Table S3 outlines a tiered roadmap that prioritizes feasibility and safety by gradually expanding user-facing autonomy. To clarify how these risks may translate into downstream consequences in dementia care, Figure 3 maps representative failure modes to mitigation strategies embedded within the proposed governance model.

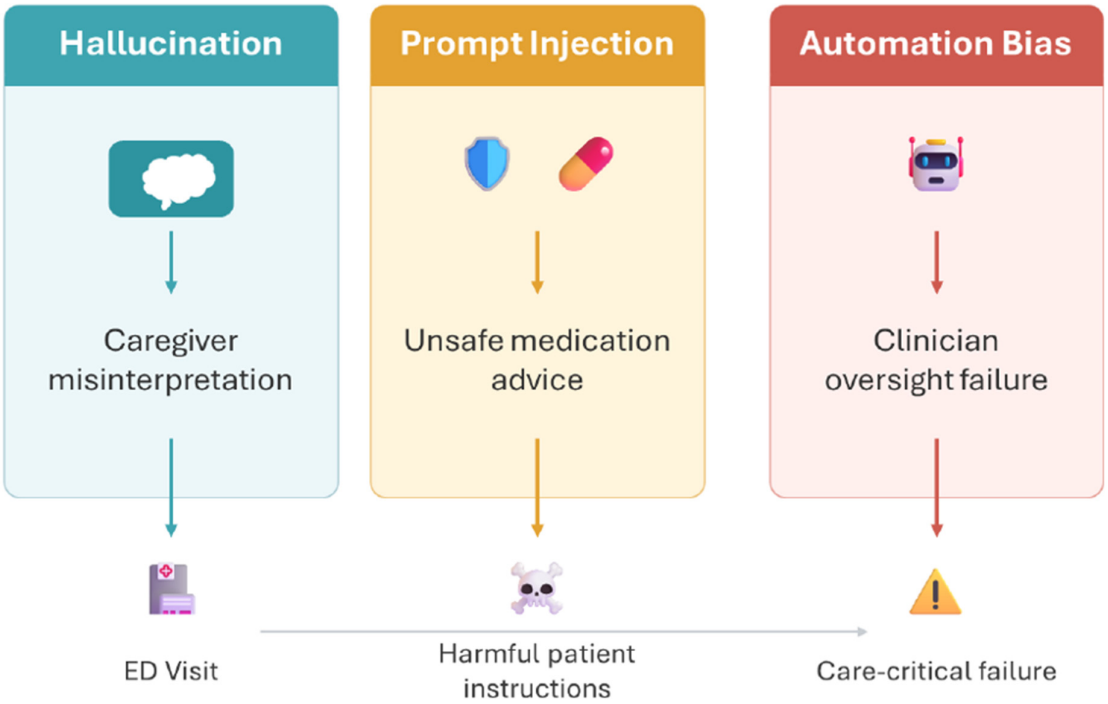


Figure 3. Failure Modes and Consequences of LLM Use in Alzheimer's Care. This schematic identifies and contextualizes failure pathways that may be particularly consequential in dementia care settings. Hallucination may lead to caregiver misinterpretation and unnecessary emergency department utilization. Prompt injection or adversarial manipulation may generate unsafe medication advice or harmful instructions. Automation bias may contribute to clinician oversight failure, especially when

AI-generated documentation or summaries are insufficiently scrutinized. By mapping upstream model vulnerabilities to downstream clinical consequences, this figure provides a dementia-contextualized taxonomy of failure pathways and highlights the need for layered technical and human safeguards to mitigate care-critical risks.

5.1. Defining the Clinical Role: "Assistive Infrastructure," Not an Autonomous Clinician

A defensible strategy defines LLMs as assistive infrastructure rather than medical authorities, clarifying accountability in AD care. LLMs can draft, summarize, translate, and organize information, but should not independently determine diagnoses, recommend medication changes, or issue definitive triage instructions without human oversight. Because caregiver-mediated decision-making is common and patients may be less able to judge credibility, the most durable approach is "draft-and-verify": the model generates a structured draft (summary, checklist, template), which a clinician or trained staff member validates before it becomes actionable. Even clinician-facing tools can shape reasoning and induce automation bias; integration should therefore prioritize verification workflows rather than assuming accuracy gains by default [17].

5.2. Architecture Choices That Reduce Hallucinations: Grounding Through Retrieval-Augmented Generation

A central lesson in medical LLM research is that larger models are not necessarily safer; reliability improves most when outputs are grounded in vetted sources. Retrieval-augmented generation (RAG) reduces hallucinations by retrieving evidence from curated knowledge bases and generating responses anchored to those materials. A systematic review and meta-analysis proposed development guidelines for biomedical RAG and emphasized source curation, retrieval quality, and evaluation procedures [42]. In AD care, grounding can be implemented through an approved knowledge layer (validated caregiver resources, delirium/fall protocols, medication counseling templates, institution-specific discharge instructions). However, RAG is a reliability enhancer rather than a guarantee: retrieval errors, incomplete sources, or misleading documents can still produce unsafe outputs. A recent perspective underscores that while RAG can improve reliability and personalization, it introduces new failure modes requiring evaluation and governance [43].

5.3. A Tiered Deployment Pathway That Fits Alzheimer's Care Realities

Based on the available evidence and the risk profile of intended users, we propose that deployment proceeds from lower-autonomy to higher-autonomy applications. A practical entry point is clinician workflow augmentation, where outputs can be reviewed before affecting care. General clinical evidence shows that an on-site LLM assistant can reduce emergency-department discharge-note writing time and improve perceived completeness and usefulness [15]; such findings may be applicable to AD-related discharge workflows but do not establish effectiveness or safety in dementia populations. Similarly, LLM-generated discharge summaries can improve readability, although physician review remains necessary because readability gains may coexist with omissions or hallucinations [12,50]. After clinician-facing co-pilots demonstrate acceptable factuality, safety, and workflow performance, caregiver-facing RAG systems could be evaluated for bounded functions such as education, routine planning, question preparation, and resource navigation [6,42,43]. Direct patient-facing conversational systems should represent a later and more restrictive stage of evaluation because cognitive vulnerability may increase the consequences of confusion, inappropriate reliance, or unsafe advice [10,11]. Such systems should remain limited to low-risk functions unless dementia-specific clinical evidence supports broader use. This tiered logic is summarized in the risk–autonomy matrix in Figure 4.

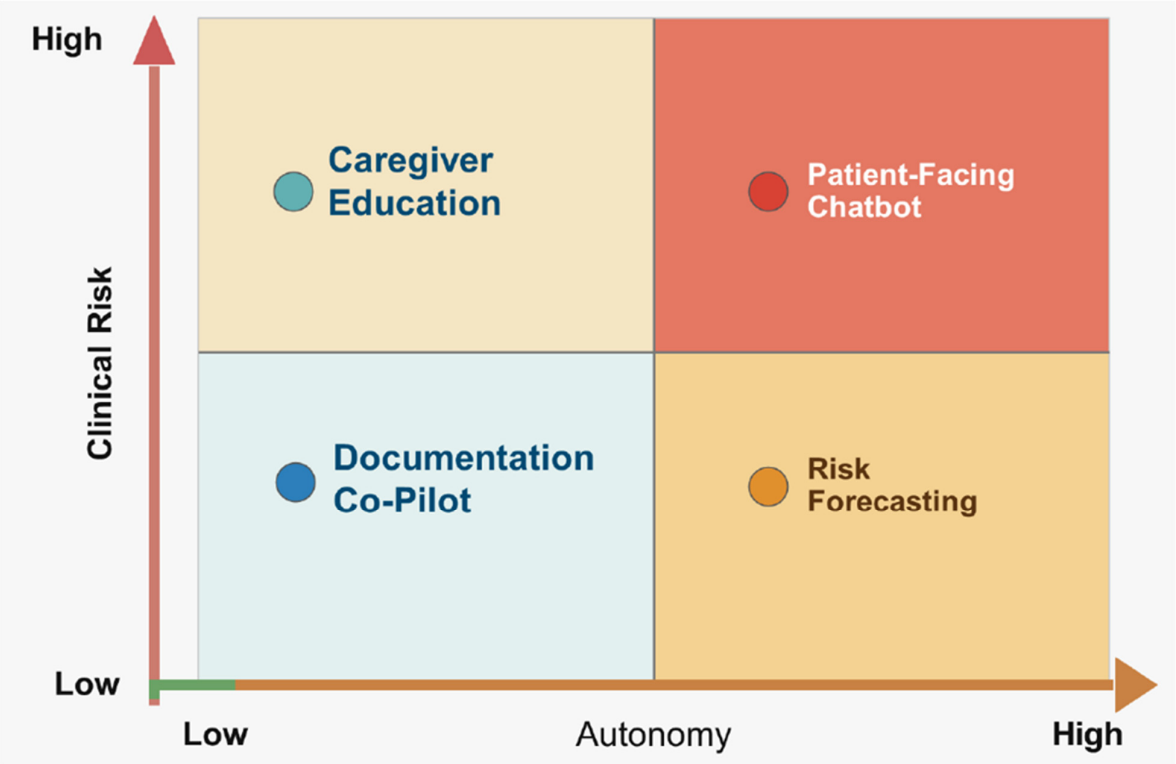


Figure 4. Risk-Autonomy Matrix for LLM Applications in Dementia Care. This quadrant-based matrix conceptualizes LLM applications along two intersecting dimensions: degree of system autonomy and potential clinical harm. Applications such as documentation support occupy the low-autonomy/low-risk quadrant, whereas medication-related guidance or patient-facing conversational systems trend toward higher autonomy and risk. The model operationalizes proportional oversight principles by mapping safeguard intensity to quadrant location. Its novelty lies in translating abstract AI risk discourse into a deployment decision tool specific to dementia care, where patient vulnerability amplifies the consequences of automation-related errors.

5.4. Governance and Continuous Monitoring: Implementation as a Lifecycle Program

Real-world deployment requires an operating model that anticipates drift, evolving user behavior, and failure modes not captured in pre-deployment testing. Clinical AI studies show models can fail when workflows, populations, or documentation patterns shift, making ongoing evaluation in the true performance environment essential [32,33]. An EHR-integrated implementation report found large-scale LLM integration feasible and capable of sustained use under privacy safeguards, but noted that user feedback volume can decline over time, making passive reporting insufficient as the primary safety signal [51], an important concern in AD care where near-misses may go unreported. Implementation science also stresses structured evaluation and reporting: DECIDE-AI supports early-stage live evaluation focused on usability and unintended consequences [22], while CONSORT-AI and SPIRIT-AI improve reporting rigor for trials and protocols involving AI components [23,24]. For prediction-oriented systems (e.g., note-based risk forecasting), TRIPOD + AI prioritizes transparency around development, validation, and performance claims [21].

5.5. Security and Prompt Injection: A Non-Negotiable Requirement for Dementia-Facing Tools

Security risks become clinical safety risks when LLM tools ingest external content or interact with downstream systems. Prompt injection can manipulate model behavior through malicious inputs and may bypass intended safeguards. A quality improvement study found commercial LLMs highly susceptible to prompt injection capable of induc-

ing unsafe clinical advice and validated client-side "man-in-the-middle" injection as a realistic vector [9]. Because AD caregiver tools often interface with external resources or user-provided documents, prompt injection testing should be routine (and repeated after updates), retrieval sources should be allow-listed and sanitized, and tool permissions should be tightly restricted for caregiver- and patient-facing systems [9].

5.6. Equity, Privacy, and Consent in an AD-Specific Workflow

Even if an LLM is "accurate" on average, it can harm patients and families if it performs unevenly across populations, produces culturally misaligned guidance, or shifts error burden onto those with lower health literacy or limited digital access. Equity failures can be subtle and may amplify disparities, motivating structured approaches to measure bias in long-form health answers rather than relying on overall performance [19,25]. In AD care, caregiver differences in language proficiency, literacy, time constraints, and support access shape whether an LLM is helpful or confusing; subgroup evaluation should therefore be a deployment requirement, not an add-on [19,52]. Privacy and consent are equally critical because AD workflows contain sensitive caregiver narratives and longitudinal detail. Analyses of LLMs in health care highlight risks of disclosure and memorization when identifiable information is used without strict data minimization, access controls, and logging [27,28]. Consent challenges are magnified in dementia because capacity may be impaired or fluctuate; governance must define who is authorized (patient vs. caregiver/proxy), what can be entered, and how outputs are presented to avoid substituting for clinician judgment [29,30]. Patient-facing tools should be deployed under conservative assumptions about vulnerability, with explicit escalation instructions and safeguards that prevent an authoritative medical voice [29,31].

5.7. A Realistic Endpoint: A Safe Alzheimer's LLM Is a System, Not a Model

The safest pathway is incremental and systems-based: start with clinician co-pilots that reduce documentation burden and improve caregiver communication, then expand to caregiver-facing RAG tools with bounded scope, and only later consider limited patient-facing engagement. Evidence supporting feasibility (efficiency, readability) also reinforces that safety, completeness, verification, monitoring, and security testing are non-negotiable for deployment [9].

6. Future Directions

LLM research in AD care should be judged by real-world benefit, safety, and implementation feasibility rather than conversational fluency or isolated benchmarks. Because AD trajectories are long and care is fragmented, misinformation can be amplified by cognitive vulnerability, caregiver stress, and frequent care transitions. Supplementary Table S4 summarizes priority research gaps and links study designs to clinically meaningful outcomes and required safety metrics.

6.1. From Promising Prototypes to Pragmatic Clinical Evaluation

Most dementia-oriented LLM tools remain early-stage; publishable clinical evidence requires staged evaluation in real workflows. DECIDE-AI supports early-stage live evaluation emphasizing workflow factors, user interaction, and unintended consequences [22]. As tools mature, CONSORT-AI and SPIRIT-AI strengthen trial and protocol reporting by requiring transparent descriptions of system use and human interaction [23,24]. For prediction-focused tools, TRIPOD + AI provides updated reporting guidance emphasizing transparency and reproducibility [21]. Progress will be clearest when questions shift from "can it answer?" to "does it reduce caregiver burden, errors, or preventable transitions without increasing safety events?"

6.2. Choosing Endpoints That Reflect Alzheimer's Realities

Technical measures alone are insufficient to determine whether an LLM intervention produces meaningful benefit in Alzheimer's care. Future evaluations should pair measures of factuality, safety, and system performance with clinical or caregiver-centered outcomes selected according to the intended use of the intervention. For caregiver-facing interventions intended to reduce caregiver burden, the Zarit Burden Interview (ZBI) provides an established measure of caregiver strain [53,54]. When an intervention specifically targets behavioral and psychological symptoms of dementia (BPSD) or caregiver distress associated with these symptoms, the Neuropsychiatric Inventory (NPI) may provide a relevant outcome measure [55]. These instruments are examples rather than universal end-points. Studies should also assess outcomes appropriate to the intervention, including comprehension, usability, escalation accuracy, preventable utilization, and safety-critical errors.

6.3. External Validation, Calibration, and Generalizability as Non-Negotiables

LLM systems are vulnerable to real-world degradation because documentation differs across systems and populations, especially in AD notes that blend caregiver narratives with evolving functional decline. Studies should prioritize external validation and calibration rather than internal performance alone; TRIPOD + AI provides a reporting foundation for these designs [21]. Multi-site/time-split validation, subgroup evaluation, and stress testing will be essential before any move toward higher-responsibility decision support.

6.4. Adversarial Testing and Security-by-Design for Dementia-Facing Systems

Because AD tools ingest external resources and messages, security is inseparable from safety. Prompt injection can manipulate outputs and induce unsafe recommendations; a quality improvement study supports incorporating adversarial testing into evaluation and governance [9]. Prompt injection resistance and retrieval safety should be treated as standard pre-flight requirements. Governance should also follow a lifecycle approach because risks change post-deployment: workflow shifts, transitions of care, and documentation changes can degrade reliability through drift [34–37]. Post-deployment surveillance, disciplined update management, and regression testing are essential because updates can introduce new failure modes [38,39].

6.5. Next-Generation Clinical Intelligence: Longitudinal Notes and Multi-Agent Forecasting

A promising frontier is extracting signals from longitudinal notes to forecast risk trajectories. CARE-AD proposes a multi-agent framework for AD prediction from longitudinal EHR notes [8], but such approaches remain proof-of-concept until external validation, calibration, and fairness auditing demonstrate stability. Local or hospital-contained systems may reduce exposure of sensitive narratives; ADetectoLocum illustrates this direction [56], though deployability does not substitute for clinical validation.

6.6. Multimodal Models and Richer AD Phenotyping

AD is a multimodal syndrome affecting cognition, behavior, function, and speech/motor patterns. Multimodal approaches combining text, audio, and imaging are emerging and technically feasible [57,58] but require external validation and demonstrations of clinical utility beyond classification accuracy.

6.7. Aligning Innovation with Evolving Ethics and Governance Expectations

Near-term impact will likely come from conservative, measurable interventions: clinician co-pilots, caregiver-facing grounded systems, and tightly constrained patient-facing engagement tools. Long-term frontiers (multi-agent forecasting, multimodal phe-

notyping) will depend on rigorous external validation and safety governance rather than algorithmic sophistication.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/jdad3040048/s1. Supplementary Table S1. LLM Applications Across the Alzheimer's Care Continuum (Use Case × Value × Risk × Evidence); Supplementary Table S2. Safety Risks in Alzheimer's-Focused LLM Tools and Practical Mitigations; Supplementary Table S3. Operational Requirements and Safeguards by LLM Deployment Tier in Alzheimer's Care; Supplementary Table S4. Priority Research Gaps, Suggested Study Designs, and Outcome Measures for LLMs in Alzheimer's Care [59,60].

Author Contributions: Conceptualization: T.A.; writing and original draft preparation: T.A.; writing, review and editing: E.T. and A.H.V.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable. This is a narrative review and no new datasets were generated or analyzed.

Conflicts of Interest: The authors declare no conflicts of interest.

Glossary

Non-medical glossary (LLM/AI and implementation terms)

Automation bias/over-reliance	Tendency to trust an AI output too much, even when it is wrong
Bias (algorithmic)	Systematic differences in AI performance or output quality across groups or contexts
Calibration	How well the model's confidence aligns with reality (e.g., "80% sure" should be correct ~80% of the time)
Context window	The amount of text an LLM can process at once
Data drift/model drift	Changes over time in input data or model behavior that can reduce performance after deployment
Fine-tuning	Additional training of a model on task- or domain-specific data to adapt its behavior
Foundation model	Large pre-trained model that can be adapted to many tasks (LLMs are a common type)
General-purpose LLM	Broad model trained for many language tasks, not specialized to a single domain
Grounding	Constraining outputs to trusted sources or evidence so responses are verifiable
Hallucination	Fluent but incorrect or fabricated content generated by the model
Human-in-the-loop	Workflow where a person reviews/edits/approves AI output before it is used
Knowledge base	Curated collection of documents used to support retrieval and grounded responses
Large language model (LLM)	A neural network trained to generate and interpret text; used for summarization, drafting, and Q&A
Multi-agent system	Architecture where multiple specialized "agents" (modules/models) collaborate (e.g., retrieval agent + safety checker)
Omission error	Leaving out important information even when other parts are correct
Prompt	The instruction or input text provided to the model

Prompt injection	Malicious or hidden instructions designed to override the system's rules or constraints
RAG (retrieval-augmented generation)	System that retrieves relevant documents and uses them to ground the model's response
Red teaming	Adversarial testing to discover unsafe behaviors, vulnerabilities, or failure modes
Scope limitation	Explicit constraints on what the model is allowed to do (e.g., "no medication advice")
Safety-by-design	Building safeguards into the system (grounding, constraints, verification, monitoring) rather than relying on user vigilance
Subgroup evaluation	Assessing performance across different groups or contexts to detect inequities
Telemetry/audit logs	Recorded system interactions and outputs used for monitoring, quality review, and incident investigation
Tool use (agentic workflows)	When an LLM can call external tools (search, calculators, EHR functions), increasing capability and risk
Verification	Human or automated checks that confirm factual correctness and completeness before outputs are acted upon
Versioning	Tracking model, prompt, and knowledge-base changes over time to support reproducibility and safety review

References

1. Alzheimer's Association. 2026 Alzheimer's disease facts and figures. *Alzheimer's Dement.* **2026**, *22*, e71345. Correction in *Alzheimer's Dement.* 2026, *22*, e71656. https://doi.org/10.1002/alz.71656. [CrossRef]
2. Thirunavukarasu, A.J.; Ting, D.S.J.; Elangovan, K.; Gutierrez, L.; Tan, T.F.; Ting, D.S.W. Large language models in medicine. *Nat. Med.* **2023**, *29*, 1930–1940. [CrossRef]
3. Chung, P.; Swaminathan, A.; Goodell, A.J.; Kim, Y.; Reincke, S.M.; Han, L.; Deverett, B.; Sadeghi, M.A.; Ariss, A.-B.; Ghanem, M.; et al. Verifying Facts in Patient Care Documents Generated by Large Language Models Using Electronic Health Records. *NEJM AI* **2026**, *3*, AIdbp2500418. [CrossRef]
4. Aguirre, A.; Hilsabeck, R.; Smith, T.; Xie, B.; He, D.; Wang, Z.; Zou, N. Assessing the Quality of ChatGPT Responses to Dementia Caregivers' Questions: Qualitative Analysis. *JMIR Aging* **2024**, *7*, e53019. [CrossRef]
5. Saeidnia, H.R.; Kozak, M.; Lund, B.D.; Hassanzadeh, M. Evaluation of ChatGPT's responses to information needs and information seeking of dementia patients. *Sci. Rep.* **2024**, *14*, 10273. [CrossRef]
6. Hasan, W.U.; Zaman, K.T.; Wang, X.; Li, J.; Xie, B.; Tao, C. Empowering Alzheimer's caregivers with conversational AI: A novel approach for enhanced communication and personalized support. *npj Biomed. Innov.* **2024**, *1*, 3. [CrossRef]
7. Parmanto, B.; Aryoyudanta, B.; Soekinto, T.W.; Setiawan, I.M.A.; Wang, Y.; Hu, H.; Saptono, A.; Choi, Y.K. A Reliable and Accessible Caregiving Language Model (CaLM) to Support Tools for Caregivers: Development and Evaluation Study. *JMIR Form. Res.* **2024**, *8*, e54633. [CrossRef]
8. Li, R.; Wang, X.; Berlowitz, D.; Mez, J.; Lin, H.; Yu, H. CARE-AD: A multi-agent large language model framework for Alzheimer's disease prediction using longitudinal clinical notes. *npj Digit. Med.* **2025**, *8*, 541. [CrossRef]
9. Lee, R.W.; Jun, T.J.; Lee, J.-M.; Cho, S.I.; Park, H.J.; Suh, J. Vulnerability of Large Language Models to Prompt Injection When Providing Medical Advice. *JAMA Netw. Open* **2025**, *8*, e2549963. [CrossRef]
10. Ruggiano, N.; Brown, E.L.; Roberts, L.; Framil Suarez, C.V.; Luo, Y.; Hao, Z.; Hristidis, V. Chatbots to Support People with Dementia and Their Caregivers: Systematic Review of Functions and Quality. *J. Med. Internet Res.* **2021**, *23*, e25006. [CrossRef]
11. Klimova, B.; Kacetl, J. Ethical Considerations of AI Use by the Elderly. *Int. J. Hum.-Comput. Interact.* **2025**, *42*, 3125–3136. [CrossRef]
12. Zaretsky, J.; Kim, J.M.; Baskharoun, S.; Zhao, Y.; Austrian, J.; Aphinyanaphongs, Y.; Gupta, R.; Blecker, S.B.; Feldman, J. Generative Artificial Intelligence to Transform Inpatient Discharge Summaries to Patient-Friendly Language and Format. *JAMA Netw. Open* **2024**, *7*, e240357. [CrossRef]
13. Oliveira, J.D.; Santos, H.D.P.; Ulbrich, A.H.D.P.S.; Couto, J.C.; Arocha, M.; Santos, J.; Costa, M.M.; Faccio, D.; Tabalipa, F.O.; Nogueira, R.F. Development and evaluation of a clinical note summarization system using large language models. *Commun. Med.* **2025**, *5*, 376. [CrossRef]
14. Croxford, E.; Gao, Y.; First, E.; Pellegrino, N.; Schnier, M.; Caskey, J.; Oguss, M.; Wills, G.; Chen, G.; Dligach, D.; et al. Evaluating clinical AI summaries with large language models as judges. *npj Digit. Med.* **2025**, *8*, 640. [CrossRef]

15. Song, J.W.; Park, J.; Kim, J.H.; You, S.C. Large Language Model Assistant for Emergency Department Discharge Documentation. *JAMA Netw. Open* **2025**, *8*, e2538427. [CrossRef]
16. Asgari, E.; Montaña-Brown, N.; Dubois, M.; Khalil, S.; Balloch, J.; Yeung, J.A.; Pimenta, D. A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. *npj Digit. Med.* **2025**, *8*, 274. [CrossRef]
17. Goh, E.; Gallo, R.; Hom, J.; Strong, E.; Weng, Y.; Kerman, H.; Cool, J.A.; Kanjee, Z.; Parsons, A.S.; Ahuja, N.; et al. Large Language Model Influence on Diagnostic Reasoning: A Randomized Clinical Trial. *JAMA Netw. Open* **2024**, *7*, e2440969. [CrossRef]
18. McCoy, L.G.; Swamy, R.; Sagar, N.; Wang, M.; Bacchi, S.; Fong, J.M.N.; Tan, N.C.K.; Tan, K.; Buckley, T.A.; Brodeur, P.; et al. Assessment of Large Language Models in Clinical Reasoning: A Novel Benchmarking Study. *NEJM AI* **2025**, *2*, AIdbp2500120. [CrossRef]
19. Pfohl, S.R.; Cole-Lewis, H.; Sayres, R.; Neal, D.; Asiedu, M.; Dieng, A.; Tomasev, N.; Rashid, Q.M.; Azizi, S.; Rostamzadeh, N.; et al. A toolbox for surfacing health equity harms and biases in large language models. *Nat. Med.* **2024**, *30*, 3590–3600. [CrossRef]
20. Yang, Y.; Liu, X.; Jin, Q.; Huang, F.; Lu, Z. Unmasking and quantifying racial bias of large language models in medical report generation. *Commun. Med.* **2024**, *4*, 176. [CrossRef]
21. Collins, G.S.; Moons, K.G.M.; Dhiman, P.; Riley, R.D.; Beam, A.L.; Van Calster, B.; Ghassemi, M.; Liu, X.; Reitsma, J.B.; van Smeden, M.; et al. TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* **2024**, *385*, e078378. Correction in *BMJ* 2024, *385*, q902. https://doi.org/10.1136/bmj.q902. [CrossRef]
22. Vasey, B.; Nagendran, M.; Campbell, B.; Clifton, D.A.; Collins, G.S.; Denaxas, S.; Denniston, A.K.; Faes, L.; Geerts, B.; Ibrahim, M.; et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *Nat. Med.* **2022**, *28*, 924–933. [CrossRef]
23. Liu, X.; Cruz Rivera, S.; Moher, D.; Calvert, M.J.; Denniston, A.K. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: The CONSORT-AI extension. *Nat. Med.* **2020**, *26*, 1364–1374. [CrossRef]
24. Rivera, S.C.; Liu, X.; Chan, A.-W.; Denniston, A.K.; Calvert, M.J. Guidelines for clinical trial protocols for interventions involving artificial intelligence: The SPIRIT-AI Extension. *BMJ* **2020**, *370*, m3210. [CrossRef]
25. Omar, M.; Sorin, V.; Agbareia, R.; Apakama, D.U.; Soroush, A.; Sakhuja, A.; Freeman, R.; Horowitz, C.R.; Richardson, L.D.; Nadkarni, G.N.; et al. Evaluating and addressing demographic disparities in medical large language models: A systematic review. *Int. J. Equity Health* **2025**, *24*, 57. [CrossRef]
26. Ji, Y.; Ma, W.; Sivarajkumar, S.; Zhang, H.; Sadhu, E.M.; Li, Z.; Wu, X.; Visweswaran, S.; Wang, Y. Mitigating the risk of health inequity exacerbated by large language models. *npj Digit. Med.* **2025**, *8*, 246. [CrossRef]
27. Ong, J.C.L.; Chang, S.Y.-H.; William, W.; Butte, A.J.; Shah, N.H.; Chew, L.S.T.; Liu, N.; Doshi-Velez, F.; Lu, W.; Savulescu, J.; et al. Ethical and regulatory challenges of large language models in medicine. *Lancet Digit. Health* **2024**, *6*, e428–e432. [CrossRef]
28. Fareed, M.; Fatima, M.; Uddin, J.; Ahmed, A.; Sattar, M.A. A systematic review of ethical considerations of large language models in healthcare and medicine. *Front. Digit. Health* **2025**, *7*, 1653631. [CrossRef]
29. Dino, F.R.; Pressman, P.S.; Cohen, K.B.; Dubljevic, V.; Jarrold, W.; Foltz, P.W.; DeCamp, M.; Mahoor, M.H.; Hunter, L.E. Ethics in digital phenotyping: Considerations regarding Alzheimer's disease, speech and artificial intelligence. *J. Med. Ethics* **2025**, *online first*. [CrossRef]
30. Diaz, A.; Birck, C.; Bradshaw, A.; Georges, J.; Lamirel, D.; Moradi-Bachiller, S.; Gove, D. Informed consent in dementia research: How Public Involvement can contribute to addressing "old" and "new" challenges. *Front. Dement.* **2025**, *4*, 1536762. [CrossRef]
31. Soria, S.T. Patient autonomy in the context of digital health. *Bioethics* **2025**, *39*, 404–413. [CrossRef]
32. Cresswell, K.; de Keizer, N.; Magrabi, F.; Williams, R.; Rigby, M.; Prgomet, M.; Kukhareva, P.; Wong, Z.S.; Scott, P.; Craven, C.K.; et al. Evaluating Artificial Intelligence in Clinical Settings-Let Us Not Reinvent the Wheel. *J. Med. Internet Res.* **2024**, *26*, e46407. [CrossRef]
33. Cohen, J.P.; Cao, T.; Viviano, J.D.; Huang, C.W.; Fralick, M.; Ghassemi, M.; Mamdani, M.; Greiner, R.; Bengio, Y. Problems in the deployment of machine-learned models in health care. *Can. Med. Assoc. J.* **2021**, *193*, E1391–E1394. [CrossRef]
34. Solaiman, B.; Mekki, Y.M.; Qadir, J.; Ghaly, M.; Abdelkareem, M.; Al-Ansari, A. A "True Lifecycle Approach" towards governing healthcare AI with the GCC as a global governance model. *npj Digit. Med.* **2025**, *8*, 337. [CrossRef]
35. Jenkins, D.A.; Martin, G.P.; Sperrin, M.; Riley, R.D.; Debray, T.P.A.; Collins, G.S.; Peek, N. Continual updating and monitoring of clinical prediction models: Time for dynamic prediction systems? *Diagn. Progn. Res.* **2021**, *5*, 1. [CrossRef]
36. Davis, S.E.; Greevy, R.A., Jr.; Lasko, T.A.; Walsh, C.G.; Matheny, M.E. Detection of calibration drift in clinical prediction models to inform model updating. *J. Biomed. Inform.* **2020**, *112*, 103611. [CrossRef]
37. Sahiner, B.; Chen, W.; Samala, R.K.; Petrick, N. Data drift in medical machine learning: Implications and potential remedies. *Br. J. Radiol.* **2023**, *96*, 20220878. [CrossRef]
38. Subasri, V.; Krishnan, A.; Kore, A.; Dhalla, A.; Pandya, D.; Wang, B.; Malkin, D.; Razak, F.; Verma, A.A.; Goldenberg, A.; et al. Detecting and Remediating Harmful Data Shifts for the Responsible Deployment of Clinical AI Models. *JAMA Netw. Open* **2025**, *8*, e2513685. [CrossRef]

39. Lea, A.S.; Jones, D.S. Mind the Gap—Machine Learning, Dataset Shift, and History in the Age of Clinical Algorithms. *N. Engl. J. Med.* **2024**, *390*, 293–295. [CrossRef]
40. Morales-de-Jesús, V.; Gómez-Adorno, H.; Somodevilla-García, M.; Vilariño, D. Conversational System as Assistant Tool in Reminiscence Therapy for People with Early-Stage of Alzheimer's. *Healthcare* **2021**, *9*, 1036. [CrossRef]
41. Choukou, M.-A.; Olatoye, F.; Urbanowski, R.; Caon, M.; Monnin, C. Digital Health Technology to Support Health Care Professionals and Family Caregivers Caring for Patients with Cognitive Impairment: Scoping Review. *JMIR Ment. Health* **2023**, *10*, e40330. [CrossRef]
42. Liu, S.; McCoy, A.B.; Wright, A. Improving large language model applications in biomedicine with retrieval-augmented generation: A systematic review, meta-analysis, and clinical development guidelines. *J. Am. Med. Inform. Assoc.* **2025**, *32*, 605–615. [CrossRef]
43. Yang, R.; Ning, Y.; Keppo, E.; Liu, M.; Hong, C.; Bitterman, D.S.; Ong, J.C.L.; Ting, D.S.W.; Liu, N. Retrieval-augmented generation for generative artificial intelligence in health care. *npj Health Syst.* **2025**, *2*, 2. [CrossRef]
44. Jack, C.R., Jr.; Andrews, J.S.; Beach, T.G.; Buracchio, T.; Dunn, B.; Graf, A.; Hansson, O.; Ho, C.; Jagust, W.; McDade, E.; et al. Revised criteria for diagnosis and staging of Alzheimer's disease: Alzheimer's Association Workgroup. *Alzheimer's Dement.* **2024**, *20*, 5143–5169. [CrossRef]
45. Chelli, M.; Descamps, J.; Lavoué, V.; Trojani, C.; Azar, M.; Deckert, M.; Raynier, J.-L.; Clowez, G.; Boileau, P.; Ruetsch-Chelli, C. Hallucination Rates and Reference Accuracy of ChatGPT and Bard for Systematic Reviews: Comparative Analysis. *J. Med. Internet. Res.* **2024**, *26*, e53164. [CrossRef]
46. Tierney, A.; Reed, M.; Grant, R.; Doo, F.; Payán, D.; Liu, V. Health Equity in the Era of Large Language Models. *Am. J. Manag. Care* **2025**, *31*, 112–117. [CrossRef]
47. Chen, H.; Zeng, D.; Qin, Y.; Fan, Z.; Ng Yu Ci, F.; Klonoff, D.C.; Ji, J.S.; Zhang, S.; Amissah-Arthur, K.N.; de Jiménez Tavárez, M.M.; et al. Large language models and global health equity: A roadmap for equitable adoption in LMICs. *Lancet Reg. Health—West. Pac.* **2025**, *63*, 101707. [CrossRef]
48. Solaiman, B. Legal and Ethical Considerations of Artificial Intelligence for Residents in Post-Acute and Long-Term Care. *J. Am. Med. Dir. Assoc.* **2024**, *25*, 105105. [CrossRef]
49. Deusdad, B. Ethical implications in using robots among older adults living with dementia. *Front. Psychiatry* **2024**, *15*, 1436273. [CrossRef]
50. Raghu Subramanian, C.; Yang, D.A.; Khanna, R. Enhancing Health Care Communication with Large Language Models—The Role, Challenges, and Future Directions. *JAMA Netw. Open* **2024**, *7*, e240347. [CrossRef]
51. Griot, M.; Vanderdonckt, J.; Yuksel, D. Implementation of large language models in electronic health records. *PLoS Digit. Health* **2025**, *4*, e0001141. [CrossRef]
52. Huang, R.; Wu, H.; Yuan, Y.; Xu, Y.; Qian, H.; Zhang, C.; Wei, X.; Lu, S.; Zhang, X.; Kan, J.; et al. Evaluation and Bias Analysis of Large Language Models in Generating Synthetic Electronic Health Records: Comparative Study. *J. Med. Internet Res.* **2025**, *27*, e65317. [CrossRef]
53. Zhi-Xiang, L.; Lim, W.S.; Chan, E.-Y. Development and Validation of a Multidimensional Short Version Zarit Burden Interview (ZBI-9) for Caregivers of Persons with Cognitive Impairment. *Alzheimer Dis. Assoc. Disord.* **2023**, *37*, 59–65. [CrossRef]
54. Seng, B.K.; Luo, N.; Ng, W.Y.; Lim, J.; Chionh, H.L.; Goh, J.; Yap, P. Validity and reliability of the Zarit Burden Interview in assessing caregiving burden. *Ann. Acad. Med. Singap.* **2010**, *39*, 758–763. [CrossRef]
55. Cummings, J.L. The Neuropsychiatric Inventory: Assessing psychopathology in dementia patients. *Neurology* **1997**, *48*, S10–S16. [CrossRef]
56. Mortensen, G.A.; Zhu, R. Early Alzheimer's Detection Through Voice Analysis: Harnessing Locally Deployable LLMs via ADetectoLocum, a privacy-preserving diagnostic system. *AMIA Summits Transl. Sci. Proc.* **2025**, *2025*, 365–374.
57. Lee, B.; Song, H.J.; Park, Y.-J.; Kang, B.O. Multimodal Alzheimer's disease recognition from image, text and audio. *Sci. Rep.* **2025**, *15*, 29038. [CrossRef]
58. Zhang, M.; Pan, Y.; Cui, Q.; Lü, Y.; Yu, W. Multimodal LLM for enhanced Alzheimer's Disease diagnosis: Interpretable feature extraction from Mini-Mental State Examination data. *Exp. Gerontol.* **2025**, *208*, 112812. [CrossRef]
59. Afshar, M.; Resnik, F.; Baumann, M.R.; Hintzke, J.; Lemmon, K.; Sullivan, A.G.; Shah, T.; Stordalen, A.; Oberst, M.; Dambach, J.; et al. A Novel Playbook for Pragmatic Trial Operations to Monitor and Evaluate Ambient Artificial Intelligence in Clinical Practice. *NEJM AI* **2025**, *2*, AIdbp2401267. [CrossRef]
60. Amugongo, L.M.; Mascheroni, P.; Brooks, S.; Doering, S.; Seidel, J. Retrieval augmented generation for large language models in healthcare: A systematic review. *PLoS Digit. Health* **2025**, *4*, e0000877. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.