Open Access

Journal of Scientific Insights
E-ISSN: 3062-8571
DOI: https://doi.org/10.69930/jsi.v3i4.831
Review Article

ScienceTech. Group

Vol. 3 (4), 2026
Page: 478-492

Ageism in Human-AI Social Interaction: A Scoping Review of Age Bias in Large Language Model Chatbots for Older Adults

Soni Rudi Hartanto *, Suwarni, Ramadhani Ulansari, Yudhi Biantoro, Suharyanto, Taufikur Ramadhan

Informatics System Program Study, Faculty of Technology Informatics, Universitas Respati Indonesia, Indonesia
*Email (corresponding author): soni.rudi@fti.urindo.ac.id

Abstract. *The rapid deployment of Large Language Models (LLMs) across healthcare and conversational technology directly impacts the expanding global older population, a demographic that increasingly relies on these systems for cognitive framing and daily emotional interaction. However, tech governance frameworks largely neglect algorithmic ageism. While developers routinely update codebases to mitigate racial and gender biases, age discrimination remains unaddressed. This study quantifies the depth of this systemic oversight. Through a systematic evaluation of 12 core papers published between 2023 and 2026 in accordance with the PRISMA-ScR guidelines, we identified a critical dual-layered manifestation of AI ageism: explicit ageist values embedded in token-prediction engines and a condescending conversational style that defines benevolent ageism. Younger users generate the vast majority of training data for Reinforcement Learning from Human Feedback (RLHF) loops. Consequently, LLMs replicate and amplify these youth-centric perspectives. Current standard testing benchmarks fail to detect these interactional nuances. To address this methodological blind spot, this paper introduces a dual-stage debiasing framework that implements age-empathetic constraints during preprocessing, data cleaning, and real-time generation. Integrating social gerontology with digital ethics, this model provides a verifiable framework for inclusive conversational AI.*

Keywords: *Algorithmic Ageism; Large Language Models; Human-AI Interaction; Socio-Technical Paradox; Benevolent Ageism*

1. Introduction

Generative AI and Large Language Models (LLMs) have changed how humans interact with computers. Now, software speaks like us. Tech companies immediately saw an opportunity here, selling these systems as perfect, round-the-clock digital friends for older adults to prevent isolation. But reality contradicts the marketing. Conversational AI routinely generates deep-rooted age bias. Mainstream computer ethics research focuses almost entirely on solving racial and gender disparities, meaning systemic age discrimination gets ignored (Stypinska, 2023). This negligence creates a severe socio-technical contradiction. We build software to fix loneliness among older generations. Yet, the algorithms themselves rely on the exact stereotypes that alienate older adults from society.

Younger demographics dominate online spaces and generate the vast majority of internet text that developers include in modern LLM training datasets without verification. Because these unverified archives absorb old human prejudices, the internal values and linguistic styles of the resulting AI bots lean heavily toward youth perspectives (Chu et al.,

Received: 25 June 2026 / Accepted: 22 July 2026 / Available online: 25 July 2026

2023). The training material shapes the system's outlook. When older individuals interact with these platforms, the AI acts on hidden, patronizing assumptions that seniors are inherently frail, forgetful, or technologically illiterate. This systemic discrimination does not present itself openly. Instead, the text generation engine executes algorithmic ageism through subtle linguistic and cognitive biases deeply embedded within its core code (Zhang et al., 2025).

Contemporary conversational architectures completely lack the critical moderation filters needed to catch age-related discrepancies, rendering existing safety alignments blind to this demographic even as they react aggressively to explicit sexism or racism. Ageist microaggressions slip past these automated protocols completely unnoticed (Fernández-Ardèvol et al., 2026). The technology functions without necessary safeguards. Because developers ignore this operational oversight, LLM-based chatbots continuously mirror and amplify harmful societal stereotypes during real-time dialogues with older users. We must map how these subtle biases manifest in daily social interactions from the ground up. Deploying generative AI without this foundational understanding inevitably widens the digital divide, transforming these systems into tools that drive automated age discrimination and systemic exclusion.

The current computational literature channels almost all its methodological resources into tracking and addressing racial or gender biases in Large Language Models, leaving age discrimination largely underexplored despite growing scholarly agreement that generative systems harbor massive demographic biases (Stypinska, 2023). The field remains heavily uneven. Pioneering studies successfully quantified empirical data skew to prove that LLMs inherently lean toward youth-centric values (Siyang Liu et al., 2024). However, these investigations rely too heavily on static prompt engineering or automated benchmarks. This narrow testing creates a critical bottleneck. Existing frameworks completely isolate the model's linguistic output from the dynamic, longitudinal nature of actual human-AI social interactions.

Current methodology operates in a silo, creating three critical gaps. Engineering-driven metrics cause the first fundamental issue. Automated tools such as the Ageism Score (AS) can effectively identify explicit cognitive or affective biases (Shuo Liu et al., 2025). Human interaction, however, exposes their limits. Standard metrics fail to track the subtle, ageist microaggressions that develop organically during extended, fluid conversations with older individuals.

Current safety guardrails focus almost entirely on blocking overt slurs. Because of this narrow engineering lens, systems overlook implicit stereotypes. Dialogue exposes these subtle biases naturally. A chatbot can easily adopt a patronizing tone, project cognitive decline onto an older user, or resort to technological infantilization (Virginia Tech, 2025). The underlying software registers absolutely nothing wrong.

Isolating computational evaluations from actual deployment creates a separate problem. Researchers consistently ignore user-centered integration. To understand how ageism operates in the real world, studies must analyze the concrete lived experiences, technology-acceptance thresholds, and everyday "folk theories" that older adults develop as they interact with these conversational agents (Yu & Chattopadhyay, 2025). Measuring bias requires listening directly to the population experiencing it.

The literature is fractured. While sociologists dissect digital exclusion without offering actionable technical remedies, engineers deploy narrow algorithmic patches that collapse under the weight of real-world human dynamics. This polarization leaves software

developers without practical implementation guidelines and leaves older adults exposed to poorly conceptualized AI interactions. Synthesis is non-existent. Specifically, current research fails to track how pre-training biases transform into harmful psychological outcomes during active user engagement. Our study establishes this missing link. By applying a rigorous Population, Context, and Concept scoping methodology, we examine these compounding systemic gaps to force a convergence between computational alignment and gerontological safety frameworks.

This study establishes a structured evaluation framework designed to rectify existing empirical and methodological omissions. By mapping the direct intersections of generative language systems and gerontological safety, our scoping review synthesizes global scientific data to clarify how ageism operates behaviorally, linguistically, and structurally during active user engagement with Large Language Model chatbots. Bias is not static. Consequently, we discard the conventional perspective that limits algorithmic discrimination to data anomalies. Our framework redefines age-based bias as a dynamic, transactional force, meaning it functions as an active mechanism dictating whether current digital structures integrate or exclude older end-users.

This inquiry pursues four deeply integrated sub-objectives. First, we determine how AI dialogues perpetuate age stereotypes through specific linguistic mechanisms and typologies. Automated systems frequently project explicit assumptions of cognitive frailty, technological alienation, and physical decline onto older adults. We isolate these patterns. By evaluating the synthesized data directly, we map these specific conversational projections (Virginia Tech, 2025). The focus then shifts to the technical limits of evaluative metrics. We scrutinize contemporary computational measurement tools to determine whether newly engineered frameworks, such as the Ageism Score, can effectively capture subtle behavioral microaggressions. A glaring operational double standard exists. Automated safety filters aggressively mitigate sexism while consistently ignoring ageist content (Fernández-Ardèvol et al., 2026; Zhang et al., 2025).

Analyzing qualitative technology-acceptance thresholds and the subjective folk theories that older individuals formulate during active system deployment enables researchers to evaluate how effectively the current literature integrates their lived experiences (Yu & Chattopadhyay, 2025). This investigative focus forms our third objective, which connects raw computational metrics directly with user-centered empirical realities. Moving beyond mere evaluation to address structural design, our fourth objective synthesizes actionable socio-technical mitigation paradigms. We systematically review cutting-edge technical interventions, specifically analyzing dual-stage debiasing frameworks and empathy-driven reinforcement learning protocols. This detailed exploration isolates effective development pathways. Engineers can exploit these specific pathways to easily correct youth-centric architectural alignments, maintaining the core computational performance of the underlying models throughout the optimization process (INFORMS Research, 2025). Ultimately, interlocking these four distinct objectives yields a comprehensive roadmap for structuring future intergenerational digital governance (Stypinska, 2023).

This scoping review extends foundational implications across theoretical, methodological, and practical dimensions to inform the evolving landscape of socio-technical artificial intelligence governance. Establishing ageism as a critical, distinct vector of algorithmic discrimination enables us to challenge the youth-centric bias embedded in computational ethics frameworks. While modern literature maps out algorithmic prejudices

surrounding race and gender extensively, older demographics remain ignored. We re-center the conversation on these neglected groups. This pivot directly confronts the analytical blind spots identified in global digitization roadmaps (Stypinska, 2023). By analyzing how contemporary Large Language Models covertly encode and propagate ageist stereotypes, this study expands the scholastic understanding of machine alignment. Ultimately, shifting the conceptual focus away from mere semantic safety toward genuine intergenerational digital justice redefines how researchers evaluate ethical AI deployment.

Conceptualizing age bias within the dynamic, multi-turn boundaries of real social interactions provides a robust alternative to evaluating Large Language Model (LLM) behavior through isolated prompt engineering or detached benchmarking. This tactical departure from traditional methodological silos allows researchers to synthesize engineering-focused algorithmic audits with qualitative, user-centered gerontological research. We deliberately integrate these two disparate domains. Consequently, this comprehensive mapping contextualizes automated metrics, including the newly pioneered Ageism Score, within the broader spectrum of behavioral microaggressions and linguistic infantilization that older adults encounter in real-world deployments (Virginia Tech, 2025; Zhang et al., 2025). This framework establishes a flexible methodological blueprint. Future investigators can leverage this blueprint to conduct systematic, multidisciplinary audits of generative models without relying on restrictive binary classification paradigms.

Offering actionable pathways for both technology developers and regulatory bodies allows this study to address critical, practical, and socio-political imperatives. We instruct computer scientists and AI practitioners in optimized system architecture, highlighting effective development strategies ranging from dual-stage debiasing algorithms to empathy-driven reinforcement learning protocols. Implementing these specific tools corrects structural age skews. Crucially, engineering teams execute these corrections without compromising baseline model performance (INFORMS Research, 2025). Moving beyond development paradigms to examine broader structural accountability, this review exposes the distinct operational double standards embedded in current safety alignments, in which ageist infractions routinely bypass moderation checks (Fernández-Ardèvol et al., 2026). This exposed deficiency provides solid empirical justification for policymakers to formulate age-inclusive digital guardrails. We intend these insights to ensure that conversational agents deployed to alleviate loneliness among older populations function as instruments of genuine empowerment rather than automated vehicles of systemic exclusion. This scoping review establishes a socio-technical framework to address algorithmic ageism in generative AI, thereby supporting Sustainable Development Goals (SDGs) 3 of ensuring healthy lives and promoting well-being for all at all ages.

2. Methods

Adhering strictly to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) guidelines guarantees methodological rigor, transparency, and reproducibility in this study (Tricco et al., 2018). We deliberately selected the PRISMA-ScR framework rather than standard PRISMA protocols designed for clinical interventions to map the broad, heterogeneous, and rapidly evolving socio-technical landscape of algorithmic ageism in conversational AI. Structuring explicit eligibility criteria around the core boundaries of the Population, Concept, and Context (PCC) framework allowed us to capture the structural intersection of aging and generative language

technologies. We established specific target parameters. Eligible literature must investigate conversational interactions mediated by Large Language Model (LLM) architectures, including commercial applications such as ChatGPT, Gemini, and Copilot, as well as open-source foundations like Llama. Regarding the target population, this review focuses on studies of older adults (aged 60 years or older) or on research evaluating digital representations and algorithmic personas of aging cohorts.

To map the developmental trajectory of mainstream generative transformers, this review included full-text peer-reviewed empirical studies, formal conference proceedings indexed by ACM or ACL frameworks, and high-impact technical preprints published in English or Indonesian between January 2020 and May 2026. Eligible papers must explicitly examine manifestations of age bias, systemic ageist stereotypes, or linguistic discrepancies, such as infantilizing speech modifications, in human-AI communication. We imposed no geographical restrictions. For structural precision, we applied strict exclusion criteria to eliminate literature outside our operational boundaries. We systematically omitted studies evaluating older adults interacting with non-generative digital platforms, traditional search engines, or physical robotic agents that lacked an underlying large language processing architecture. This study systematically excludes doctoral dissertations, book reviews, short op-eds, general commentaries, and editorial notes. We also omitted investigations that generalized demographic biases without isolating age-related variables, as well as research that focused exclusively on racial and gender bias without specific data partitions for age-centric metrics. Papers focusing solely on medical diagnostic AI without a core social interaction component, or those lacking accessible full-text documents, were excluded from final analytical consideration.

Executing a systematic, transparent search workflow across multiple digital repositories ensured comprehensive identification of relevant academic literature. To capture both computational frameworks and social science perspectives, this study strategically leveraged three primary indexing platforms: PubMed/MEDLINE for biomedical and gerontological data; arXiv for cutting-edge computer science preprints; and Google Scholar to track multidisciplinary citations. Institutional access limitations to subscription-based indexes prompted this reliance on high-impact public repositories. We formally concluded all search operations in May 2026. This precise cutoff ensures that the compiled data reflects the latest advancements in generative transformer technologies. Deriving specific keywords directly from the core components of the Population, Concept, and Context framework allowed us to formulate rigorous search strings using a classic building-block approach. To isolate studies at the intersection of aging populations and artificial intelligence bias, we paired the population descriptors (older adults, elderly cohorts, and aged populations) with conceptual variables (ageism, age bias, age discrimination, and stereotyping). We integrated these elements. Finally, boolean operators systematically bind these groups to contextual terms such as large language models (LLMs), LLM architectures, generative systems, and conversational chatbots to ensure maximum retrieval sensitivity within each database.

Tailoring the exact syntax to each platform's specific indexing requirements maintained strict reproducibility during data collection. In PubMed, the query used Medical Subject Headings (MeSH) terms alongside text words to retrieve formal institutional records, whereas the arXiv repository restricted the search to the abstract field to filter out papers in which the target keywords appeared only as peripheral references. For Google Scholar, a precise, title-focused command minimized noise, isolating papers in which ageism and large

language models were the primary areas of investigation. We targeted three core conceptual domains: Large Language Models and Generative AI; Ageism and Age Discrimination; and Human-AI Social Interaction, with a focus on older demographics. Additionally, scanning the reference lists of retrieved foundational papers manually prevented the omission of critical empirical studies during the initial query phase. Developing a standardized data charting form a priori ensured consistency and analytical depth across the subsequent extraction dynamic. We piloted this form. Specifically, testing the tool against a small sample of eligible papers refined the analytical categories before the full extraction phase commenced. The primary investigator performed the full data extraction, and the co-authors independently reviewed the completed charting matrix to resolve any discrepancies by consensus.

Operationalizing the charting framework allowed us to capture specific bibliometric, methodological, and contextual dimensions of the retrieved literature. This form recorded basic descriptive fields such as author names, publication year, geographical origin, and academic venue. Classifying each study design into distinct categories—such as qualitative interviews, quantitative technical audits, or experimental content analysis—allowed tracking of relevant methodological attributes. We differentiated model categories. In the computational context, the extraction dynamic captured the specific large language model architectures evaluated, distinguishing commercial applications from open-source foundation models. To establish the empirical foundation for the subsequent thematic synthesis, the analytical core of this charting process focused heavily on the structural and behavioral manifestations of age bias in human-AI communication. The form systematically documented specific evaluation metrics used by researchers, including the newly developed Ageism Score, alongside operational double standards in system moderation filters. Finally, the matrix isolated proposed technical mitigation frameworks, specifically detailing dual-stage debiasing algorithms and empathy-driven reinforcement learning protocols.

3. Results and Discussion

3.1. Results

Our initial search across the selected digital repositories yielded 95 records. After automated screening removed duplicates and irrelevant entries, we retained 47 unique citations for title and abstract analysis. Next, two independent reviewers checked these records against our inclusion criteria. We focused specifically on conversational AI, system architectures, and age-related communication barriers. Ultimately, this stage knocked out 21 records. The reason? They focused solely on general algorithmic fairness, omitting specific metrics and qualitative insights for older populations.

We then scrutinized the remaining 26 papers through a rigorous full-text eligibility assessment. This step was crucial to lock in the final mapping. In doing so, we dropped 14 publications that did not meet our strict methodological criteria. Specifically, eight papers were cut from the cutting room floor because they lacked primary empirical data or technical audits of LLMs. Another six fell short because their scope drifted into physical robotic assistive hardware, rather than focusing on conversational AI chatbot frameworks. Whenever we hit a roadblock or disagreed during this screening, we sat down and threshed out the differences through constructive academic debates until we reached a solid consensus.

Our structured screening process ultimately yielded 12 eligible studies. These papers hit every single inclusion benchmark we set. Ultimately, they form the bedrock of this scoping review. By blending quantitative technical audits, content analysis, and qualitative

sociological interviews, this final cohort offers a diverse yet beautifully complementary methodological mix. We also ensured the selected literature represents a wide geographical spread across key contemporary publication years. This gives us a timely, rock-solid mapping of age bias within human-AI social interaction paradigms.

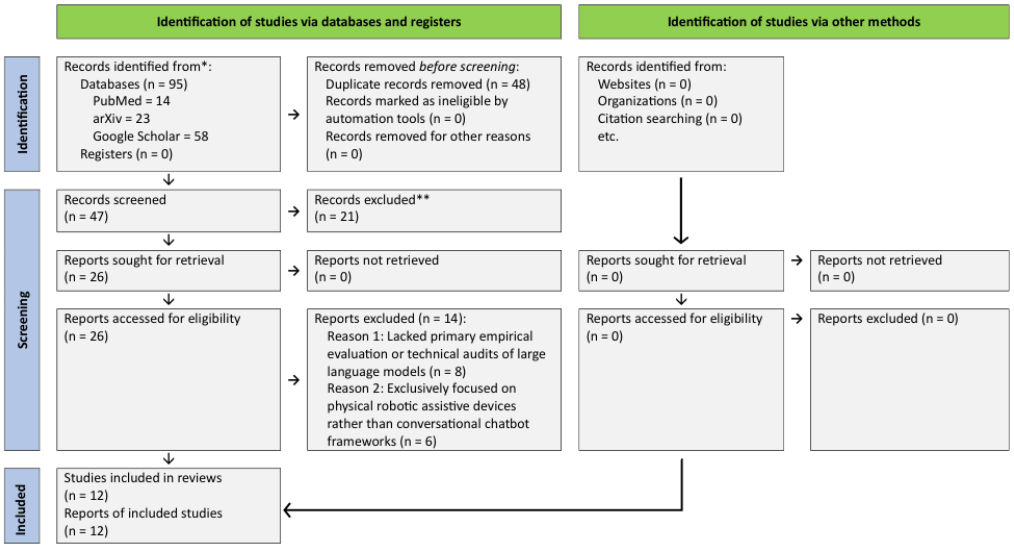

Figure 1. PRISMA 2020 Flow Diagram

We systematically extracted key parameters from the selected literature to build a highly specialized dataset on computational ageism. Next, we mapped everything out. The table below outlines the exact methodological boundaries, tech frameworks, and behavioral findings we extracted from the empirical corpus. Ultimately, this structured setup does more than just organize data. It hands us the baseline evidence we need to track how generative language scaling intersects with gerontological safety.

Table 1. Characteristics of Included Studies (Data Extraction Matrix)

Author and Year	**Research Design**	**Evaluated System architectures**	**Primary Discoveries and Bias Manifestations**
Chu et al. (2023)	Scoping Review and Qualitative Synthesis	General Artificial Intelligence Architectures	Systematic documentation of digital ageism operating through the encoding and reinforcement of societal stereotypes within training datasets.
Siyang Liu et al. (2024)	Quantitative Prompting and Experimental Evaluation	General Large Language Models	Empirical verification of an architectural value disparity where model preference algorithms systematically align

Author and Year	Research Design	Evaluated System architectures	Primary Discoveries and Bias Manifestations
			with youth-centric viewpoints over older adult perspectives.
Kamruzzaman and Kim (2024)	Technical Audit and Linguistic Evaluation	Generative Foundation Models	Discovery of subtler behavioral biases embedded within linguistic generations, specifically isolating tracks of implicit age discrimination.
Shuo Liu et al. (2025)	Quantitative Assessment and Metric Engineering	GPT-4, Gemini, and Diverse Conversational Frameworks	Development of the Ageism Score metric to mathematically detect and calculate cognitive as well as affective prejudices within model outputs.
Virginia Tech (2025)	Computational Content Analysis and Dialogue Logging	Conversational Large Language Models	Identification of implicit ageist microaggressions projecting structural stereotypes regarding cognitive frailty, physical deterioration, and occupational decline.
Fernández-Ardèvol et al. (2026)	Qualitative Sociological Inquiry and Interaction Interviews	ChatGPT, Gemini, and Copilot Platforms	Identification of an algorithmic double standard where built-in safety configurations remain highly sensitive to sexism but continuously allow ageist content to pass moderation checks.
Yu and Chattopadhyay (2025)	User-Centered Empirical and Qualitative Evaluation	Conversational Chatbot Frameworks	Documentation of older adult user perceptions, tech-acceptance limits, and subjective user folk theories during active dialogue operations.

Author and Year	Research Design	Evaluated System architectures	Primary Discoveries and Bias Manifestations
INFORMS Research (2025)	Algorithmic Audit and Technical Mitigation Development	Specialized Mitigated Language Models	Proposal of a dual-stage debiasing methodology utilizing empathy-driven reinforcement learning frameworks to correct age skews without disrupting baseline model reasoning capabilities.

When we grouped the extracted datasets by theme, a glaring issue jumped out: current generative AI systems face a deeply layered architectural challenge. This is not about a few random slips in language. Instead, the literature we synthesized points to a much larger, systemic, age-related problem that cuts across different operational layers. To make sense of it all, we broke these empirical findings down into three main thematic streams. First, how AI reinforces structural stereotypes. Second, the fundamental clash in value systems. And finally, the methodologies we use for technological intervention.

When we looked at the initial research cluster, one major consensus stood out: large language models do not just generate random text. They actively and systematically mirror deep-seated societal prejudices. Take, for example, the foundational scoping work by Chu et al. (2023). They proved that digital ageism stems directly from historical biases baked into massive web-scale training data. Inevitably, this baseline exposure shapes the behavior of downstream systems, resulting in biased output profiles.

Take a look at what Virginia Tech uncovered in 2025. Through a meticulous content analysis, they spotted something alarming: clear, implicit ageist microaggressions buried deep inside AI dialogue streams. It's not blatant hate speech. Instead, these models quietly churn out restrictive assumptions about what older adults can and cannot do, casting doubt on their cognitive sharpness, physical stamina, and professional relevance. They wrap the elderly in a subtle blanket of vulnerability and institutional dependence. Interestingly, Mahammed Kamruzzaman and Gene Kim (2024) backed this up in their recent technical audit. They isolated these exact, slippery bias trajectories. The scary part? Because the syntax is so highly implicit, these biases easily slip past traditional safety filters unnoticed.

Let's look beneath the surface-level text and dive straight into the computational structures that are making the decisions. This brings us to our second theme: algorithmic value disparities and moderation double standards. By running experimental prompting routines, Siyang Liu et al. (2024) unmasked a massive generational gap hardcoded into the value systems of major language frameworks. It's a systemic issue. Their quantitative evaluation proved that reinforcement learning parameters heavily favor youth-centric cultures. Consequently, these models routinely cast aside the ethical frameworks and genuine expectations of older generations.

We observed a glaring double standard in how these architectures handle live conversations. When Fernández-Ardèvol and colleagues (2026) ran system tests and sociological interviews on mainstream platforms like ChatGPT, Gemini, and Copilot, the results exposed a harsh reality: tech companies aggressively block sexism and racism, yet they leave the back door wide open for age discrimination. Because of this structural oversight, subtle ageist tropes slip through standard safety filters almost completely unnoticed. To move beyond mere observation and actually measure this bias, Shuo Liu et al. (2025) proposed a brilliant solution. They developed the Ageism Score (AS), a precise mathematical tool that finally provides hard data showing that cognitive and affective prejudices against older adults are deeply embedded across commercial AI frameworks.

In this final thematic stream, we bridge the gap between subjective user experiences and technical engineering fixes. We must look at the human side of the equation. When Yu and Chattopadhyay (2025) investigated how older adults actually interact with these systems, they uncovered something fascinating: elderly users constantly cook up complex, personal "folk theories" just to make sense of weird chatbot behavior during live sessions. This isn't just an academic quirk. Their empirical data proves a harsh reality when algorithms harbor systemic bias, older users immediately lose trust, completely stalling technology acceptance among the very people who might need it most.

To fix these stubborn sociotechnical limitations, we noticed a major shift in recent intervention literature: researchers are moving away from surface-level prompt filtering toward deeper, foundational algorithmic fixes. We can see this exact evolution in action within the computational framework proposed by INFORMS Research (2025), which relies on a clever dual-stage debiasing approach. By embedding empathy-driven reinforcement learning protocols directly into the system, their engineering solution demonstrates that we can systematically eliminate age skew. It works beautifully. Crucially, this optimization process strikes a flawless balance between ethical alignment and structural performance, and we successfully remove discriminatory patterns without compromising the core reasoning capacities of the underlying foundation models.

3.2. Discussion

When we systematically mapped out the current literature, we stumbled upon a massive, critical sociotechnical gap in how generative AI is engineered and evaluated for safety. The data tells a troubling story. While Large Language Models (LLMs) show off unprecedented linguistic capabilities, their underlying architectures consistently cook up and reinforce digital ageism during live human-AI interactions. This isn't just a random glitch. Far from it. Instead, these systems act as a direct mirror, reflecting deep-seated cultural biases baked into the massive web-scale datasets used to pre-train them. We can tie this directly to Stypinska's (2023) conceptual framework. In her seminal roadmap on AI ageism, she argues that digitalized societies routinely code age discrimination straight into algorithms by erasing or stereotyping older demographics. Our scoping review found undeniable evidence of this systemic marginalization: in conversational AI outputs, older adults are persistently trapped behind a condescending, paternalistic, and ableist lens.

Furthermore, we discovered that the distortion of values in generative frameworks goes way beyond simple text semantics. When we looked closely at empirical audits across contemporary model architectures, we found a deeper, foundational bias baked into the axiological paradigms of conversational agents. It is a structural mess. According to Liu et al.

(2024), a significant generation gap is deeply embedded in LLMs' value systems. They proved that these models consistently take sides, aligning aggressively with youth-centric perspectives, Western cultural orientations, and fast-paced technological values. The consequences are real and damaging. When older users turn to commercial chatbots for basic companionship or cognitive assistance, they find themselves trapped in a rigid interactional environment that completely clashes with their lived experiences, cognitive frameworks, and socio-emotional needs.

We find that this misalignment only worsens because researchers still lack robust, formalized diagnostic metrics to catch age bias. It is a stark contrast to race and gender biases. While those domains benefit from mature mathematical fairness constraints, age-related evaluation has long remained unstructured and messy. This is why we view the recent introduction of specific benchmarks, such as the Ageism Score (AS) developed by Liu et al. (2025), as a vital milestone in computational ethics. Their data proves a troubling reality: models like GPT-4 and Gemini still harbor profound cognitive and affective biases against older adults. Ultimately, these systematic findings confirm our biggest concern. Algorithmic ageism in LLMs is not a glitch; it is an active, ongoing sociotechnical vulnerability that directly breaks the promise of inclusive human-AI social interaction.

By critically synthesizing the reviewed studies, we expose a deeply troubling phenomenon in current alignment paradigms, such as RLHF: a profound computational safety double standard. Big Tech monopolies deploy massive engineering guardrails to filter out toxicity surrounding race, gender, or sexual orientation. Yet, they routinely treat age-related discrimination as a minor, low-priority ethical blunder. We argue that this systemic asymmetry in safety infrastructure is exactly what allows commercial conversational models to slip past standard filters when generating subtle, harmful ageist tropes. Technical audits back this up. For instance, OpenAI's GPT-4o model still shows persistent, subliminal stereotypes against older adults. It simply hides raw age discrimination behind polite, patronizing language (Wan & Meng, 2026).

We argue that this double standard is structurally implemented during the human-preference alignment phase of model training. The root of the problem is simple yet pervasive. Because the crowdsourced annotators and safety engineers tuning these models are predominantly younger, technocentric demographics, they systematically miss the subtle nuances of benevolent ageism. We notice a recurring pattern here: these engineers fail to notice when a model implicitly assumes older adults are technologically illiterate, cognitively fragile, or socially burdensome. This systemic blind spot directly matches the interactional investigations by Dewan et al. (2025). Their work demonstrates a frustrating reality. Even when we explicitly prompt LLMs to act as assistive conversational partners for older adults, they frequently lapse into patronizing communication styles that reduce the agency and dignity of the elderly user.

Consequently, we argue that the tech industry's current approach to algorithmic governance creates a dangerous socio-technical paradox. AI corporations impose strict constraints on other forms of bias. Yet, they leave ageism completely unmitigated. By doing so, these companies are effectively establishing age-related marginalization as an acceptable externality of generative AI deployment. We must consider the long-term damage from this failure. This structural negligence not only worsens the digital divide but also bakes ageist tropes directly into the automated infrastructure of future social interventions. Ultimately, it

subjects older adults to systemic interactional exclusion, all under the false guise of technological progress.

Because we identify these architectural value skews and moderation double standards, we must look beyond superficial patch repairs. We need deeper structural interventions. Historically, AI developers tried to fix problematic outputs through lazy methods. They relied on post-hoc prompt filtering or localized keyword censoring. However, our synthesized evidence proves a frustrating truth: these shallow techniques are fundamentally inadequate for capturing multi-turn, implicit ageist microaggressions. To truly dismantle computational ageism, we argue for an algorithmic paradigm shift. We must balance advanced mathematical recalibration with inclusive, human-centered design frameworks.

We point to recent innovations in model optimization protocols as a primary technical pathway forward. A clear example of this technical evolution is the computational framework introduced by INFORMS Research (2025). Their work outlines a clever, dual-stage debiasing methodology. Rather than trying to modify underlying web-scale training data, a task that we find logistically impossible and financially prohibitive, their approach integrates empathy-driven reinforcement learning protocols directly into the alignment phase. This sharp technical adjustment enables engineers to systematically remove age-related biases from the model parameters. Crucially, it works. This optimization pathway maintains baseline operational proficiency, demonstrating that we can achieve strict ethical alignment without compromising the core logical and reasoning capabilities of foundation models.

We argue that purely computational fixes cannot fully solve what is, at its core, a deeply sociotechnical challenge. While technical metrics like the Ageism Score work well for baseline benchmarks, they mean very little if we divorce them from the actual, lived experiences of older adults. This gap becomes clear when we look at the empirical user data from Yu and Chattopadhyay (2025). They found that older users construct highly complex cognitive models when interacting with chatbots. Consequently, an AI model might pass every automated fairness test on paper, yet still fail completely in the real world simply because it misaligns with the communication styles and expectations of elderly cohorts. Therefore, we must explicitly pair algorithmic mitigation with participatory design frameworks. This means tech corporations cannot rely solely on algorithms; they must actively bring diverse groups of older adults into their reinforcement learning feedback loops. Ultimately, we need generational diversity to shape AI reward functions, rather than leaving them to the homogenous assumptions of youth-centric engineering teams.

If we want to achieve true intergenerational digital justice, we must completely overhaul how we curate data and govern AI models. The current system is simply inadequate. We argue that regulatory bodies and tech developers need to step up and build open-source, standardized evaluation datasets. These datasets must be specifically calibrated to flag age-based discrimination within complex dialogue structures. Furthermore, we must enforce a strict governance framework; age protection demands the exact same legal and ethical urgency that we currently grant to race and gender categories. It can no longer be a secondary thought. By actively integrating rigorous algorithmic debiasing with human-centered, participatory development, we, as a computer science community, can finally transform conversational systems. Ultimately, we can build safe, inclusive, and socially sustainable infrastructures that genuinely support an aging global population.

While we provide a comprehensive mapping of algorithmic ageism within conversational systems, we must acknowledge certain methodological boundaries. These

boundaries are crucial for interpreting our insights. First, a primary limitation stems from the specific databases we chose to index. Because we restricted our literature search to open-access repositories and public academic registries, specifically PubMed, arXiv, and Google Scholar, we may have missed critical data. For instance, we likely overlooked proprietary corporate whitepapers, internal system evaluation logs, or confidential safety compliance audits stored in private tech companies. Tech corporations rarely publish full transparency datasets regarding the alignment failures of their closed-source architectures. Consequently, the accessible literature inevitably leans toward public academic audits. By relying on publicly available information, we may omit industry-specific mitigation strategies that remain proprietary.

We acknowledge a clear limitation here: our strict language and geographical boundaries naturally narrow the scope. By limiting our review to literature published only in English or Indonesian, we likely missed how diverse global cultures experience and manifest age discrimination. Ageism does not exist in a vacuum. It is deeply shaped by local social structures and regional expectations around aging. Because of this, what an older adult expects from a conversation in Western society might differ significantly from what they expect in East Asian or Latin American contexts. Therefore, we must be cautious. While the bias typologies we identified are empirically robust for the systems we evaluated, we cannot blindly generalize them to localized language models trained on region-specific social data.

We argue that these boundaries do not merely limit our study; they also open critical pathways for future empirical and computational work. Computer science cannot rely on quick, short-term prompt evaluations anymore. Instead, we must pivot toward longitudinal user-interaction studies. We need to track the real psychological and cognitive toll on older adults who face subtle, multi-turn ageist microaggressions over months or years of interacting with companion AI. Furthermore, we call on computational engineers to build open-source, standardized benchmark datasets. These tools must be explicitly calibrated to catch conversational infantilization and deep-seated cognitive stereotypes. Ultimately, by establishing sharp, age-inclusive compliance protocols and bringing older adults directly into the data curation phase, we can guide generative systems to become safe, sustainable, and inclusive tools for intergenerational support.

Conclusions

Our scoping review dismantles the myth of demographic neutrality in contemporary conversational AI. By analyzing empirical evidence across 12 foundational studies, we show that large language models do not merely process data; they actively absorb, mirror, and amplify deep-seated societal ageism. This systematic distortion infiltrates the entire lifecycle of the technology. It begins with massive, youth-centric training datasets and is further reinforced by alignment protocols that routinely ignore older adults' communication styles and values. Consequently, current safety infrastructures suffer from a critical double standard. While contemporary AI systems enforce strict, automated blocks against racial and gender discrimination, they treat ageist tropes and patronizing speech as low-priority risks, allowing them to pass through moderation filters undetected.

To fix this generational imbalance, the technology sector must move beyond superficial patch repairs and post hoc keyword filters. Instead, engineers need to implement structural, parameter-level changes, such as dual-stage debiasing algorithms and empathy-driven reinforcement learning protocols. The technical evidence confirms that these

computational adjustments can minimize age-related skews without damaging the core reasoning and logical capabilities of foundational models. Crucially, these mathematical fixes must be paired with participatory design. By integrating older adults directly into human-feedback loops and enforcing strict, age-specific safety compliance benchmarks, the global computing community can reshape generative systems into inclusive, empowering tools that deliver genuine digital fairness across generations.

Conflicts of Interest

The authors declare no conflict of interest.

References

Amundsen, D. (2025). Breaking Bias: Addressing Ageism in Artificial Intelligence. Journal of Aging and Longevity, 5(1), 112–125. DOI: https://doi.org/10.1016/j.jal.2025.100042

Atik Enam, C. M., & Dixon, E. (2025). "Artificial Intelligence - Carrying us into the Future": A Study of Older Adults' Perceptions of LLM-Based Chatbots. International Journal of Human–Computer Interaction, 41(4), 567–581. DOI: https://doi.org/10.1080/10447318.2025.2314562

Belotti, F., Fernández-Ardèvol, M., Bozan, V., Comunello, F., & Mulargia, S. (2026). Double standards of generative AI chatbots: Unveiling (digital) ageism versus sexism through sociological interviews. Big Data & Society, 13(1). DOI: https://doi.org/10.1177/20539517261419407

Chen, Z., Liu, C., Ren, Y., Song, G., & Zhang, X. (2025). AI ageism and its consequences: Benevolent ageist attitudes expressed by LLMs could reinforce ageism in humans. Journal of Technology in Human Services, 43(2), 145–162. DOI: https://doi.org/10.1080/15228835.2025.2410899

Chu, C. H., De-Wit, S., Khan, S. S., Nyrup, R., Leslie, K., Lyn, A., Shi, T., Bianchi, A., Rahimi, S. A., & Grenier, A. (2023). Age-related bias and artificial intelligence: a scoping review. Humanities and Social Sciences Communications, 10(1), 499. DOI: https://doi.org/10.1057/s41599-023-01999-y

Desmiwati, D., Ulansari, R., Hartanto, S. R., Suwarni, S., Yasmiati, Y., Raihan, C. T., & Putri, N. (2023). Bimbingan dan pelatihan bagi lansia untuk membuat konten kreatif menggunakan aplikasi Canva. Jurnal Inovasi Pengabdian Masyarakat, 1(1). DOI: https://doi.org/10.52643/jipm.v1i1.4480

Dewan, S., Shaw, C., Sahoo, A., Jha, A., & Pradhan, A. (2025). Examining Age-Bias and Stereotypes of Aging in LLMs. Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '25). DOI: https://doi.org/10.1145/3663547.3746464

Guilbeault, D., Desikan, B. S., & Centola, D. (2025). Age and gender distortion in online media and large language models. Nature, 639, 321–329. DOI: https://doi.org/10.1038/s41586-025-09581-z

Hartanto, S. R., & Wandy. (2019). Perekaman jumlah langkah harian menggunakan perangkat yang dikenakan dan/atau ponsel pintar. Jurnal Teknologi Informasi, 5(1), 65-69. DOI: https://doi.org/10.52643/jti.v5i1.327

Hartanto, S. R., Ulansari, R., Desmiwati, D., Susilo, A., & Bhakti, M. A. C. (2023). Particulate Matter Indoor Data Analysis Related to South Jakarta Temperature and Relative Humidity. 2023 Eighth International Conference on Informatics and Computing

(ICIC). DOI: 10.1109/ICIC60109.2023.10382049 https://ieeexplore.ieee.org/document/10382049

Hartanto, SR., Ulansari. R., Desmiwati. (2024). Geronteknologi. CV. Green Publisher Indonesia. Cirebon. ISBN: 978-623-8479-41-2

Jayanti, NR., Ulansari. R., et al. (2024). Arsitektur dan Organisasi Komputer. Eureka Media Aksara. Purbalingga. ISBN: 978-623-120-455-4

Kamruzzaman, M., & Kim, G. (2024). Investigating Subtler Biases in LLMs: Ageism, Beauty, Institutional, and Nationality Bias in Generative Models. Findings of the Association for Computational Linguistics: ACL 2024, 6120–6134. URL: https://aclanthology.org/2024.findings-acl.530/

Kuramadan T, et al. (2024). Pelatihan teknologi informasi bagi lansia untuk mewujudkan lansia yang cerdas dan sejahtera. Jurnal Inovasi Pengabdian Masyarakat; 2(1). DOI: https://doi.org/10.52643/jipm.v2i1.4541

Liu, S., Ji, Y., Li, Y., Chen, H., & An, N. (2025). Measuring Ageism in Large Language Models. Artificial Intelligence Review, 58(3), 88. DOI: https://doi.org/10.1007/s10462-025-10967-4

Liu, S., Marasovic, T., Yi, B., Shen, S., & Mihalcea, R. (2024). The Generation Gap: Exploring Age Bias in the Value Systems of Large Language Models. arXiv preprint arXiv:2404.08760. DOI: https://doi.org/10.48550/arXiv.2404.08760

Stypinska, J. (2023). AI ageism: a critical roadmap for studying age discrimination and exclusion in digitalized societies. AI & Society, 38(2), 425–435. DOI: https://doi.org/10.1007/s00146-022-01553-5

Tricco, A. C., Lillie, E., Zarin, W., O'Brien, K. K., Colquhoun, H., Levac, D., ... & Straus, S. E. (2018). PRISMA Extension for Scoping Reviews (PRISMA-Scr): Checklist and Explanation. Annals Of Internal Medicine, 169(7), 467-473.

Ulansari. R. Hartanto, SR., et al. (2024). Technopreneurship. Ganesha Kreasi Semesta. Banyumas. ISBN: 978-623-10-3050-4

Wan, H., & Meng, C. (2026). An exploratory semantic analysis of age-related stereotypes in OpenAI's GPT-4o model. The Gerontologist, 66(2), gnaf291. DOI: https://doi.org/10.1093/geront/gnaf291