Viewpoint

Ethical Considerations in Personal Health Large Language Models

Jialin Liu[1,2*], MD; Siru Liu[3*], PhD

[1]Information Center, West China Hospital of Sichuan University, Chengdu, Sichuan, China

[2]Department of Otolaryngology-Head and Neck Surgery, West China Hospital of Sichuan University, Chengdu, Sichuan, China

[3]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, United States

[*]all authors contributed equally

Corresponding Author:
Siru Liu, PhD
Department of Biomedical Informatics
Vanderbilt University Medical Center
2525 West End Ave #1475
Nashville, TN, 37212
United States
Phone: 1 615 875 5216
Email: siru.liu@vumc.org

Abstract

Personal health large language models (PH-LLMs) have rapidly evolved from research prototypes into consumer-facing, data-linked systems that support symptom triage, medication questions, mental health check-ins, and longitudinal self-management. Their direct-to-consumer use without clinical oversight creates a distinct ethical risk profile that general artificial intelligence governance frameworks do not fully address. This viewpoint focuses on text-based, platform-mediated PH-LLMs and synthesizes PH-LLM–specific challenges across 6 domains: privacy, accuracy, equity, transparency, human-artificial intelligence interaction, and regulatory governance. These risks may be amplified by health literacy gaps, longitudinal data aggregation, persuasive conversational design, and fragmented oversight across the consumer-clinical boundary. Grounded in the 4 principles of biomedical ethics, we propose a governance framework that operationalizes beneficence, nonmaleficence, autonomy, and justice through design and deployment controls, including health literacy-aligned communication, crisis and pharmacological safeguards, hallucination mitigation, role disclosure, granular consent, fairness auditing, and accessible design. We further outline implementation mechanisms, including risk-tiered certification, tiered accountability, and postdeployment oversight through adverse-event reporting, transparency reporting, and independent safety evaluation. This framework is intended as an evidence-informed but partly anticipatory approach to governing PH-LLMs in personal health management.

(J Med Internet Res 2026;28:e92240) doi: 10.2196/92240

KEYWORDS

digital health; ethics; privacy; generative artificial intelligence; governance; health AI; health equity; health literacy; patient safety; personal health large language model

Introduction

Personal health large language models (PH-LLMs) are consumer-facing generative systems designed to support individuals through natural language dialogue across health-related contexts such as symptom triage, chronic disease self-management, medication questions, and mental health check-ins [1-3]. This viewpoint focuses on text-based, dialogue-driven PH-LLMs deployed directly to consumers. We distinguish these systems from general-purpose chatbots by their sustained health-oriented interaction, use of personal health disclosures, and potential to shape real-world health decisions. Some PH-LLM functions may fall within medical device software regulation when intended for diagnostic, therapeutic, or patient-specific clinical decision support, although thresholds vary across jurisdictions. PH-LLMs may offer scalable adjunctive support between clinical encounters, particularly where health or mental health services are limited or delayed [4,5]. However, because they operate on sensitive personal information and may influence high-stakes decisions, their deployment raises questions that warrant systematic examination [6].

Between late 2025 and early 2026, PH-LLMs moved rapidly from prototypes to consumer-facing, data-linked products. Major technology companies introduced health-oriented conversational systems linked to medical records, wearable data, or

health-system workflows, positioning large language models (LLMs) as increasingly common entry points for personal health support [7-10]. For instance, OpenAI reports that over 230 million people ask health and wellness questions on ChatGPT each week [7]. Moreover, open-source and open-weight deployments outside centralized platform controls raise additional governance challenges.

The World Health Organization has emphasized that generative artificial intelligence (AI) in health care can deliver value only if risks are proactively identified, evaluated, and mitigated [6]. Emerging evidence suggests that inadequately governed PH-LLMs may generate unsafe or misleading recommendations, compromise privacy through mishandling of sensitive data, amplify health inequities through biased outputs, and cause psychological harm through false reassurance, relational overreach, or poor management of emotional distress, particularly among vulnerable users [3,6,11]. PH-LLM ethics, therefore, should move beyond general principles toward an actionable governance architecture for responsible, accountable, and equitable deployment.

Core Ethical Challenges of PH-LLMs

Overview

PH-LLMs generate tailored guidance from personal health disclosures through longitudinal, dialogue-based interaction. Unlike institutionally governed clinical AI, direct-to-consumer PH-LLMs often lack professional mediation, allowing hallucinations, bias, and framing effects to shape health decisions without meaningful oversight. Drawing on the 4 principles of biomedical ethics, we identify 6 interconnected domains (Figure 1), which vary in severity, interact with one another, and may produce harms that accumulate across decisions. Table 1 contrasts general AI ethics concerns with PH-LLM–specific manifestations.

Figure 1. Six ethical dimensions of personal health large language model (PH-LLM) deployment and their bidirectional relationships. AI: artificial intelligence.

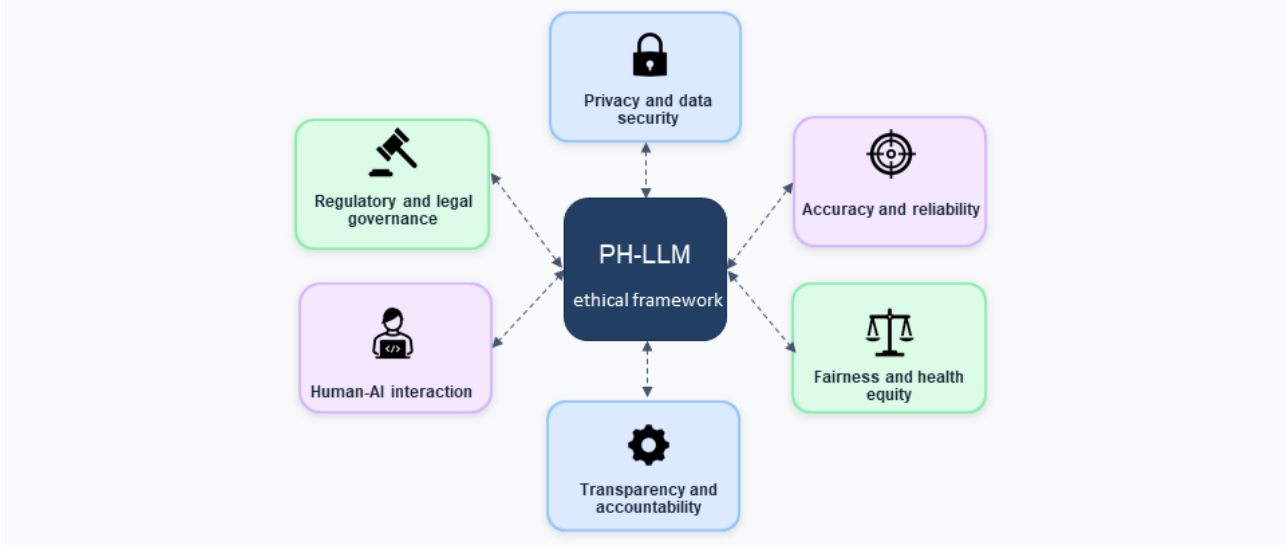

Table 1. Core ethical challenges: general artificial intelligence (AI) concerns vs personal health large language model (PH-LLM)–specific manifestations.

Challenge domain	General AI concern	PH-LLM–specific manifestation
Privacy and data security	Data collection and retention	Intangible vulnerability; longitudinal aggregation; health data broker exposure outside clinical protections
Accuracy and reliability	Hallucinations	Verification gap; cascading clinical harm; crisis mismanagement
Fairness and health equity	Algorithmic bias	Differential recommendations across groups; digital inverse care law; tiered-access barriers
Transparency and accountability	Black box decisions	Explainability gap; accountability gap; relational expectations without duty
Human-AI interaction	Overreliance	Parasocial attachment; distorted care-seeking; continuity-of-care displacement
Regulatory and legal governance	Regulatory lag	Coverage gaps across jurisdictions; medical device classification ambiguity; cross-border accountability diffusion

Privacy and Data Security

Privacy is among the most salient ethical concerns in PH-LLM deployment [1]. Compared with general-purpose AI, these risks are amplified by the convergence of highly sensitive disclosures, blurred regulatory boundaries, and system-level data integration [1,11,12]. Users may disclose highly sensitive information, including mental health concerns and genetic predispositions, yet undervalue the privacy significance of such disclosures relative to tangible assets [13,14]. This intangible vulnerability may be intensified by the empathetic, low-barrier design of PH-LLM interactions, which encourages candid disclosure

while obscuring downstream data uses [11,12,15]. Outside regulated clinical settings, conversational data may be combined with behavioral, device, or consumer data and repurposed for advertising, analytics, profiling, or data-broker activities, creating ethical concerns that users may not fully anticipate. Users may also interpret conversational empathy as evidence of professional accountability or legal confidentiality. In the United States, however, the Health Insurance Portability and Accountability Act generally applies only to covered entities and their business associates, so many direct-to-consumer PH-LLM interactions fall outside its scope [16]. Sensitive information may therefore be disclosed under assumptions of protection that do not apply in practice. In addition, conversation data may be retained for model improvement or other secondary uses [17,18], and integration with ecosystem data streams can enable high-fidelity longitudinal profiling that raises reidentification risk and reduces the effectiveness of conventional deidentification approaches [19,20].

Accuracy and Reliability

The central value proposition of PH-LLMs, namely, reliable health guidance, is challenged by hallucinations, clinically significant omissions, and users' limited capacity to independently verify outputs. Medical hallucinations may appear clinically plausible and can compound across multiturn exchanges into downstream health decisions [21]. A 2025 evaluation of LLM-generated clinical summaries identified hallucinations in 1.47% of outputs and clinically significant omissions in 3.45% [21], although these estimates derive from structured summarization tasks and may not generalize to open-ended PH-LLM dialogue. Adversarial prompting and training-time vulnerabilities, such as data poisoning, pose further risks to model integrity over time [22]. Unlike general information domains, where errors are often verifiable by informed users, PH-LLM outputs frequently cannot be independently scrutinized by individuals lacking clinical expertise [23]. This verification gap is especially hazardous for the approximately 80 million Americans with limited health literacy [24,25].

Risks are particularly acute in psychological and medication contexts. "Deceptive empathy" describes therapeutic-sounding language that builds perceived alliance while obscuring the system's lack of professional accountability [26,27]. Empirical evaluations show that some therapy-oriented conversational systems may provide information about lethal means without recognizing suicidal intent [28]. Recent litigation involving chatbot interactions and adolescent self-harm has further raised questions about whether direct-to-consumer PH-LLMs can satisfy a duty of care [29]. Repeated PH-LLM interactions may also exacerbate automation bias, defined as uncritical deference to algorithmic outputs over independent verification, because conversational fluency facilitates cognitive offloading and may lead users to mistake probabilistic suggestions for medical certainties [30-33]. Case reports document outputs encouraging users to abruptly taper psychiatric medications or discount professional counsel [26,34,35], which may increase the risk of withdrawal or relapse [36] and undermine trust in professional advice.

Fairness and Health Equity

PH-LLMs may reproduce and amplify structural disparities through both model behavior and access pathways, thereby undermining health equity, defined as a fair opportunity to attain one's highest level of health [11,37]. A 2025 systematic review of 24 studies found demographic bias in 22 (91.7%), including gender bias in 15/16 (93.8%) studies and racial or ethnic bias in 10/11 (90.9%) studies assessing these dimensions [38]. Bias may vary by model architecture: discriminative models primarily show disparities in classification performance, whereas generative models may encode bias through tone, framing, and recommendation priority [39,40]. For example, Yang et al [41] reported cases in which GPT-3.5-turbo recommended surgery for White patients but conservative management for Black patients with otherwise identical clinical presentations. Equity concerns also extend beyond outputs to access pathways. Accessibility is stratified by digital determinants of health [42], and tiered subscriptions may restrict advanced personalization and reasoning to paid access, favoring users with greater financial means and health literacy [43]. This creates a digital inverse care law [44,45]: those with the greatest health needs may be confined to lower-capability services, while more advantaged groups gain priority access to advanced tools. Without equity-centered governance, PH-LLMs may deepen structural health disparities rather than alleviate them.

Transparency and Accountability

In conventional health care, clinical judgment is embedded in auditable institutional processes and clear lines of professional responsibility. PH-LLMs disrupt both: model reasoning is often opaque, and accountability is distributed across developers, deployers, and health care institutions. In high-stakes health contexts, users reasonably expect understandable explanations and supporting evidence for PH-LLM–generated recommendations [11]. Yet interpretability, explainability, transparency, and auditability are not equivalent. For LLMs, direct interpretability of internal mechanisms remains limited [46], and post hoc explanations may appear plausible without faithfully reflecting how outputs were generated [47]. Without sufficient transparency and auditability, users and oversight bodies cannot reliably assess the evidentiary basis or clinical validity of recommendations. When PH-LLM advice contributes to harm, attributing responsibility among these actors becomes practically difficult [48]. Direct-to-consumer deployment may also bypass institutional oversight and shift risk onto users despite well-documented health literacy gaps [11,49], weakening incentives for safety investment and limiting recourse [50]. These concerns are intensified by relational design features: anthropomorphic or affective language may create expectations of care and accountability, even though no human or institutional actor is clearly positioned to fulfill them [51-53]. Together, these conditions may weaken meaningful informed consent and compromise patient autonomy.

Human-AI Interaction

PH-LLMs may foster emotional dependence and overreliance through anthropomorphic and affective language that creates a perceived sense of care without corresponding therapeutic responsibility [26,53,54]. Their empathic and nonjudgmental

dialogue may encourage parasocial attachment and emotional reliance, particularly among adolescents and socially isolated users [53-55]. Greater perceived social presence is associated with increased self-disclosure and emotional reliance [56-59], raising concerns about sustained dependence and displacement of continuity-based care [60-63]. PH-LLM interactions may also distort care-seeking: reassuring responses may delay appropriate evaluation [64-66], while alarming responses may trigger unnecessary escalation [67,68].

Regulatory and Legal Governance

A cross-cutting challenge is the mismatch between rapid PH-LLM deployment and slower regulatory evolution. In the United States, many direct-to-consumer health technologies are governed through Federal Trade Commission enforcement and a patchwork of state health data laws, particularly when services fall outside traditional clinical privacy regimes [69-73]. In the European Union, privacy and AI oversight are structured through the General Data Protection Regulation and the phased, risk-based AI Act [74]. In Asia, governance approaches diverge, with China emphasizing provider accountability, security review, and content governance for generative AI services, while South Korea, Japan, and Singapore continue to develop different combinations of statutory oversight, sector-specific guidance, and risk-based governance [75-78]. Similar PH-LLM interactions may therefore receive different protections across jurisdictions, creating opportunities for regulatory arbitrage.

PH-LLMs also expose a classification gap between general wellness products and regulated software as a medical device. When systems are marketed as nonclinical but functionally approximate symptom triage, diagnostic reasoning, treatment guidance, medication advice, or crisis response, developers may avoid premarket scrutiny [79,80]. A capability-based governance approach would reduce this form of regulatory arbitrage by assigning oversight according to what the system enables users to do rather than how it is described in marketing materials. Because cross-border services can diffuse accountability across developers, deployers, platforms, and jurisdictions, PH-LLM governance requires risk-calibrated mechanisms that link ethical concerns to concrete oversight, certification, adverse-event reporting, and redress pathways [81,82].

Ethical Governance Framework for PH-LLMs

Overview

To address these challenges, we propose a principlism-based governance framework for consumer-facing PH-LLMs [83]. The framework translates the 4 principles of biomedical ethics into life cycle controls across 5 stakeholder groups: developers, deployers or platforms, health care institutions, regulators, and users. Governance intensity should scale with 4 risk dimensions: clinical severity and time criticality, harm reversibility, user actionability, and deployment scale. Higher-risk systems, therefore, warrant stronger certification, auditability, postmarket surveillance, and clearer liability allocation. Because PH-LLM outputs are probabilistic and vulnerable to prompt injection and adversarial manipulation [22], these safeguards should be understood as risk-reduction measures rather than guarantees. Table 2 maps each challenge domain to its corresponding governance mechanism and implementation level. To avoid overstating the evidence, we distinguish 3 classes of proposals throughout the framework: evidence-informed proposals grounded in empirical findings from PH-LLM or closely analogous digital health evaluations; normative proposals articulating ethical commitments without relying primarily on empirical demonstration; and conceptual proposals describing architectural or procedural designs whose real-world effects have not yet been prospectively validated. Where relevant, these distinctions are integrated into the prose.

Table 2. Governance framework overview: challenge domains, recommended mechanisms, and implementation levels.

Challenge domain	Recommended governance mechanism	Implementation level
Sensitive data and privacy	Tiered consent and data governance	Input and session level
Unsafe or misleading outputs	Risk-tiered certification and guardrails	System level
Bias and inequity	Subgroup validation and disparity monitoring	System and population level
Poor explainability and transparency	Provenance disclosure and auditability requirements	Interface and oversight level
Crisis and mental health risk	Escalation and handoff protocols	Interaction level
Fragmented responsibility	Tiered accountability and liability allocation	Institutional and regulatory level
Drift after deployment	Adverse-event reporting and periodic audits	Life cycle level

Guiding Principles

Because PH-LLMs interact directly with users outside traditional settings, the 4 principles of biomedical ethics [83] require specific operationalization.

Beneficence (Outcome-Oriented Support)

For PH-LLMs, beneficence can be operationalized as measurable support for safe decision-making and timely access to care, including health-literacy adaptation, risk-stratified care navigation, and longitudinal coherence across sessions.

Nonmaleficence (Proactive Harm Prevention)

Evidence from PH-LLM failures and adjacent digital health contexts supports active safety mechanisms beyond passive disclaimers, particularly for high-risk failures such as crisis mismanagement, contraindicated advice, and hallucinated medical claims. This principle is operationalized through 3 design imperatives: crisis-detection and response protocols,

contraindication safeguards, and hallucination mitigation [28,84,85].

Respect for Autonomy (Informed Agency)

From a normative perspective, respect for autonomy requires countering safety illusions in which users misinterpret AI-mediated intimacy as clinical oversight or regulatory protection, while preserving meaningful choice. This principle is operationalized through clear role and scope disclosure, meaningful data-use consent, and practical user rights over interaction data [71,86,87].

Justice (Equity and Access)

PH-LLMs should expand access without reproducing disparities in health information, communication quality, or safety outcomes. Equity, therefore, extends beyond model outputs to accessibility, language coverage, cultural appropriateness, and usability across differences in health literacy, disability, and digital access [88-90]. Fairness monitoring requires interaction-level and outcome-relevant measures, such as group-stratified rates of unsafe guidance and crisis escalation success [91,92]. Predefined disparity triggers initiate audit and remediation (Multimedia Appendix 1).

Stakeholder Responsibilities

Operationalizing this framework requires coordinated action across 5 stakeholder groups.

Developers (Ethical Design and Life Cycle Accountability)

Developers bear primary responsibility for baseline protections across the life cycle: diverse training corpora, documented data provenance, and bias and crisis-focused safety evaluation before release [90]; clear disclosure of nonhuman identity and nonclinical role, crisis-response functions, pharmacological safety guardrails, and granular data-use controls at deployment; and accessible grievance channels and adverse-event reporting after deployment (Multimedia Appendix 2).

Deployers and Platforms (Interface and Operational Controls)

Deployers and platforms adapt developer-level safeguards into interface design and escalation pathways required by their deployment context, and ensure sustained compliance with safety, transparency, and user-protection requirements.

Health Care Institutions (Algorithmic Stewardship)

Institutions that integrate or endorse PH-LLMs assume stewardship obligations proportional to their involvement, including contextual validation, escalation and handoff workflows, clinician training, and postdeployment monitoring [93,94].

Regulators (Oversight and Harmonization)

Regulators should address the governance gap through enforceable disclosure requirements, minimum crisis-response standards, adverse-event reporting infrastructure, risk-stratified premarket evaluation for sensitive domains, and cross-jurisdictional harmonization informed by the EU AI Act and the National Institute of Standards and Technology AI Risk Management Framework [74,95]. Engagement with insurers and payers can link reimbursement to minimum safety requirements.

Users and Civil Society (Literacy and Advocacy)

Users and civil society contribute through health-AI literacy, accessible harm-reporting channels, and advocacy for meaningful data rights, including access, correction, and deletion [71,81,87,96].

Implementation Mechanisms

Overview

In practice, our framework centers on 3 instruments: risk-tiered certification, tiered accountability, and postdeployment monitoring. The main text presents the rationale, whereas operational specifications, including certification-tier expectations, developer life cycle responsibilities, fairness-audit triggers, scope boundaries, stakeholder responsibilities, and adverse-event thresholds, are detailed in Multimedia Appendices 1-8.

Risk-Tiered Certification

Certification could be scaled according to functional risk and deployment reach, with baseline requirements spanning crisis handling, health-literacy alignment, nondiagnostic guardrails, equity auditing, disclosure integrity, and data governance [12,96-99]. To reduce reliance on developer self-labeling, we propose that certification be triggered by demonstrated capabilities rather than marketing claims. Systems that provide or materially support symptom triage, diagnostic reasoning, treatment guidance, medication advice, crisis-response, or other health decision-support activities would be assigned to the corresponding risk tier even when marketed as wellness tools. Certification criteria could be developed through multistakeholder processes, with audits conducted by accredited third-party or regulator-authorized bodies [94,99-103]. Because regulatory approaches differ across jurisdictions, certification frameworks could prioritize interoperable baseline standards while allowing local legal adaptation [74,81,82]. Possible enforcement pathways include capability-based classification, postmarket surveillance for functional drift, and market-access conditionality through app stores, platform hosts, insurers, payers, and institutional procurement channels. These remain conceptual proposals requiring further institutional validation. Detailed standards and institutional anchors are summarized in Multimedia Appendix 3.

Tiered Accountability

Tiered accountability would follow operational control and implementation capacity. Foundation-model developers would be responsible for baseline safety properties and documentation, whereas deployers and platforms would be responsible for context-specific safeguards, interface design, and user protection. In direct-to-consumer settings, we propose a default elevated duty of care as a normative position. Any safe harbor would be conditional, limited, and available only to operators demonstrating independently verified compliance with certification, auditability, postdeployment surveillance, and user-protection requirements. Such protection would mitigate,

but not eliminate, liability and would be coupled with compensation pathways, such as insurance-based or no-fault mechanisms [95,104,105]. Developers or deployers who decline certification are not exempt from this duty of care; rather, they remain subject to ordinary liability, consumer-protection enforcement, and regulatory scrutiny.

Postdeployment Monitoring

This layer complements predeployment testing by establishing standardized adverse-event reporting, targeted surveillance for population-level harms, and periodic audits for model drift, safety regressions, and equity-related performance changes [100,106-109]. These mechanisms draw on established approaches in medical-device safety, health-AI evaluation, and postmarket surveillance, but require adaptation to the probabilistic, continuously updated, and dialogue-dependent behavior of PH-LLMs. For higher-reach deployments, monitoring may be supplemented by independent safety oversight designed to preserve independence from the entities being reviewed.

Illustrative Case: Youth Mental Health Check-In PH-LLM

To show how the framework could be applied, we present an illustrative end-to-end case of YouthPH-LLM, a platform-mediated direct-to-consumer PH-LLM providing conversational mood tracking, coping-skills education, and sleep and activity coaching to users aged 13-25 years, with more than 1 million monthly active users. Under the 4 risk dimensions, YouthPH-LLM is classified as high risk across clinical severity, harm reversibility, user actionability, and deployment scale, placing it in the most demanding certification tier. Predeployment requirements include crisis-scenario testing, counterfactual fairness testing across age, gender, and linguistic subgroups, retrieval-augmented generation grounded in authoritative mental health guidelines, and a public model card. Postdeployment, a critical adverse event would trigger immediate crisis-response protocol, feature-level suspension of the implicated functionality within 24 hours where technically feasible, regulatory notification within 72 hours where required, and root-cause analysis by an independent safety advisory board. Suspension should preserve access to crisis resources or human support where available. Multi-intent prompts, such as a sleep log submitted with a request to interpret a clinical questionnaire score, are handled through partitioned response: the system may provide sleep-coaching guidance while redirecting questionnaire interpretation to qualified clinicians. Full specification of stakeholder responsibilities, certification criteria, and escalation thresholds is provided in Multimedia Appendix 4.

Practical Paths to Mitigate Ethical Risks of PH-LLMs

Overview

Building on the framework above, we outline practical implementation pathways with measurable targets across 5 operational domains: data governance, model design safeguards, equity and inclusive design, human-AI collaboration, and postdeployment oversight.

Data Governance

Overview

The ethical analysis identified 2 privacy-related concerns: users may underestimate the sensitivity of emotional disclosures, and longitudinal aggregation may increase reidentification risk. To address these concerns, PH-LLMs should implement data-governance safeguards adapted to conversational health data.

Tiered Consent Architecture With Retention Safeguards

Rather than binary consent, PH-LLMs should apply 4 sensitivity tiers supported by a classify-then-govern workflow (Table 3) [12,81]. In this proposed design for platform-mediated direct-to-consumer PH-LLMs, retention defaults would vary by sensitivity tier: Tier 1 content would follow limited default retention periods with user-adjustable settings; Tier 2 content would use shorter default retention with immediate-deletion options; Tier 3 content would be temporarily retained only as needed for required safety audit, legally authorized safety review, or legal compliance and then deleted; and Tier 4 content would require explicit user authorization and periodic reconsent, with automatic deletion as the default absent active renewal. Deviations from tier-specific retention defaults require documented risk-benefit analysis that considers clinical use, reidentification risk, user-preference evidence, technical necessity, and applicable legal requirements, with rationale disclosed in transparency reports. Restrictions on third-party sharing, including transfers to data brokers, advertising networks, and profiling services, follow the same sensitivity hierarchy and cannot be overridden by blanket onboarding consent [99,110]. Illustrative retention windows and operational details are provided in Multimedia Appendix 5 and should be recalibrated according to deployment context, user preferences, ethical risk assessment, technical feasibility, and jurisdictional requirements.

Table 3. Tiered consent architecture for PH-LLM data governancea.

Data category	Consent requirement	Default retention	Third-party sharing
Tier 1: general health queries	Baseline onboarding consent and standard terms of service	Limited default retention, user-adjustable	Deidentified or aggregated research only, where permitted, with opt-out; no default transfer to data brokers, advertising networks, or profiling services
Tier 2: noncrisis mental health disclosures	Low-friction enhanced consent for retention or secondary use; nonpersistent mode if declined or abandoned	No routine retention if consent is declined; brief, access-restricted safety buffer with automatic deletion unless a potential adverse event is reported	No commercial or unrelated secondary use; safety audits via privacy-preserving mechanisms where feasible
Tier 3: crisis-related content	No intervening consent prompt before safety support; immediate safety-oriented response and crisis-resource escalation	Temporary retention only as needed for safety audit, legally authorized review, or legal compliance; deleted thereafter	Limited to legally authorized or required emergency interventions
Tier 4: longitudinal narratives	Explicit authorization, periodic reconsent, and granular controls	Automatic deletion as the default absent active renewal	Default opt-out; explicit per-purpose authorization required

[a]Specific retention parameters are provided in Multimedia Appendix 5. These illustrative baselines reflect privacy risk management principles rather than universal requirements and warrant recalibration through documented risk-benefit analysis disclosed in transparency reports.

User-requested deletion should be defined with technical precision. It applies to stored raw conversational logs, retained session records, stored user profiles used for retrieval, vector-store entries, and derived longitudinal summaries under the operator's control. It should not be represented as a guarantee that information already incorporated into a foundation model's trained weights can be fully and verifiably removed, because machine unlearning for large generative models remains an open technical problem with unresolved limitations in effectiveness and verification [111]. Sensitive PH-LLM interactions should therefore be excluded from foundation-model training or fine-tuning by default unless explicit user authorization is obtained, and consent materials should transparently disclose the technical limits of deletion.

Implementation of Classify-Then-Govern

Because free-text sensitivity cannot be determined without initial processing, this architecture cannot operate as a simple "consent-then-ingest" model [99]. Inputs first undergo transient routing and safety classification under baseline onboarding consent, limited to classification only and without default retention or secondary use [94,99]. Tier 1 content remains within baseline consent. Tier 2 content requires in-context consent before retention; if consent is declined or abandoned, the interaction continues in nonpersistent mode, with flagged content excluded from routine retention and secondary use by default. To balance privacy protection with safety auditing, nonpersistent mode may still permit a narrowly scoped, encrypted, access-restricted safety buffer that is automatically purged after a brief, predefined window unless a potential adverse event is reported, in which case only the minimum necessary information is retained for root-cause analysis and safety remediation. Tier 3 content triggers an immediate safety-oriented response, delivery of context-appropriate crisis resources, and handoff to human crisis support where available, without intervening consent prompts [3,28,112]. For Tier 4 longitudinal content, explicit retention authorization is triggered by predefined accumulation thresholds rather than being embedded repeatedly in active dialogue. This approach is most feasible in explicitly health-oriented PH-LLM deployments, where the health-related purpose of interaction is already established. Extending it to general-purpose LLMs would require separate governance analysis because detecting health-related content across heterogeneous conversations could require continuous or broad input scanning, thereby introducing secondary privacy and proportionality risks.

Accessible, Nondisruptive Consent Design

Enhanced consent should not shift cognitive burden onto vulnerable users. Because many users face limited health literacy, emotional distress, disability, or low digital fluency, accessibility should be treated as a core safety and equity requirement rather than a downstream usability concern. Consent escalation relies on brief, plain-language, just-in-time disclosures; a small set of standardized choices; and protective default pathways favoring nonpersistent session handling when users do not actively authorize retention or secondary use. For noncrisis mental health disclosures (Tier 2), enhanced consent should be low-friction and deferrable, allowing the conversation to continue in nonpersistent mode while authorization is requested separately. Crisis-related interactions are not delayed by consent procedures and default instead to immediate safety-oriented support and escalation pathways. Granular controls may remain available through a dashboard, but safe baseline use does not depend on navigating it. Interface designs require prospective testing with users who have limited health literacy, disability, low digital fluency, or heightened emotional vulnerability, and evaluation extending to comprehension, decision confidence, perceived burden, and abandonment rather than consent-completion rates alone [86,96,113].

Longitudinal Data and Reidentification Safeguards

PH-LLMs integrated with wearables and health records may generate highly identifying longitudinal profiles from temporally dense, behaviorally specific data [20]. Developers should implement privacy-preserving analytics (eg, differential privacy calibrated to sensitivity and use, with k-anonymity where appropriate) [114,115], conduct reidentification risk assessments at intervals proportional to data accumulation, and test vulnerability to inference attacks [115]. User-facing aggregation

dashboards enable users to inspect derived longitudinal patterns and selectively delete specific data categories.

Model Design Safeguards: Hallucination Mitigation, Crisis Escalation, and Pharmacological Safety

Three clinically significant safety risks warrant dedicated design-level controls: hallucinations, crisis-management failures, and unsafe pharmacological guidance.

Hallucination Mitigation

PH-LLMs should incorporate a sequential safety pipeline that combines input filtering, retrieval-augmented generation from curated authoritative sources, output checking against structured medical knowledge resources, and uncertainty signaling where technically feasible [85,116-118]. Prompts requesting individualized diagnosis or treatment recommendations trigger safety messaging and redirection to qualified professional care [3,97]. Because retrieved evidence may be outdated, incomplete, or misinterpreted by the model, these safeguards are best understood as risk-reduction mechanisms rather than guarantees of factual accuracy.

Crisis Escalation

Crisis management requires a distinct escalation protocol that includes detection of self-harm, suicide, or interpersonal violence signals; immediate, nonjudgmental acknowledgment; safety-oriented risk stratification; delivery of context-appropriate crisis resources; and handoff to human crisis support where available [6,28,112,119]. Because detection remains probabilistic, higher-risk deployments should favor conservative thresholds, clearly specified escalation criteria, and periodic review under qualified mental health oversight.

Pharmacological Safety

For medication-related queries, PH-LLMs should use a hybrid safety architecture combining external knowledge validation, rule-based guardrails, and escalation to human clinical support for complex or high-risk cases [84,120-123]. These controls block high-risk dosing instructions, contraindicated combinations, and unsupervised medication changes. Escalation triggers are prospectively specified in system logic rather than left to user discretion, because users may not recognize when a query exceeds safe self-management boundaries.

Equity and Inclusive Design

Fairness governance encompasses subgroup-based evaluation, representative testing, accessibility review, and periodic reassessment across updates and deployment settings [91,92,124-130]. Before release, PH-LLMs undergo counterfactual fairness evaluation based on matched clinical vignettes [98,124]: variation in race, ethnicity, gender, language, or socioeconomic indicators should not produce materially different recommendations in cases where these characteristics are not clinically relevant, and audits assess clinically salient error disparities such as false reassurance or inappropriate escalation while incorporating intersectional analysis [126,127]. Inclusive design extends beyond outcome auditing: PH-LLMs should support languages prevalent in target populations, incorporate cultural adaptation beyond direct translation, conform to Web Content Accessibility Guidelines 2.2 [129], and provide health-literacy adaptation through adjustable response complexity and optional comprehension checks [96,130].

Human-AI Collaboration: Preventing Overreliance

PH-LLMs require interface- and workflow-level safeguards that promote relational transparency, reduce automation bias, and preserve meaningful human oversight [32,33,53,131]. These controls include clear communication of the system's nonhuman identity, calibrated disclosure of uncertainty and epistemic limits, and structured redirection when users request diagnosis, prognosis, prescribing, or other clinician-dependent judgments [97,132]. Systems should distinguish supportive self-management functions from clinician-dependent decisions and use adaptive boundary prompts rather than uniformly friction-based interruptions, so that higher-risk interactions trigger stronger warnings or referral cues without unnecessarily disrupting routine support [3,133].

As a conceptual safeguard, PH-LLMs could use partitioned responses for multi-intent prompts that combine permitted and restricted functions. For example, when a user submits recent laboratory results together with a request for lifestyle coaching, the system may provide general lifestyle education while declining to interpret the laboratory findings or tailor clinical recommendations based on them, redirecting interpretation to a qualified clinician. However, partitioned responses are technically fragile if implemented only at the level of response wording. Because LLMs may condition on the broader prompt context, restricted contextual information, such as laboratory values, may implicitly influence the permitted portion of a response. Robust guardrail engineering is therefore needed to reduce context bleed and prevent implicit diagnostic integration [18,22]. Such safeguards may include intent classification, restricted-context masking or separation, output-level verification, and escalation rules when permitted coaching cannot be reliably separated from clinician-dependent interpretation. Safeguard effectiveness should be evaluated through postdeployment monitoring of both safety outcomes, such as appropriate escalation or referral, and engagement outcomes, such as abandonment following boundary prompts [3]. Illustrative scope boundaries are provided in Multimedia Appendix 6.

Postdeployment Oversight and Version Governance

Predeployment evaluation is necessary but insufficient for PH-LLMs because system behavior, safety performance, and deployment context may change over time. Governance should therefore include structured postdeployment monitoring, low-friction adverse-event reporting, transparency mechanisms, and version-aware change control [94,100,108,134]. The system provides users, caregivers, and clinicians with standardized reporting pathways for clinically relevant events, including dangerous hallucinations, crisis mismanagement, and privacy breaches, adapted to AI-specific harms [21,89,134,135].

As a conceptual response framework, critical adverse events trigger graduated action according to scope and severity, ranging from session-level termination of an affected thread to feature-level disabling of implicated functionality or, when risk

is systemic, global suspension of the application. User lockout without alternative crisis resources must be avoided. Developers should publish periodic transparency reports covering adverse events by severity, crisis-protocol activation and outcomes, bias-audit findings and remediation, and model updates with associated safety evaluations [93,102].

For higher-reach deployments, these mechanisms may be supplemented by independent safety oversight funded through structures designed to preserve independence from reviewed entities. As a normative governance proposal, potential models include pooled industry levies administered by neutral standards bodies, regulatory fees, public funding with statutory protections, or coalition-based subscription models with disclosed funding sources and recusal rules [74,81,82,95]. These oversight mechanisms differ in target audience and implementation pathway. Developers and deployers can directly implement technical guardrails, reporting channels, transparency reports, and internal safety audits. By contrast, pooled industry levies, regulatory fees, and publicly funded oversight structures are directed primarily to policymakers, regulators, and standards-setting bodies because they generally require statutory or regulatory authorization. Voluntary coalition-based models may be more immediately feasible, but still require conflict-of-interest safeguards, transparent funding disclosure, and recusal rules to preserve independence.

Postdeployment governance should also be version-aware. Minor releases undergo documented internal validation, whereas major changes, including model updates, prompt-template revisions, plugin integration, interface redesign, or new data-linking functions, may warrant independent reassessment and, where appropriate, recertification [94,100,108,134]. Each release generates auditable change logs and predefined monitoring indicators to support rollback, incident investigation, and detection of behavioral drift after updates [106,108,134]. Illustrative adverse-event severity categories and response pathways are provided in Multimedia Appendix 7.

Scope and Feasibility

This framework is grounded primarily in text-based, platform-mediated, direct-to-consumer PH-LLM deployment. Multimodal systems introduce additional governance challenges, including biometric-data protection, modality-specific failure modes, and cross-modal inference of sensitive attributes, and therefore require modality-specific validation and corresponding adjustments to consent, accountability, and monitoring structures [12,136]. Feasibility also varies across institutions, jurisdictions, and developer types because certification, audits, postdeployment surveillance, and incident reporting depend on unevenly distributed financial, technical, regulatory, and workforce capacity [100,101].

As the YouthPH-LLM case illustrates, youth-facing PH-LLMs expose an additional limitation of current digital privacy-law frameworks when applied to adolescent mental health support. Age-of-consent and parental authorization requirements, such as those under the General Data Protection Regulation [87], may protect minors' privacy and parental rights, but strict implementation can conflict with the ethical goal of providing confidential, low-friction access to adolescent mental health support. Routine parental disclosure or burdensome consent workflows may deter help-seeking, particularly among adolescents experiencing psychological distress, stigma, family conflict, or safety concerns [137]. Future governance should therefore distinguish ordinary data-processing consent from immediate safety-oriented support, preserving rapid access to crisis resources while developing jurisdiction-specific mechanisms for age-appropriate privacy, parental involvement, and safeguarding obligations.

The framework should balance underregulation, which may allow safety-critical failures to propagate at scale and shift risk onto vulnerable users, against overregulation, which may consolidate markets around large incumbents, reduce innovation for underrepresented groups, and divert resources from substantive safety work to compliance documentation. Because validated PH-LLM audit and certification cost data are not yet available, cost implications are described as qualitative categories: lower-cost safeguards include identity disclosure, scope labeling, basic adverse-event reporting, and crisis-resource signposting, whereas independent audits, red-team testing, certification preparation, continuous monitoring, and independent safety oversight are likely to impose moderate to high recurring costs. Large technology platforms may absorb these recurring costs, whereas startups, academic deployers, hospitals, and resource-constrained settings may face substantial barriers without proportionate pathways such as regulatory sandboxes, shared evaluation infrastructure, audit reciprocity, market-access incentives, and tiered certification with reduced fixed costs for smaller deployers [138,139].

We distinguish near-term feasible actions, such as identity disclosure and scope labeling, Tier 3 crisis exclusion from training data, and standardized adverse-event reporting, from longer-term aspirational elements, such as harmonized certification reciprocity, sustainably funded independent safety advisory boards, and population-level longitudinal outcomes monitoring. Qualitative cost categories, institutional capacity profiles, and implementation trade-offs are detailed in Multimedia Appendix 8.

Current Evidence Base and Limitations

The evidence base for ethical and governance concerns in PH-LLMs remains heterogeneous. Established empirical findings include hallucinations and clinically significant omissions in health-related LLM tasks, persistent demographic bias, privacy and confidentiality concerns in direct-to-consumer digital health settings, and failures of some conversational systems to manage crisis-related interactions safely [21,28,38]. Automation bias in PH-LLM contexts is supported mainly by adjacent-domain evidence from clinical decision-support systems, with limited direct PH-LLM evidence [32]. Emerging signals include increasing user reliance on PH-LLMs for health-related guidance, emotionally dependent use patterns, early indicators of therapeutic displacement, and the influence of interface design and deployment context on safety, trust, and care-seeking behavior [53,56-59]. Forward-looking risks include population-level safety effects, large-scale longitudinal profiling, governance challenges in multimodal and open-source

deployments, and health-related use of general-purpose LLMs, where evidence remains limited and many safeguards have not yet been prospectively validated. This framework should therefore be understood as evidence-informed but partly anticipatory, identifying plausible governance needs where empirical certainty remains uneven.

Conclusion

PH-LLMs are rapidly emerging as widely used interfaces for health sensemaking and self-management. Their direct-to-consumer deployment introduces a distinct ethical risk profile spanning 6 domains: privacy, accuracy, equity, transparency, human-AI interaction, and regulatory governance. Because failures across these domains may propagate harm at scale, governance for PH-LLMs should be not only principle-based but also operational and life cycle-oriented. We therefore propose a principlism-grounded framework that translates beneficence, nonmaleficence, autonomy, and justice into deployable safeguards across developers, deployers, health care institutions, regulators, and users. This framework is intended as an evidence-informed but partly anticipatory governance approach rather than an empirically validated solution set. Ongoing oversight will likely require continuous monitoring, adverse-event reporting, and, where appropriate, independent safety evaluation. Future work should prospectively examine the feasibility, effects, and implementation trade-offs of these mechanisms, including whether privacy-preserving forms of health-context activation could support limited adaptation of selected framework components to general-purpose LLMs used for health-related queries.

Acknowledgments

Following initial drafting, the authors used ChatGPT (OpenAI) for language editing to improve clarity. All artificial intelligence–generated suggestions were critically reviewed and revised as needed. The authors assume full responsibility for the final content and integrity of the manuscript.

Data Availability

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this article.

Authors' Contributions

JL and SL contributed to the conceptualization and design of the viewpoint. JL and SL conducted the literature analysis and drafted the original manuscript. All authors critically revised the manuscript for important intellectual content, approved the final version, and agreed to be accountable for all aspects of the work.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Protected-attribute handling for fairness monitoring.
[DOCX File , 21 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Developer lifecycle responsibilities and adverse-event reporting.
[DOCX File , 18 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Standards and institutional anchors for risk-tiered certification.
[DOCX File , 20 KB-Multimedia Appendix 3]

Multimedia Appendix 4

YouthPH-LLM case: full stakeholder specification.
[DOCX File , 20 KB-Multimedia Appendix 4]

Multimedia Appendix 5

Illustrative retention windows for tiered consent architecture.

[DOCX File , 19 KB-Multimedia Appendix 5]

Multimedia Appendix 6

Illustrative scope boundaries: permitted versus clinician-dependent functions.
[DOCX File , 20 KB-Multimedia Appendix 6]

Multimedia Appendix 7

Adverse-event severity categories and response pathways.
[DOCX File , 18 KB-Multimedia Appendix 7]

Multimedia Appendix 8

Feasibility, cost categories, and implementation trade-offs.
[DOCX File , 22 KB-Multimedia Appendix 8]

References

1. Khasentino J, Belyaeva A, Liu X, Yang Z, Furlotte NA, Lee C, et al. A personal health large language model for sleep and fitness coaching. Nat Med. 2025;31(10):3394-3403. [doi: 10.1038/s41591-025-03888-0] [Medline: 40813712]
2. Aydin S, Karabacak M, Vlachos V, Margetis K. Large language models in patient education: a scoping review of applications in medicine. Front Med (Lausanne). 2024;11:1477898. [FREE Full text] [doi: 10.3389/fmed.2024.1477898] [Medline: 39534227]
3. Huo B, Boyle A, Marfo N, Tangamornsuksan W, Steen JP, McKechnie T, et al. Large language models for chatbot health advice studies: a systematic review. JAMA Netw Open. 2025;8(2):e2457879. [FREE Full text] [doi: 10.1001/jamanetworkopen.2024.57879] [Medline: 39903463]
4. Improving service access and quality. World Health Organization. URL: https://www.who.int/activities/improving-service-access-and-quality [accessed 2026-01-20]
5. Li J, Li Y, Hu Y, Ma DCF, Mei X, Chan EA, et al. Chatbot-delivered interventions for improving mental health among young people: a systematic review and meta-analysis. Worldviews Evid Based Nurs. 2025;22(4):e70059. [doi: 10.1111/wvn.70059] [Medline: 40662463]
6. WHO releases AI ethics and governance guidance for large multi-modal models. World Health Organization. 2024. URL: https://www.who.int/news/item/18-01-2024-who-releases-ai-ethics-and-governance-guidance-for-large-multi-modal-models [accessed 2026-01-20]
7. OpenAI. Introducing ChatGPT Health. OpenAI Blog. 2026. URL: https://openai.com/index/introducing-chatgpt-health [accessed 2026-01-20]
8. What is ChatGPT Health? OpenAI help center. OpenAI. 2026. URL: https://help.openai.com/en/articles/20001036-what-is-chatgpt-health [accessed 2026-01-20]
9. Anthropic. Advancing Claude in healthcare and the life sciences. Anthropic News. 2026. URL: https://www.anthropic.com/news/healthcare-life-sciences [accessed 2026-01-20]
10. Abramson A. Fitbit's personal health coach in public preview is here. Google (The Keyword). 2025. URL: https://blog.google/products-and-platforms/devices/fitbit/personal-health-coach-public-preview/ [accessed 2026-01-12]
11. Wang C, Liu S, Yang H, Guo J, Wu Y, Liu J. Ethical considerations of using ChatGPT in health care. J Med Internet Res. 2023;25:e48009. [FREE Full text] [doi: 10.2196/48009] [Medline: 37566454]
12. Ethics and governance of artificial intelligence for health: guidance on large multi-modal models. World Health Organization. Geneva. WHO; 2025. URL: https://www.who.int/publications/i/item/9789240084759 [accessed 2026-01-10]
13. What is special category data? UK GDPR guidance. Information Commissioner's Office. 2024. URL: https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/lawful-basis/special-category-data/what-is-special-category-data/ [accessed 2026-01-10]
14. Kwesi J, Cao J, Manchanda R, Emami-Naeini P. Exploring user security and privacy attitudes and concerns toward the use of general-purpose LLM chatbots for mental health. 2025. Presented at: SEC '25: Proceedings of the 34th USENIX Conference on Security Symposium; 2025 August 13 - 15:6007-6024; Seattle WA USA.
15. Papneja H, Yadav N. Self-disclosure to conversational AI: A literature review, emergent framework, and directions for future research. Pers Ubiquit Comput. 2024;29(2):119-151. [doi: 10.1007/s00779-024-01823-7]
16. Marks M, Haupt CE. AI chatbots, health privacy, and challenges to HIPAA compliance. JAMA. 2023;330(4):309-310. [doi: 10.1001/jama.2023.9458] [Medline: 37410450]
17. King J, Klyman K, Capstick E, Saade T, Hsieh V. User privacy and large language models: an analysis of frontier developers' privacy policies. AIES. 2025;8(2):1465-1477. [doi: 10.1609/aies.v8i2.36646]
18. OWASP Top 10 for large language model applications 2025. OWASP Foundation. 2024. URL: https://owasp.org/www-project-top-10-for-large-language-model-applications/ [accessed 2026-01-15]

19. Doherty C, Baldwin M, Lambe R, Altini M, Caulfield B. Privacy in consumer wearable technologies: a living systematic analysis of data policies across leading manufacturers. NPJ Digit Med. 2025;8(1):363. [FREE Full text] [doi: 10.1038/s41746-025-01757-1] [Medline: 40517175]
20. Chikwetu L, Miao Y, Woldetensae MK, Bell D, Goldenholz DM, Dunn J. Does deidentification of data from wearable devices give us a false sense of security? A systematic review. Lancet Digit Health. 2023;5(4):e239-e247. [FREE Full text] [doi: 10.1016/S2589-7500(22)00234-5] [Medline: 36797124]
21. Asgari E, Montaña-Brown N, Dubois M, Khalil S, Balloch J, Yeung JA, et al. A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. NPJ Digit Med. 2025;8(1):274. [FREE Full text] [doi: 10.1038/s41746-025-01670-7] [Medline: 40360677]
22. Lee RW, Jun TJ, Lee J, Cho SI, Park HJ, Suh J. Vulnerability of large language models to prompt injection when providing medical advice. JAMA Netw Open. 2025;8(12):e2549963. [FREE Full text] [doi: 10.1001/jamanetworkopen.2025.49963] [Medline: 41632124]
23. Yao J, Zhou Z, Cui H, Ouyang Y, Han W. Trust transfer from medical AI to doctors and hospitals: integrating digital, AI, and scientific literacy in a cross-sectional framework. BMC Med Ethics. 2025;26(1):144. [FREE Full text] [doi: 10.1186/s12910-025-01300-7] [Medline: 41107871]
24. Keene Woods N, Ali U, Medina M, Reyes J, Chesser AK. Health literacy, health outcomes and equity: a trend analysis based on a population survey. J Prim Care Community Health. 2023;14:21501319231156132. [FREE Full text] [doi: 10.1177/21501319231156132] [Medline: 36852725]
25. McDonald M, Shenkman LJ. Health literacy and health outcomes of adults in the United States: implications for providers. IJAHSP. 2018;16(4):1-5. [doi: 10.46743/1540-580x/2018.1689]
26. Iftikhar Z, Xiao A, Ransom S, Huang J, Suresh H. How LLM counselors violate ethical standards in mental health practice: a practitioner-informed framework. AIES. 2025;8(2):1311-1323. [FREE Full text] [doi: 10.1609/aies.v8i2.36632]
27. Shen J, DiPaola D, Ali S, Sap M, Park HW, Breazeal C. Empathy toward artificial intelligence versus human experiences and the role of transparency in mental health and social support chatbot design: comparative study. JMIR Ment Health. 2024;11:e62679. [FREE Full text] [doi: 10.2196/62679] [Medline: 39321450]
28. Pichowicz W, Kotas M, Piotrowski P. Performance of mental health chatbot agents in detecting and managing suicidal ideation. Sci Rep. 2025;15(1):31652. [FREE Full text] [doi: 10.1038/s41598-025-17242-4] [Medline: 40866537]
29. Duffy C. Character.AI and Google agree to settle lawsuits over teen mental health harms and suicides. CNN Business. URL: https://www.cnn.com/2026/01/07/business/character-ai-google-settle-teen-suicide-lawsuit [accessed 2026-01-15]
30. Carnat I. Human, all too human: accounting for automation bias in generative large language models. Int Data Priv Law. 2024;14(4):299-314. [doi: 10.1093/idpl/ipae018]
31. Shekar S, Pataranutaporn P, Sarabu C, Cecchi GA, Maes P. People overtrust AI-generated medical advice despite low accuracy. NEJM AI. 2025;2(6):AIoa2300015. [doi: 10.1056/aioa2300015]
32. Khera R, Simon MA, Ross JS. Automation bias and assistive AI: risk of harm from AI-driven clinical decision support. JAMA. 2023;330(23):2255-2257. [doi: 10.1001/jama.2023.22557] [Medline: 38112824]
33. Wingerter TL, Straub T, Schweitzer S. Mitigating automation bias in generative AI through nudges: a cognitive reflection test study. Procedia Comput Sci. 2025;270:2106-2114. [doi: 10.1016/j.procs.2025.09.331]
34. Al-Sibai N. ChatGPT Is telling people with psychiatric problems to go off their meds. Futurism. 2025. URL: https://futurism.com/chatgpt-mental-illness-medications [accessed 2026-01-20]
35. Hough J, Culley N, Erganian C, Alahdab F. Potential risks of GenAI on medical education. BMJ Evid Based Med. 2025;30(6):406-408. [doi: 10.1136/bmjebm-2025-114339] [Medline: 41326281]
36. Lewis G, Marston L, Duffy L, Freemantle N, Gilbody S, Hunter R, et al. Maintenance or discontinuation of antidepressants in primary care. N Engl J Med. 2021;385(14):1257-1267. [doi: 10.1056/NEJMoa2106356] [Medline: 34587384]
37. Rodriguez JA, Alsentzer E, Bates DW. Leveraging large language models to foster equity in healthcare. J Am Med Inform Assoc. 2024;31(9):2147-2150. [doi: 10.1093/jamia/ocae055] [Medline: 38511501]
38. Omar M, Sorin V, Agbareia R, Apakama DU, Soroush A, Sakhuja A, et al. Evaluating and addressing demographic disparities in medical large language models: a systematic review. Int J Equity Health. 2025;24(1):57. [FREE Full text] [doi: 10.1186/s12939-025-02419-0] [Medline: 40011901]
39. Pfohl SR, Cole-Lewis H, Sayres R, Neal D, Asiedu M, Dieng A, et al. A toolbox for surfacing health equity harms and biases in large language models. Nat Med. 2024;30(12):3590-3600. [doi: 10.1038/s41591-024-03258-2] [Medline: 39313595]
40. Hanna JJ, Wakene AD, Johnson AO, Lehmann CU, Medford RJ. Assessing racial and ethnic bias in text generation by large language models for health care-related tasks: cross-sectional study. J Med Internet Res. 2025;27:e57257. [FREE Full text] [doi: 10.2196/57257] [Medline: 40080818]
41. 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 (Lond). 2024;4(1):176. [FREE Full text] [doi: 10.1038/s43856-024-00601-z] [Medline: 39256622]
42. van Kessel R, Seghers L, Anderson M, Schutte N, Monti G, Haig M, et al. A scoping review and expert consensus on digital determinants of health. Bull World Health Organ. 2025;103(2):110-125H. [doi: 10.2471/BLT.24.292057] [Medline: 39882497]

43. Ong JCL, Seng BJJ, Law JZF, Low LL, Kwa ALH, Giacomini KM, et al. Artificial intelligence, ChatGPT, and other large language models for social determinants of health: current state and future directions. Cell Rep Med. 2024;5(1):101356. [FREE Full text] [doi: 10.1016/j.xcrm.2023.101356] [Medline: 38232690]
44. Davies AR, Honeyman M, Gann B. Addressing the digital inverse care law in the time of COVID-19: potential for digital technology to exacerbate or mitigate health inequalities. J Med Internet Res. 2021;23(4):e21726. [FREE Full text] [doi: 10.2196/21726] [Medline: 33735096]
45. Hart JT. The inverse care law. Lancet. 1971;1(7696):405-412. [FREE Full text] [doi: 10.1016/s0140-6736(71)92410-x] [Medline: 4100731]
46. Mesinovic M, Watkinson P, Zhu T. Explainability in the age of large language models for healthcare. Commun Eng. 2025;4(1):128. [FREE Full text] [doi: 10.1038/s44172-025-00453-y] [Medline: 40676176]
47. Savage T, Nayak A, Gallo R, Rangan E, Chen JH. Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine. NPJ Digit Med. 2024;7(1):20. [FREE Full text] [doi: 10.1038/s41746-024-01010-1] [Medline: 38267608]
48. Shumway DO, Hartman HJ. Medical malpractice liability in large language model artificial intelligence: legal review and policy recommendations. J Osteopath Med. 2024;124(7):287-290. [FREE Full text] [doi: 10.1515/jom-2023-0229] [Medline: 38295300]
49. Mennella C, Maniscalco U, De Pietro G, Esposito M. Ethical and regulatory challenges of AI technologies in healthcare: a narrative review. Heliyon. 2024;10(4):e26297. [FREE Full text] [doi: 10.1016/j.heliyon.2024.e26297] [Medline: 38384518]
50. Noto LDG, Bezerra LCT. Can there be responsible AI without AI liability? Incentivizing generative AI safety through ex-post tort liability under the EU AI liability directive. Int J Law Inf Technol. 2024;32:eaae021. [doi: 10.1093/ijlit/eaae021]
51. Peter S, Riemer K, West JD. The benefits and dangers of anthropomorphic conversational agents. Proc Natl Acad Sci U S A. 2025;122(22):e2415898122. [FREE Full text] [doi: 10.1073/pnas.2415898122] [Medline: 40378006]
52. Sobowale K, Humphrey DK, Zhao SY. Evaluating generative AI psychotherapy chatbots used by youth: cross-sectional study. JMIR Ment Health. 2025;12:e79838. [FREE Full text] [doi: 10.2196/79838] [Medline: 41370787]
53. Laestadius L, Bishop A, Gonzalez M, Illenčík D, Campos-Castillo C. Too human and not human enough: a grounded theory analysis of mental health harms from emotional dependence on the social chatbot Replika. New Media Soc. 2022;26(10):5923-5941. [doi: 10.1177/14614448221142007]
54. Hua Y, Siddals S, Ma Z, Galatzer-Levy I, Xia W, Hau C, et al. Charting the evolution of artificial intelligence mental health chatbots from rule-based systems to large language models: a systematic review. World Psychiatry. 2025;24(3):383-394. [FREE Full text] [doi: 10.1002/wps.21352] [Medline: 40948070]
55. Zhang K, Xie Y, Chen D, Ji Z, Wang J. Effects of attractions and social attributes on peoples' usage intention and media dependence towards chatbot: the mediating role of parasocial interaction and emotional support. BMC Psychol. 2025;13(1):986. [FREE Full text] [doi: 10.1186/s40359-025-03284-w] [Medline: 40883845]
56. Talk, trust, and trade-offs: how and why teens use AI companions. Common Sense Media. 2025. URL: https://www.commonsensemedia.org/sites/default/files/research/report/talk-trust-and-trade-offs_2025_web.pdf [accessed 2026-01-15]
57. Faverio M, Sidoti O. Teens, social media and AI chatbots 2025. Pew Research Center. 2025. URL: https://www.pewresearch.org/internet/2025/12/09/teens-social-media-and-ai-chatbots-2025/ [accessed 2026-01-15]
58. Ischen C, Butler J, Ohme J. Chatting about the unaccepted: self-disclosure of unaccepted news exposure behaviour to a chatbot. Behav Inf Technol. 2023;43(10):2044-2056. [doi: 10.1080/0144929x.2023.2237605]
59. Skjuve M, Følstad A, Brandtzæg PB. A longitudinal study of self-disclosure in human–chatbot relationships. Interact Comput. 2023;35(1):24-39. [doi: 10.1093/iwc/iwad022]
60. Catapan SDC, Sazon H, Zheng S, Gallegos-Rejas V, Mendis R, Santiago PHR, et al. A systematic review of consumers' and healthcare professionals' trust in digital healthcare. NPJ Digit Med. 2025;8(1):115. [FREE Full text] [doi: 10.1038/s41746-025-01510-8] [Medline: 39984678]
61. Birkhäuer J, Gaab J, Kossowsky J, Hasler S, Krummenacher P, Werner C, et al. Trust in the health care professional and health outcome: a meta-analysis. PLoS One. 2017;12(2):e0170988. [FREE Full text] [doi: 10.1371/journal.pone.0170988] [Medline: 28170443]
62. Beatty C, Malik T, Meheli S, Sinha C. Evaluating the therapeutic alliance with a free-text CBT conversational agent (Wysa): a mixed-methods study. Front Digit Health. 2022;4:847991. [FREE Full text] [doi: 10.3389/fdgth.2022.847991] [Medline: 35480848]
63. Martinengo L, Lin X, Jabir AI, Kowatsch T, Atun R, Car J, et al. Conversational agents in health care: expert interviews to inform the definition, classification, and conceptual framework. J Med Internet Res. 2023;25:e50767. [FREE Full text] [doi: 10.2196/50767] [Medline: 37910153]
64. Saenger JA, Hunger J, Boss A, Richter J. Delayed diagnosis of a transient ischemic attack caused by ChatGPT. Wien Klin Wochenschr. 2024;136(7-8):236-238. [FREE Full text] [doi: 10.1007/s00508-024-02329-1] [Medline: 38305909]
65. Riboli-Sasco E, El-Osta A, Alaa A, Webber I, Karki M, El Asmar ML, et al. Triage and diagnostic accuracy of online symptom checkers: systematic review. J Med Internet Res. 2023;25:e43803. [FREE Full text] [doi: 10.2196/43803] [Medline: 37266983]

66. Woodcock C, Mittelstadt B, Busbridge D, Blank G. The impact of explanations on layperson trust in artificial intelligence-driven symptom checker apps: experimental study. J Med Internet Res. 2021;23(11):e29386. [FREE Full text] [doi: 10.2196/29386] [Medline: 34730544]
67. Wallace W, Chan C, Chidambaram S, Hanna L, Iqbal FM, Acharya A, et al. The diagnostic and triage accuracy of digital and online symptom checker tools: a systematic review. NPJ Digit Med. 2022;5(1):118. [FREE Full text] [doi: 10.1038/s41746-022-00667-w] [Medline: 35977992]
68. Tele-triage. US Department of Health and Human Services. 2025. URL: https://telehealth.hhs.gov/providers/best-practice-guides/telehealth-for-emergency-departments/tele-triage [accessed 2026-01-15]
69. Covered entities and business associates. US Department of Health and Human Services. 2024. URL: https://www.hhs.gov/hipaa/for-professionals/covered-entities/index.html [accessed 2026-01-15]
70. Tovino SA. Artificial intelligence and the HIPAA privacy rule: a primer. Hous J Health Law Policy. 2024;24:77-140. [FREE Full text]
71. Office for Civil Rights, U.S. Department of Health and Human Services. URL: https://www.hhs.gov/hipaa/for-professionals/privacy/guidance/access-right-health-apps-apis/ [accessed 2026-01-15]
72. Savage M, Savage LC. Doctors routinely share health data electronically under HIPAA, and sharing with patients and patients' third-party health apps is consistent: interoperability and privacy analysis. J Med Internet Res. 2020;22(9):e19818. [FREE Full text] [doi: 10.2196/19818] [Medline: 32876582]
73. FTC to ban betterHelp from revealing consumers' data, including sensitive mental health information, to Facebook and others for targeted advertising. Federal Trade Commission. 2023. URL: https://www.ftc.gov/news-events/news/press-releases/2023/03/ftc-ban-betterhelp-revealing-consumers-data-including-sensitive-mental-health-information-facebook [accessed 2026-01-15]
74. Regulation (EU) 2024/1689 of the European parliament and of the council of 13 june 2024. European Parliament and Council of the European Union. URL: https://eur-lex.europa.eu/eli/reg/2024/1689/oj [accessed 2026-01-15]
75. Interim measures for administration of generative artificial intelligence services. State Council of the People's Republic of China. 2023. URL: https://english.www.gov.cn/archive/statecouncilgazette/202308/30/content_WS64ee8c0cc6d0868f4e8deeea.html [accessed 2026-01-12]
76. Framework act on the development of artificial intelligence and the creation of a foundation for trust. Korea Legislation Research Institute / Ministry of Government Legislation. 2025. URL: https://elaw.klri.re.kr/eng_service/lawView.do?hseq=71019&lang=ENG [accessed 2026-01-12]
77. Ministry of Internal Affairs and Communications. AI guidelines for business ver1.0. Ministry of Economy, Trade and Industry. Tokyo, Japan. Ministry of Economy, Trade and Industry; 2024. URL: https://www.meti.go.jp/english/press/2024/0419_002.html [accessed 2026-02-26]
78. Proposed model AI governance framework for generative AI. Infocomm Media Development Authority. Singapore. Infocomm Media Development Authority; 2024. URL: https://www.imda.gov.sg/resources/press-releases-factsheets-and-speeches/press-releases/2024/public-consult-model-ai-governance-framework-genai [accessed 2026-02-26]
79. Palaniappan K, Lin EYT, Vogel S. Global regulatory frameworks for the use of artificial intelligence (AI) in the healthcare services sector. Healthcare (Basel). 2024;12(5):562. [FREE Full text] [doi: 10.3390/healthcare12050562] [Medline: 38470673]
80. Aboy M, Minssen T, Vayena E. Navigating the EU AI Act: implications for regulated digital medical products. NPJ Digit Med. 2024;7(1):237. [FREE Full text] [doi: 10.1038/s41746-024-01232-3] [Medline: 39242831]
81. Ethics and governance of artificial intelligence for health: WHO guidance. World Health Organization. Geneva. World Health Organization; 2021. URL: https://www.who.int/publications/i/item/9789240029200 [accessed 2026-01-16]
82. Recommendation of the council on artificial intelligence. Organization for Economic Co-operation and Development. 2019. URL: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449 [accessed 2026-01-16]
83. Beauchamp TL, Childress JF. Principles of Biomedical Ethics. New York. Oxford University Press; 2019.
84. Bakker T, Klopotowska JE, Dongelmans DA, Eslami S, Vermeijden WJ, Hendriks S, et al. SIMPLIFY study group. The effect of computerised decision support alerts tailored to intensive care on the administration of high-risk drug combinations, and their monitoring: a cluster randomised stepped-wedge trial. Lancet. 2024;403(10425):439-449. [doi: 10.1016/S0140-6736(23)02465-0] [Medline: 38262430]
85. Amugongo LM, 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(6):e0000877. [FREE Full text] [doi: 10.1371/journal.pdig.0000877] [Medline: 40498738]
86. Brückner S, Dridi A, Deshmukh A, Kirsten T, Lauber-Rönsberg A, Riedel R, et al. A user-driven consent platform for health data sharing in digital health applications. NPJ Digit Med. 2025;8(1):699. [FREE Full text] [doi: 10.1038/s41746-025-02147-3] [Medline: 41298895]
87. Regulation (EU) 2016/679 of the European parliament and of the council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection

Regulation). European Parliament and Council of the European Union. URL: https://eur-lex.europa.eu/eli/reg/2016/679/oj [accessed 2026-04-30]
88. Alderman JE, Palmer J, Laws E, McCradden MD, Ordish J, Ghassemi M, et al. Tackling algorithmic bias and promoting transparency in health datasets: the STANDING together consensus recommendations. Lancet Digit Health. 2025;7(1):e64-e88. [FREE Full text] [doi: 10.1016/S2589-7500(24)00224-3] [Medline: 39701919]
89. Hatef E, Hudson Scholle S, Buckley B, Weiner J, Austin J. Development of an evidence- and consensus-based digital healthcare equity framework. JAMIA Open. 2024;7(4):ooae136. [doi: 10.1093/jamiaopen/ooae136] [Medline: 39553827]
90. Babic B, Glenn Cohen I, Stern AD, Li Y, Ouellet M. A general framework for governing marketed AI/ML medical devices. NPJ Digit Med. 2025;8(1):328. [FREE Full text] [doi: 10.1038/s41746-025-01717-9] [Medline: 40450160]
91. Liu M, Ning Y, Teixayavong S, Liu X, Mertens M, Shang Y, et al. A scoping review and evidence gap analysis of clinical AI fairness. NPJ Digit Med. 2025;8(1):360. [FREE Full text] [doi: 10.1038/s41746-025-01667-2] [Medline: 40517148]
92. Templin T, Fort S, Padmanabham P, Seshadri P, Rimal R, Oliva J, et al. Framework for bias evaluation in large language models in healthcare settings. NPJ Digit Med. 2025;8(1):414. [FREE Full text] [doi: 10.1038/s41746-025-01786-w] [Medline: 40624264]
93. Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, et al. DECIDE-AI expert group. Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. BMJ. 2022;377:e070904. [FREE Full text] [doi: 10.1136/bmj-2022-070904] [Medline: 35584845]
94. Health Canada, Medicines and Healthcare products Regulatory Agency. Good machine learning practice for medical device development: guiding principles. U.S. Food and Drug Administration. 2021. URL: https://www.fda.gov/media/153486/download [accessed 2026-01-15]
95. Artificial Intelligence Risk Management Framework (AI RMF 1.0). National Institute of Standards and Technology. 2023. URL: https://doi.org/10.6028/NIST.AI.100-1 [accessed 2026-01-15]
96. What is health literacy? Centers for Disease Control and Prevention. 2024. URL: https://www.cdc.gov/health-literacy/php/about/index.html [accessed 2026-01-15]
97. Clinical decision support software: guidance for industry and food and drug administration staff. US Food and Drug Administration. 2026. URL: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/clinical-decision-support-software [accessed 2026-02-10]
98. Bouguettaya A, Stuart EM, Aboujaoude E. Racial bias in AI-mediated psychiatric diagnosis and treatment: a qualitative comparison of four large language models. NPJ Digit Med. 2025;8(1):332. [FREE Full text] [doi: 10.1038/s41746-025-01746-4] [Medline: 40467886]
99. Use of online tracking technologies by HIPAA covered entities and business associates. US Department of Health and Human Services. 2024. URL: https://www.hhs.gov/hipaa/for-professionals/privacy/guidance/hipaa-online-tracking/index.html [accessed 2026-01-15]
100. Regulatory considerations on artificial intelligence for health. World Health Organization. Geneva, Switzerland. World Health Organization; 2023. URL: https://www.who.int/publications/i/item/9789240078871 [accessed 2026-03-10]
101. Rozenblit L, Price A, Solomonides A, Joseph AL, Koski E, Srivastava G, et al. Toward responsible AI governance: balancing multi-stakeholder perspectives on AI in healthcare. Int J Med Inform. 2025;203:106015. [FREE Full text] [doi: 10.1016/j.ijmedinf.2025.106015] [Medline: 40680319]
102. AI system evaluations: independent evaluations. Developing AI accountability: inputs and a deeper dive (AI accountability policy report series). National Telecommunications and Information Administration. 2024. URL: https://www.ntia.gov/issues/artificial-intelligence/ai-accountability-policy-report/developing-accountability-inputs-a-deeper-dive/ai-system-evaluations/independent-evaluations [accessed 2026-01-15]
103. Kiseleva A, Kotzinos D, De Hert P. Transparency of AI in healthcare as a multilayered system of accountabilities: between legal requirements and technical limitations. Front Artif Intell. 2022;5:879603. [FREE Full text] [doi: 10.3389/frai.2022.879603] [Medline: 35707765]
104. Sullivan HR, Schweikart SJ. Are current tort liability doctrines adequate for addressing injury caused by AI? AMA J Ethics. 2019;21(2):E160-E166. [FREE Full text] [doi: 10.1001/amajethics.2019.160] [Medline: 30794126]
105. Longpre S, Kapoor S, Klyman K, Ramaswami A, Bommasani R, Blili-Hamelin B. A safe harbor for AI evaluation and red teaming. arXiv. Preprint posted online on March 7, 2024. [doi: 10.48550/arXiv.2403.04893]
106. Davis S, Dorn C, Park D, Matheny M. Emerging algorithmic bias: fairness drift as the next dimension of model maintenance and sustainability. J Am Med Inform Assoc. 2025;32(5):845-854. [doi: 10.1093/jamia/ocaf039] [Medline: 40079820]
107. The importance of pharmacovigilance: safety monitoring of medicinal products. World Health Organization. Geneva. WHO; 2002. URL: https://iris.who.int/server/api/core/bitstreams/002b78d5-4dfc-433c-a518-f92ebdee8706/content [accessed 2026-01-15]
108. Marketing submission recommendations for a predetermined change control plan for artificial intelligence-enabled device software functions. US Food and Drug Administration. Silver Spring, MD. US Food and Drug Administration; 2024. URL: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/marketing-submission-recommendations-predetermined-change-control-plan-artificial-intelligence [accessed 2026-03-10]

109. Council for International Organizations of Medical Sciences (CIOMS). Artificial Intelligence in Pharmacovigilance: CIOMS Working Group XIV Report. Geneva. CIOMS (Council for International Organizations of Medical Sciences); 2025.
110. Protecting the privacy and security of your health information when using your personal cell phone or tablet. US Department of Health and Human Services. 2022. URL: https://www.hhs.gov/hipaa/for-professionals/privacy/guidance/cell-phone-hipaa/index.html [accessed 2026-01-15]
111. Le-Khac UN, Truong VNX. A survey on large language models unlearning: taxonomy, evaluations, and future directions. Artif Intell Rev. 2025;58(12):399. [doi: 10.1007/s10462-025-11376-7]
112. 2025 national guidelines for a behavioral health coordinated system of crisis care. SAMHSA. 2025. URL: https://www.chcs.org/resource-center-item/2025-national-guidelines-for-a-behavioral-health-coordinated-system-of-crisis-care/ [accessed 2026-01-15]
113. Nahum-Shani I, Smith SN, Spring BJ, Collins LM, Witkiewitz K, Tewari A, et al. Just-in-time adaptive interventions (JITAIs) in mobile health: key components and design principles for ongoing health behavior support. Ann Behav Med. 2018;52(6):446-462. [FREE Full text] [doi: 10.1007/s12160-016-9830-8] [Medline: 27663578]
114. Zhang S, Li X. Differential privacy medical data publishing method based on attribute correlation. Sci Rep. 2022;12(1):15725. [FREE Full text] [doi: 10.1038/s41598-022-19544-3] [Medline: 36131115]
115. Garfinkel S, Guttman B, Near J, Dajani A, Singer P. De-Identifying Government Datasets: Techniques and Governance. Gaithersburg, MD. National Institute of Standards and Technology; 2023.
116. Farquhar S, Kossen J, Kuhn L, Gal Y. Detecting hallucinations in large language models using semantic entropy. Nature. 2024;630(8017):625-630. [FREE Full text] [doi: 10.1038/s41586-024-07421-0] [Medline: 38898292]
117. Roustan D, Bastardot F. The clinicians' guide to large language models: a general perspective with a focus on hallucinations. Interact J Med Res. 2025;14:e59823. [FREE Full text] [doi: 10.2196/59823] [Medline: 39874574]
118. Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N. Retrieval-augmented generation for knowledge-intensive NLP tasks. Adv Neural Inf Process Syst. 2020;33:9459-9474.
119. Arowosegbe A, Oyelade T. Application of natural language processing (NLP) in detecting and preventing suicide ideation: a systematic review. Int J Environ Res Public Health. 2023;20(2):1514. [FREE Full text] [doi: 10.3390/ijerph20021514] [Medline: 36674270]
120. Pais C, Liu J, Voigt R, Gupta V, Wade E, Bayati M. Large language models for preventing medication direction errors in online pharmacies. Nat Med. 2024;30(6):1574-1582. [FREE Full text] [doi: 10.1038/s41591-024-02933-8] [Medline: 38664535]
121. National Library of Medicine. RxNorm. National Institutes of Health. URL: https://www.nlm.nih.gov/research/umls/rxnorm/index.html [accessed 2026-01-15]
122. FDA adverse event reporting system (FAERS). U.S. Food and Drug Administration. URL: https://www.fda.gov/drugs/drug-approvals-and-databases/fda-adverse-event-reporting-system-faers-database [accessed 2026-01-15]
123. Ong JCL, Jin L, Elangovan K, Lim GYS, Lim DYZ, Sng GGR, et al. Large language model as clinical decision support system augments medication safety in 16 clinical specialties. Cell Rep Med. 2025;6(10):102323. [doi: 10.1016/j.xcrm.2025.102323] [Medline: 40997804]
124. Zack T, Lehman E, Suzgun M, Rodriguez JA, Celi LA, Gichoya J, et al. Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study. Lancet Digit Health. 2024;6(1):e12-e22. [FREE Full text] [doi: 10.1016/S2589-7500(23)00225-X] [Medline: 38123252]
125. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019;366(6464):447-453. [FREE Full text] [doi: 10.1126/science.aax2342] [Medline: 31649194]
126. Chen IY, Szolovits P, Ghassemi M. Can AI help reduce disparities in general medical and mental health care? AMA J Ethics. 2019;21(2):E167-E179. [FREE Full text] [doi: 10.1001/amajethics.2019.167] [Medline: 30794127]
127. Chin MH, Afsar-Manesh N, Bierman AS, Chang C, Colón-Rodríguez CJ, Dullabh P, et al. Guiding principles to address the impact of algorithm bias on racial and ethnic disparities in health and health care. JAMA Netw Open. 2023;6(12):e2345050. [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.45050] [Medline: 38100101]
128. Naderbagi A, Loblay V, Zahed IUM, Ekambareshwar M, Poulsen A, Song YJC, et al. Cultural and contextual adaptation of digital health interventions: narrative review. J Med Internet Res. 2024;26:e55130. [FREE Full text] [doi: 10.2196/55130] [Medline: 38980719]
129. Web Content Accessibility Guidelines (WCAG) 2.2. World Wide Web Consortium (W3C). 2023. URL: https://www.w3.org/TR/WCAG22/ [accessed 2026-01-15]
130. Ayre J, Bonner C, Muscat DM, Cvejic E, Mac O, Mouwad D, et al. Online plain language tool and health information quality: a randomized clinical trial. JAMA Netw Open. 2024;7(10):e2437955. [FREE Full text] [doi: 10.1001/jamanetworkopen.2024.37955] [Medline: 39378036]
131. Shi R, Petrakaki D. User-AI intimacy in digital health. Soc Sci Med. 2026;389:118853. [FREE Full text] [doi: 10.1016/j.socscimed.2025.118853] [Medline: 41453777]
132. Kompa B, Snoek J, Beam AL. Second opinion needed: communicating uncertainty in medical machine learning. NPJ Digit Med. 2021;4(1):4. [FREE Full text] [doi: 10.1038/s41746-020-00367-3] [Medline: 33402680]

133. Boucher EM, Harake NR, Ward HE, Stoeckl SE, Vargas J, Minkel J, et al. Artificially intelligent chatbots in digital mental health interventions: a review. Expert Rev Med Devices. 2021;18(sup1):37-49. [FREE Full text] [doi: 10.1080/17434440.2021.2013200] [Medline: 34872429]
134. Artificial intelligence and machine learning (AI/ML)-enabled medical devices. U.S. Food and Drug Administration. Silver Spring, MD. FDA URL: https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices [accessed 2026-01-20]
135. Fossouo Tagne J, Yakob RA, Mcdonald R, Wickramasinghe N. Linking activity theory within user-centered design: novel framework to inform design and evaluation of adverse drug reaction reporting systems in pharmacy. JMIR Hum Factors. 2023;10:e43529. [FREE Full text] [doi: 10.2196/43529] [Medline: 36826985]
136. Huang SC, Jensen M, Yeung-Levy S, Lungren MP, Poon H, Chaudhari AS. A systematic review and implementation guidelines of multimodal foundation models in medical imaging. Res Sq Preprint. 2025. [doi: 10.21203/rs.3.rs-5537908/v1]
137. Chung R, Lee J, Hackell J, Alderman EM, COMMITTEE ON ADOLESCENCE. Confidentiality in the care of adolescents: policy statement. Pediatrics. 2024;153(5):e2024066326. [doi: 10.1542/peds.2024-066326] [Medline: 38646690]
138. OECD. Artificial intelligence and competitive dynamics in downstream markets. OECD Roundtables on Competition Policy Papers No. 331. Paris, France. OECD Publishing; 2025. URL: https://www.oecd.org/content/dam/oecd/en/publications/reports/2025/11/artificial-intelligence-and-competitive-dynamics-in-downstream-markets_c6e81d0e/ccf0624a-en.pdf [accessed 2026-03-10]
139. OECD. Regulatory sandboxes in artificial intelligence. OECD Digital Economy Papers No. 356. Paris, France. OECD Publishing; 2023. URL: https://www.oecd.org/content/dam/oecd/en/publications/reports/2023/07/regulatory-sandboxes-in-artificial-intelligence_a44aae4f/8f80a0e6-en.pdf [accessed 2026-03-10]

Abbreviations

AI: artificial intelligence
LLM: large language model
PH-LLM: personal health large language model

Edited by A Coristine; submitted 27.Jan.2026; peer-reviewed by L Mauri, P Lacroix, A El Ammari; comments to author 26.Feb.2026; revised version received 27.May.2026; accepted 29.May.2026; published 17.Jun.2026

Please cite as:
Liu J, Liu S
Ethical Considerations in Personal Health Large Language Models
J Med Internet Res 2026;28:e92240
URL: https://www.jmir.org/2026/1/e92240
doi: 10.2196/92240
PMID: