Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Scoping review of governance, privacy, and crisis safety in AI-supported conversational agents for perinatal mental health

Desmond O. Obe

desmondobinna.obe@my.jcu.edu.au

James Cook University

Samantha J. Teague

James Cook University

Ickjai Lee

James Cook University

Adrian B. R. Shatte

James Cook University

Research Article

Keywords: Perinatal mental health, Conversational agents, Large language models, Clinical AI governance, privacy-preserving, federated learning, crisis escalation

Posted Date: November 11th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11506514/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: Competing interest reported. S.J.T. and A.B.R.S. are co-authors of two included studies evaluating the same system (Studies 27 and 35). To manage this potential conflict, D.O.O. independently assessed eligibility, extracted the data and assigned the Governance Maturity Scores for these studies using predefined criteria, and both studies were excluded from the blinded inter-rater re-scoring subset. The remaining authors declare no competing interests.

Scoping review of governance, privacy, and crisis safety in AI-supported conversational agents for perinatal mental health

Desmond O. Obe[1,*], Samantha J. Teague[2,3,4], Ickjai Lee[1], Adrian B. R. Shatte[3,5]

[1] Department of Information Technology, College of Science and Engineering, James Cook University, Cairns, Australia

[2] Department of Psychology, College of Healthcare Sciences, James Cook University, Townsville, Australia

[3] Margaret Roderick Centre for Mental Health Research, James Cook University, Townsville, Queensland, Australia

[4] Deakin Lifespan Institute, School of Psychology, Faculty of Health, Deakin University, Burwood, Victoria, Australia

[5] Department of Information Technology, College of Science and Engineering, James Cook University, Townsville, Australia

*Corresponding author: Desmond O. Obe; desmondobinna.obe@my.jcu.edu.au

ORCID iDs: Desmond O. Obe 0000-0003-2903-6085; Samantha J. Teague 0000-0002-0487-7307; Ickjai Lee 0000-0002-6886-6201; Adrian B. R. Shatte 0000-0002-6225-9697

Acknowledgements: This work forms part of the first author's PhD research at James Cook University. We thank Benton Lewis, Liaison Librarian, James Cook University, for advice on the search strategy.

Abstract

Purpose: Perinatal mental health conditions affect nearly 20% of mothers and 10% of fathers, yet access to timely care remains limited. AI-supported conversational agents (CAs) offer scalable support, but their governance remains poorly characterised. This scoping review examined how perinatal CAs are built, governed and evaluated, whom they serve, and what safeguards they incorporate.

Methods: Following JBI guidance, PRISMA-ScR, and a registered protocol, we searched nine interdisciplinary databases (January 2015–September 2025) and tracked citations. Two reviewers independently screened records, and findings were charted and summarised descriptively.

Results: Of 6,118 records, 35 studies describing 30 unique CAs across 14 countries were included. Rule-based architectures predominated; large language model and hybrid systems emerged rapidly after 2022. Twenty-two studies described screening, treatment, or crisis-related functions, although therapeutic foundations were inconsistently reported. None reported advanced privacy-preserving computation. Fourteen studies sent conversational content to commercial application programming interfaces (APIs), none reporting data jurisdiction or retention. Crisis safeguards narrowed progressively: 18 studies reported crisis provision, ten automated risk detection, and three linked detection to human follow-up. Only one specified a response timeframe; none reported crisis-detection sensitivity. Governance reporting was limited (median exploratory Governance Maturity Score, 7/24), with monitoring and post-deployment updating least documented. Fathers, couples, non-birthing partners, linguistically diverse populations, and low-connectivity users were poorly represented. No system was developed in a low-income country.

Conclusions: We propose eight minimum standards to guide developers, clinicians, and policymakers towards governance-first development. These prioritise validated crisis-response pathways, transparent data handling, inclusive design, human oversight, and continuous post-deployment monitoring.

Keywords: Perinatal mental health; Conversational agents; Large language models; Clinical AI governance; privacy-preserving; federated learning; crisis escalation.

Introduction

Perinatal mental health (PMH) conditions, such as depression and anxiety arising during pregnancy and the first postpartum year, are important contributors to morbidity associated with childbearing[1]. They affect about one in five mothers[2, 3] and one in ten fathers[4, 5] worldwide, with evidence of increasing prevalence over time[1, 6], yet remain substantially underdiagnosed and undertreated[1, 7]. Consequences extend to parent-infant bonding[8], family functioning[9], and child development[10, 11]. In severe cases, perinatal mental disorders contribute to maternal suicide[12], which remains among the leading causes of in the first postpartum year[13]. The care gap is pronounced in low- and middle-income countries (LMICs)[14-16] and other low-resource settings, including regional, rural, and remote communities, where specialist perinatal services are scarce[17]. Workforce shortages, cost, geographic isolation, fragmented care pathways, and stigma further limit access to timely psychological support[18-20].

Digital mental health interventions, delivered through web and mobile applications, have emerged as a scalable response to this care gap, with adoption accelerating during the COVID-19 pandemic as healthcare systems expanded remote models of care[21]. Among these interventions, artificial intelligence (AI)-supported conversational agents (CAs), comprising chatbots and voice assistants, have attracted growing interest due to their capacity to provide flexible, interactive, and increasingly personalised support[22]. Through text, voice, or multimodal interfaces, they provide psychoeducation, symptom monitoring, behavioural support, and referral guidance[22, 23]. Their continuous availability, perceived anonymity, and low barriers to access may encourage engagement among perinatal parents who are unable or reluctant to seek conventional care[23]. These same features increase the importance of robust safeguards when systems operate without immediate clinical oversight.

CAs have evolved substantially over the past six decades[24]. Early rule-based systems such as ELIZA[25], introduced in 1966 to simulate a Rogerian psychotherapist, relied on predefined rules and scripted pattern matching. Contemporary systems increasingly use classical machine learning trained on task-specific data, large language models (LLMs) pretrained on large text corpora, and hybrid architectures that combine multiple computational approaches[24, 26]. These architectures differ in how responses are generated, their capacity to adapt to users and conversational context, and the predictability and transparency of their outputs[24]. Such advances have enabled more context-aware and personalised interactions that respond to changing needs across the perinatal period[27]. This broadens the application of CAs to pregnancy-related distress, postnatal depression, parenting confidence, and care navigation. These architectural differences also have implications for clinical oversight, privacy, governance, safety, and responsible use[28, 29].

These concerns are particularly consequential in perinatal care, where parents may disclose suicidality, intimate partner violence, substance use, and thoughts of infant harm[30]. Failures in data protection or clinical safeguards, such as privacy breaches, unsafe advice, or missed crisis signals, can harm both parent and infant[28]. Key data practices include the provenance of data used to develop models, where data are stored and processed, the involvement of third-party services, and how systems are updated after deployment. They also raise questions of

consent and accountability when users are unclear about how their data are retained or reused, or when model updates alter system behaviour[29]. Advanced privacy-preserving approaches have been applied in general mental health research[31]. These include federated learning (training models across decentralised data while keeping raw data local), differential privacy (limiting individual-data disclosure through calibrated noise), and on-device inference (generating outputs locally on the user's device). Their adoption in perinatal CAs remains unclear. Many digital systems instead rely on commercial cloud infrastructures with limited transparency on data storage, third-party access, and re-identification risk[32]. Because model updates and provider changes can alter system behaviour after evaluation, these questions extend beyond pre-deployment testing to continuous post-deployment monitoring, or algorithmovigilance.

Important questions also remain about who these systems serve. Paternal PMH is linked to maternal wellbeing[4] and child outcomes[33], and parents may also influence each other as an interdependent dyad[34]. Yet whether perinatal CAs include fathers/non-birthing partners or model parents dyadically has not been mapped. Evidence is also limited on whether these systems support diverse users in low-resource settings[35, 36], demonstrate real-world clinical effectiveness, or provide reliable crisis escalation and human follow-up. These dimensions are closely linked, as design choices intended to strengthen personalisation, privacy, inclusivity, or clinical safety may have consequences for the others. Such interactions may be overlooked when system design, governance, and population inclusion are considered in isolation.

Recent reviews have examined mental-health CAs across effectiveness and feasibility[37, 38], technological evolution from rule-based systems to LLMs[24], ChatGPT applications[39], generative-AI design[40], and safety[28]. Related reviews have addressed federated learning in mental-health research[41], mHealth evaluation[42], and healthcare AI governance[43, 44]. Within PMH, reviews have considered LMIC context[45], care cascades[46], emotional support[47], and CA intervention[27],[48],[49]. These strands have largely been examined separately, and to our knowledge, no prior review has jointly examined computational architecture, data practices, population inclusion, governance maturity, and design trade-offs in perinatal CAs. This matters because increasing technical capability may not be accompanied by safeguards and evidence proportionate to clinical risk.

We conducted a scoping review of AI-supported CAs for PMH, following JBI methodology[50] and PRISMA-ScR reporting guidance[51]. We asked four questions:

(i) How are computational approaches (rule-based, classical machine learning, LLM, and hybrid) implemented in perinatal mental health CAs, and what ethical and privacy principles are embedded or neglected in these systems?

(ii) How do data practices (training-data provenance, storage location, processing methods, model transparency, and updating strategies) shape privacy, consent, and accountability?

(iii) Which populations, contexts, and care needs are included or excluded in the design and evaluation of these systems; that is, whose care is left behind?

(iv) What design tensions emerge between computational efficiency, personalisation, privacy, inclusivity, and clinical safety?

To describe governance reporting across heterogeneous systems, we used a review-specific Governance Maturity Score; from the findings, we synthesised a taxonomy of recurrent design tensions and propose eight minimum safety standards. This review maps the current evidence, identifies key gaps, and provides practical guidance for developers, clinicians, and policymakers to support responsible development of perinatal CAs.

Methods

We conducted a scoping review following JBI methodology[52] and reported according to the PRISMA-ScR checklist[51] (Online Resource 1). The protocol was prospectively registered on the Open Science Framework (https://osf.io/63mpb)[53]. Post-registration amendments, including development of the Governance Maturity Score and partial dual extraction, are documented in the OSF amendment log and described below.

Search strategy

On 5 November 2025, we searched nine health, psychology, computing and engineering databases (MEDLINE, PsycINFO, Emcare, CINAHL, Cochrane CENTRAL, Scopus, Web of Science, IEEE Xplore and the ACM Digital Library) for studies published from January 2015 to September 2025. The strategy combined three concepts (perinatal period, mental health and AI-supported conversational agents), was developed with a specialist librarian (B.L.) and adapted for each database. Backward and forward citation tracking and grey-literature searches supplemented the database searches. Full strategies are provided in Online Resource 2.

Eligibility criteria

Eligibility follows the Population, Concept, Context framework[50]. The population comprised individuals during pregnancy to 12 months postpartum, their partners or both. Mixed-population studies were eligible only when perinatal findings were reported separately. The concept comprised AI-supported text, voice, or multimodal CAs enabling interactive natural-language dialogue and explicitly addressing perinatal mental health or wellbeing. The context included clinical, real-world, and development settings, with evaluations involving human participants, simulations, datasets or expert review. We included peer-reviewed primary studies, full conference papers, major institutional reports, and protocol papers reporting substantive methodological development. We excluded non-conversational systems, studies without an explicit perinatal mental-health focus, reviews, editorials, unreviewed preprints, commercial white papers, non-English publications, abstracts, and posters (Online Resource 2).

Study selection and data extraction

Records were deduplicated and screened in Rayyan. Two reviewers (D.O.O. and A.B.R.S.) independently screened titles and abstracts (agreement 87.3%) and full texts (89.5%), resolving disagreements by consensus. One report remained unavailable despite extensive searches and author contact. Studies 27 and 35, co-authored by A.B.R.S. and S.J.T., were assessed by the non-conflicted reviewer, D.O.O.

A 92-variable form covering eight domains (study context; system and clinical scope; population representation; safety and risk; conversational design and architecture; data governance, privacy and accessibility; evaluation evidence; and design tensions) was piloted on five studies (Online Resource 3, sheet 10). D.O.O. extracted all data, and A.B.R.S. independently verified a random 11% sample (four studies) with full agreement, a deviation from the planned full dual extraction. Crisis provision was classified by the highest documented escalation tier, from no mechanism to automated detection with timed human follow-up.

System-level characteristics were reported per unique CA (n = 30) and evaluation-specific characteristics, such as governance, crisis escalation, and study design, per study (N = 35). Systems were classified by the architecture generating the user-facing response (rule-based, classical machine learning, LLM, hybrid, unspecified AI, or not reported); where several publications described one system, classification used the fullest technical description.

Governance Maturity Score

The Governance Maturity Score (GMS) rated six dimensions of reported governance (privacy, safety, monitoring, regulation, bias and fairness, and consent) from 0 (absent) to 4 (comprehensive) against anchored criteria informed by health-AI governance frameworks[44]·[43], WHO guidance[54], the PMH literature[4, 12], and requirements for special-category health data[55] (Table 1). Totals (0–24) were classified descriptively as None (0–3), Basic (4–7), Moderate (8–11), Advanced (12–14), or Comprehensive (15–24). Although not planned in the protocol, the GMS served as an assessment of governance reporting quality, used to describe and band studies by how completely governance was reported. It is not a validated measure of system safety and was not used to test hypotheses. A.B.R.S. independently re-scored a random 12 of the 33 eligible studies (Studies 27 and 35 were excluded owing to co-authorship). Exact dimension-level agreement was 71% (51 of 72 ratings) and 96% differed by no more than one point; agreement on totals was high (intraclass correlation coefficient, two-way, absolute agreement, single rater [ICC(A,1)]: 0.95, 95% CI 0.82–0.98)[56]. Nine of 12 studies were assigned to the same band; the three discordant cases were one point apart at the Basic–Moderate threshold and were resolved by consensus without changing the original band (Online Resource 3, sheets 11 and 12).

Table 1: GMS: dimensions and anchored 0–4 scoring criteria.

Dimension	0 - Absent	1 - Minimal	2 - Partial	3 - Substantial	4 - Comprehensive
Privacy	No privacy measure reported	Generic statement of confidentiality only	Anonymisation or pseudonymisation at res	Encryption plus access control; data flows described	Privacy-preserving architecture (on-device, federated or differential privacy)
Safety	No safety mechanism	Static safety disclaimer only	Disclaimer plus hotline or crisis-service contact details	Crisis detection or human-in-the-loop escalation	Automated crisis detection with human follow-up within a defined timeframe

Dimension	0 - Absent	1 - Minimal	2 - Partial	3 - Substantial	4 - Comprehensive
Monitoring	No monitoring or oversight	Ad hoc developer review	Periodic human review of outputs	Defined monitoring process or clinical oversight	Continuous clinical review with audit trail
Regulation	No regulatory or ethics reference	Ethics approval mentioned	Ethics approval plus trial or data registration	Named data-protection regulation cited (for example, GDPR)	Regulatory pathway or device clearance documented
Bias and fairness	Not considered	Bias acknowledged as a limitation	Population characteristics reported	Subgroup performance described	Empirical fairness evaluation across demographic groups
Consent	No consent process	Implicit consent (for example, terms of service)	General informed consent obtained	Explicit, documented, study-specific consent	Explicit consent with data-use transparency and withdrawal

Data synthesis and analysis

Formal critical appraisal was not conducted, consistent with PRISMA-ScR[51] and JBI guidance[52]. Charted data were summarised with counts, proportions, and tabulations and synthesised narratively; design tensions were synthesised inductively using constant comparison. GMS totals and dimension scores were summarised descriptively (medians and interquartile ranges) by model class, crisis-escalation tier, and study design; no inferential comparisons between groups were performed, because the GMS was developed after protocol registration and several groups were small. Proposed minimum safety and governance standards were derived after charting by mapping the findings against the NIST AI Risk Management Framework [57], WHO guidance[54], FDA clinical decision support guidance[58], EU AI Act transparency requirement[59], GDPR requirements for special-category data[40, 60] and practices identified across the included studies. Each unique system was assessed against predefined criteria (Box 1).

Result

Study selection and characteristics

Our search identified 6,118 records. After removing 2,404 duplicates, 3,714 underwent title-and-abstract screening and 65 proceeded to full-text review. One report was unavailable, and 32 of 64 met the inclusion criteria. Citation tracking and Google Scholar identified three additional studies, yielding 35 included studies (Fig. 1). These described 30 unique CAs; *Parentbot, Woebot, ALBA*, and *Rover* were each evaluated in more than one

study. Twenty-eight studies (80%) were published in 2022–2025, and 11 (31%) were conference papers, all at conceptual or laboratory stage.

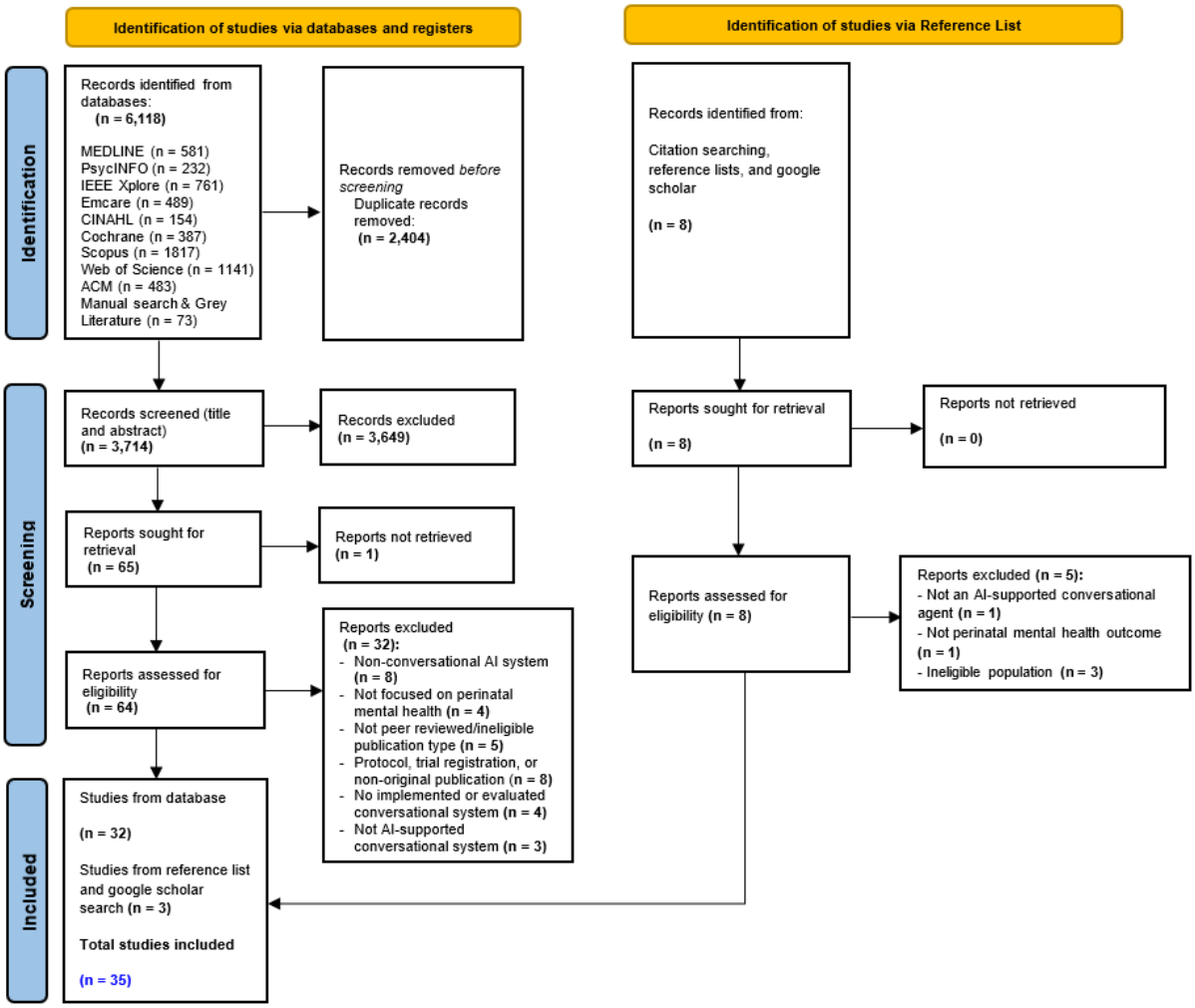


Fig. 1: PRISMA-ScR flow diagram of the study selection process.

The 30 unique systems were developed across 14 countries: 20 (67%) in high-income countries, four in upper-middle-income, and five in lower-middle-income countries, one through a USA–Kenya collaboration, and none in a low-income country (Fig. 2). Evaluation followed the same pattern: 19 of 35 studies (54%) were conducted in high-income countries, four in upper-middle-income and four in lower-middle-income settings, and eight did not report an evaluation country. Two systems were general-purpose commercial products evaluated for perinatal use. Of the 28 purpose-built systems, 12 (43%) were developed by computing or engineering teams, seven by interdisciplinary teams, five by commercial companies, three by health teams, and one by another academic discipline; computing teams accounted for 10 of the 11 conference studies. Clinical involvement in development or validation was reported in 23 of 35 studies (66%). Text was the dialogue modality in 32 studies (91%) and voice in two; 25 studies (71%) used a mobile application, two combined app and web interfaces, one used a web interface, one used SMS and a messaging platform, and six did not report the platform. Mixed-methods evaluations were the most common design (n = 11), followed by simulation studies (n = 8) and RCT (n = 5). Study samples ranged from two participants to an analytical dataset of 380,500 records (median human-user sample, 50

participants). Eleven studies (31%) enrolled no human end users, relying instead on simulation, secondary dataset analysis, or expert review.

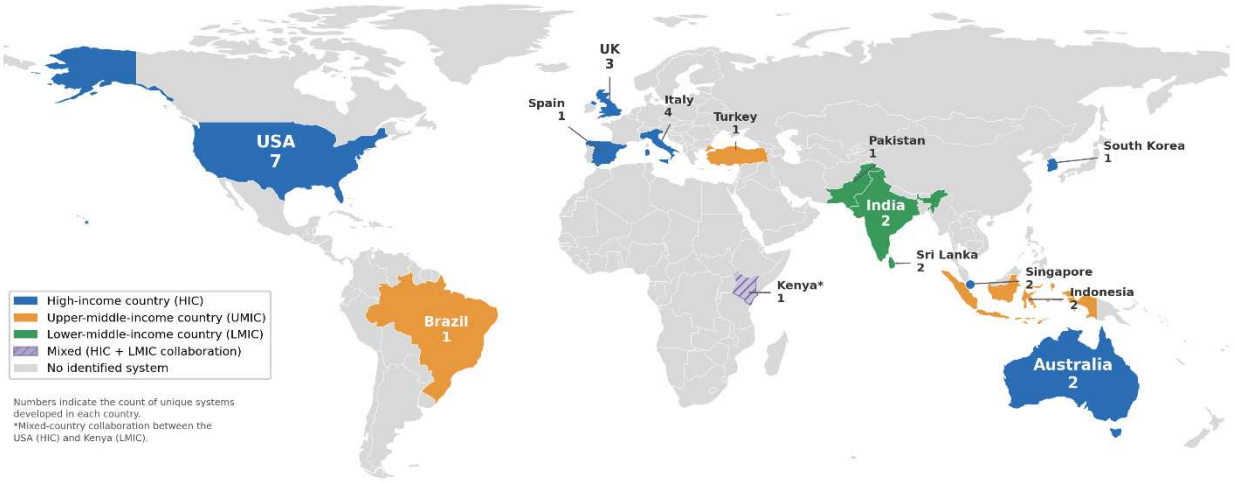


Fig. 2: Geographic distribution of perinatal CAs. The 30 unique systems were developed across 14 countries, classified by World Bank income group. Zuri was jointly developed in the USA and Kenya.

Computational approaches and embedded principles (question i)

Across the unique systems, rule-based architecture predominated (10 systems, 33%), followed by hybrid (7, 23%) and LLM systems (6, 20%); two used classical machine learning, three relied on commercial AI of undisclosed architecture, and two reported no model (Table 2). At study level, this yielded 15 rule-based, 7 hybrid, 6 LLM, 3 unspecified-AI, 2 classical machine learning, and 2 not-reported studies. All six LLM systems and five of seven hybrid systems appeared in 2022–2025 (Fig. 3). Two systems (MOMCare and DIAN) used retrieval-augmented generation.

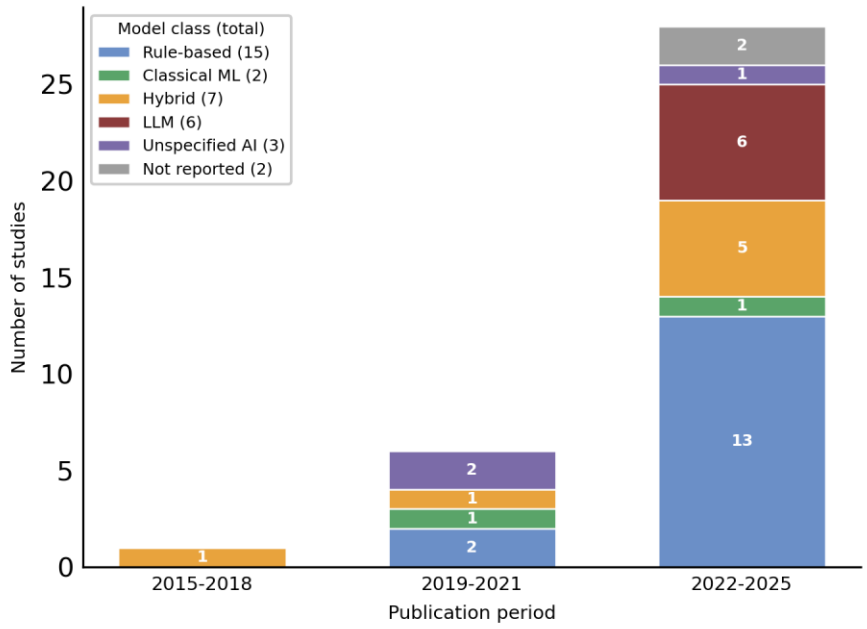


Fig. 3 Publications over time by model class (N = 35 studies)

Table 2: Model-class classification of the 30 unique CAs across 35 studies.

Model Class	**Unique Systems, *n*=30**	**Studies, *N*=35**	**Systems (study IDs; references)**
Rule-based	10 (33%)	5 (43%)	*Parentbot*(Studies 1, 17 and 34)[61-63]; *Woebot*(Studies 5 and 30)[64, 65]; *ALBA*(Studies 14 and 22)[66, 67]; *iPosyandu chatbot*(Study 8)[68]; *ClaimIt chatbot*(Study 23)[69]; *Juno*(Study 16)[70]; *Maia*(Study 26)[71]; *Rover*(Studies 27 and 35)[72, 73]; *Moment for Parents*(Study 28)[74]; *Pregnancy Companion*(Study 32)[75]
Hybrid	7 (23%)	7 (20%)	*PPD screener*(Study 4)[76]; *Baby Bump*(Study 13)[77]; *e-Midwife*(Study 15)[78]; *Dr. Joy*(Study 18)[79]; *Maria*(Study 19)[80]; *Rosie*(Study 29)[81]; *MamaBot*(Study 33)[82]
Large Language Model (LLM)	6 (20%)	6 (17%)	*MOMCare*(Study 6)[83]; *NurtureBot*(Study 7)[84]; *Breastfeeding chatbot*(Study 11)[85]; *DIAN*(Study 12)[86]; *ChatGPT intervention*(Study 20)[87]; *Storytelling LLM*(Study 24)[88]
Unspecific AI	3 (10%)	3 (9%)	*Zuri*(Study 2)[89]; *Voice assistants*(Study 3)[90]; *Wysa*(Study 31)[91]
Classical Machine Learning (ML)	2 (7%)	2 (6%)	*Chat robot*(Study 9)[92]; *SVM chatbot*(Study 10)[93]
Not reported	2 (7%)	2 (6%)	*AI chatbots*(Study 21)[94]; *HiBaby*(Study 25)[95]

Classification follows the architecture generating the user-facing response. Commercial voice assistants (Study 3) were classified as unspecified AI because their response-generating architectures were not disclosed; Woebot responses are rule-based although input processing uses machine-learning intent classification.

Depression was the most frequent target (21 studies, 60%), followed by wellbeing (16), anxiety (15), and stress (10); when classified into mutually exclusive categories, multi-condition support was most common (16 studies), followed by depression alone (10) (Fig. 4a). Informational support was the most common function (27 studies, 77%), followed by monitoring (22), treatment (16), screening (8), crisis detection or response (4), and decision support (1); 22 studies (63%) described at least one screening, treatment, or crisis function (Fig. 4b). No system performed a diagnostic function. Hybrid approaches and psychoeducation were the most common therapeutic foundations (nine systems each), while seven of 30 systems (23%) reported no explicit therapeutic foundation (Fig. 4c); interpersonal psychotherapy and acceptance and commitment therapy appeared only within hybrid systems.

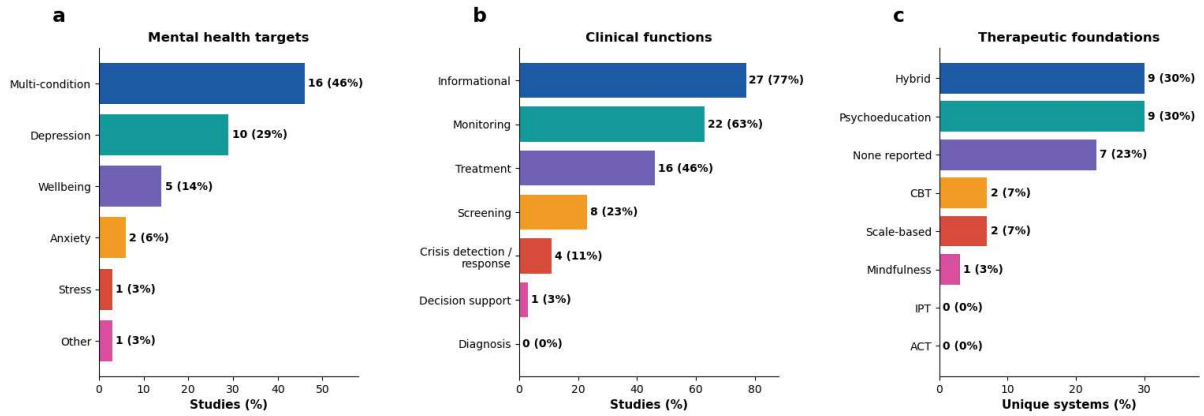


Fig. 4: Clinical scope, functions, and therapeutic foundations of perinatal CAs. a*, Mental health targets across 35 studies.* ***b****, Clinical functions across 35 studies.* ***c****, Therapeutic foundations across 30 unique systems*

Crisis provision narrowed at each step (Table 3; Fig. 5). Eighteen studies (51%) documented any crisis provision, ten studies describing nine systems used automated risk or emergency detection, three linked detection to a human response, and one specified a response timeframe: Rosie guaranteed follow-up by a clinical social worker within 24 hours. Of these three, two were specific to mental-health crises. Zuri alerted a live team member without an assured response time, and the Pregnancy Companion's emergency pathway was triggered by general medical emergencies rather than mental-health crisis language. No study reported the sensitivity, specificity, or false-negative rate of crisis detection; e-Midwife, for example, reported 97.9% accuracy for risk-keyword classification but not crisis-detection sensitivity. Ten studies (29%) documented no escalation mechanism and seven (20%) did not address crisis provision. None of the eight postnatal-only systems reported a human-involving crisis pathway.

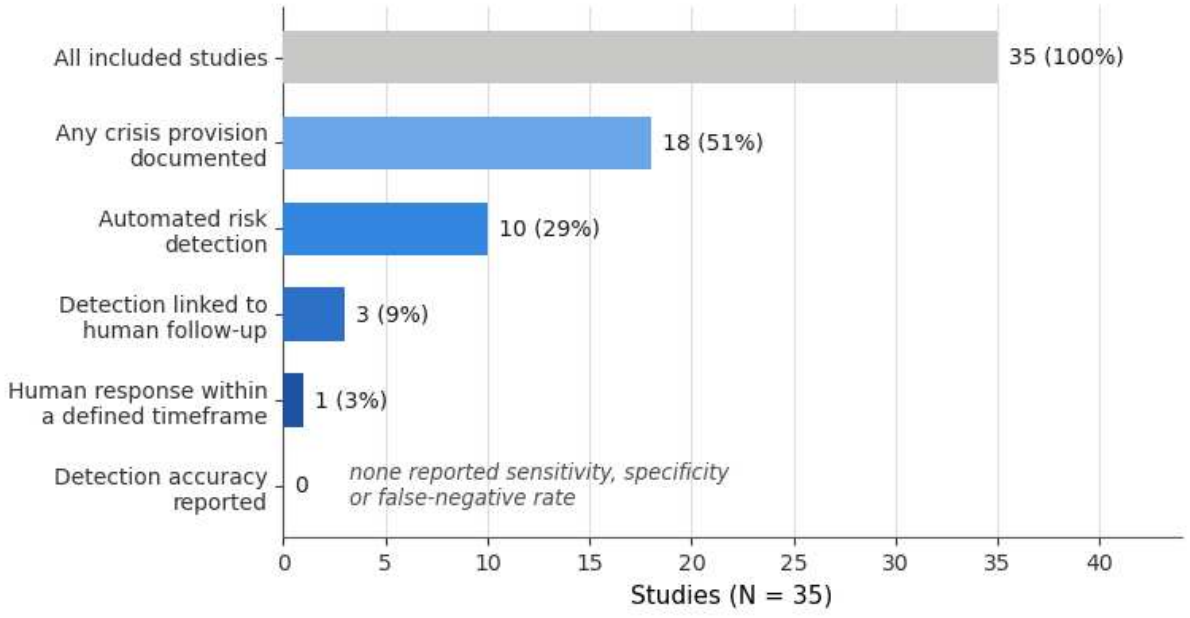


Fig. 5: Crisis provision across included studies. Each bar counts studies meeting that condition and all conditions above it; percentages are of 35 studies. No study reported the sensitivity, specificity, or false-negative rate of crisis detection

Table 3 Highest documented crisis-escalation tier (N = 35 studies)

Highest crisis-escalation tier	Studies	%
Automated detection + human follow-up	3	9%
Automated detection (no human)	4	11%
Human-in-the-loop (referral without automated detection)	6	17%
Hotline / emergency contacts	3	9%
Static disclaimer only	2	6%
None	10	29%
Not reported	7	20%
Total	35	100%

Categories are mutually exclusive. Human-in-the-loop refers to referral or transfer to a clinician or midwife without automated crisis detection.

Data practices (question ii)

No study reported advanced privacy-preserving computation such as federated learning, on-device inference, differential privacy, secure aggregation, or homomorphic encryption (Table 4). Anonymisation or pseudonymisation was reported in 13 studies (37%) and encryption in two. Fourteen studies (40%) depended on commercial application programming interfaces (APIs), including OpenAI, AWS, Kakao, Google Dialogflow, Telegram, and Twilio, routing conversational content through external servers. Half of these stated where data were stored, but none specified the jurisdiction governing the data or a retention policy; across all studies, no explicit retention policy was reported. Training-data or knowledge-base provenance was reported in 30 studies (86%), with clinical datasets underpinning four systems. Source code was openly available for one system and no study provided a model card or datasheet. Post-deployment change was assessed at two levels: model-level updating, involving retraining or learning from new data, and content or configuration updating without modifying the underlying model. Model-level updating was reported in three studies (two used continuous online learning and one iterative refinement), and content or knowledge-base updating in two. Eight studies described a static model or decision mechanism, and 21 (60%) did not report an updating strategy. The PPD screener proposed on-device processing and stronger encryption as future safeguards but, as implemented, transmitted user input to a commercial cloud API.

Table 4 Privacy-preserving measures, data practices, and transparency across included studies (N = 35)

Measure	Reported, n (%)	Details
Advanced privacy-preserving computation	0 (0%)	No federated learning, on-device inference, differential privacy, secure aggregation or homomorphic encryption reported

Measure	Reported, n (%)	Details
Anonymisation / pseudonymisation at rest	13 (37%)	–
Encryption reported	2 (6%)	–
Third-party commercial API dependency	14 (40%)	15 studies (43%) when authentication-only dependency in Study 26 is included
Jurisdiction governing data specified	1 (3%)	None among the 14 studies using commercial APIs
Explicit data-retention policy	0 (0%)	Two studies mentioned retention without a specified policy
Data-storage location stated	9 (26%)	Cloud (5); hybrid (2); local (1); unclear (1)
Open-source system code	1 (3%)	One further study shared analysis code and de-identified data; another used an open-source platform without releasing system-specific code
Named data-protection regulation	6 (17%)	GDPR (5); Australian Privacy Act 1988 (1)
Training-data or knowledge-base provenance	30 (86%)	Clinical datasets (4)
Model card / datasheet provided	0 (0%)	None reported
Post-deployment updating status reported	13 (37%)	Model-level updating: continuous online learning (2), iterative model refinement (1). Content or knowledge-base updating without model-level change (2). Static model or decision mechanism (8), of which two allowed content or configuration changes.

Third-party API dependency excludes the authentication-only digital-identity service used in Study 26.

Populations, contexts, and care needs (question iii)

Four of 30 systems addressed fathers (Parentbot, NurtureBot, Dr. Joy, and Rover), leaving 26 (87%) focused exclusively on mothers; the depth of paternal inclusion varied widely. Rover was designed specifically for fathers, with father-specific content and male-voiced guidance, and included fathers regardless of biological parenthood; Dr. Joy enrolled two male participants. Parentbot recruited couples but adjusted for clustering instead of modelling reciprocal partner effects, and NurtureBot analysed mothers and fathers independently; no study modelled dyadic interdependence. Parentbot restricted participation to heterosexual couples. Twenty-nine of 35 studies made no reference to non-birthing partners, sexual orientation, or gender identity. Nineteen systems (63%) operated exclusively in English, none supported multilingual conversation, and only Zuri supported access without mobile broadband, through SMS. Sixteen studies (46%) simultaneously focused only on mothers, operated only in English, and offered no offline capability. Rule-based studies came almost exclusively from high-income settings (14 of 15), whereas every system developed in a lower-middle-income country used a hybrid or LLM architecture. Cultural adaptation was reported in 15 studies (43%); two LLM systems delivered content in Turkish and Indonesian on base models trained predominantly on English data without reporting evaluation for linguistic or cultural bias. Stage coverage was balanced (15 systems spanned pregnancy and postpartum), but no system

targeted perinatal psychosis, obsessive-compulsive disorder, bipolar disorder, birth-related trauma, substance use, eating disorders, or perinatal loss.

Governance maturity

GMS totals ranged from 2 to 14 of 24 (median 7, IQR 5–10.5); 4 studies fell in the None band range, 16 in Basic, 10 in Moderate, and 5 in Advanced, and none reached Comprehensive (Fig. 6). Even the five highest-scoring studies (Zuri, the ALBA feasibility study, Rosie, Maia, and the Parentbot trial) reported no privacy-preserving architecture or empirical fairness evaluation. Mean dimension scores were highest for bias and fairness (2.00 of 4) and safety (1.57), and lowest for privacy (0.89) and monitoring (0.46) (Fig. 7). Clinical supervision was absent in 23 studies (66%) and no study reported continuous clinical review of conversational outputs during deployment. Explicit consent was documented in 18 studies (51%); two relied on implicit consent, three described no consent process, and 12 did not report one. Six studies cited a named data-protection regulation (the GDPR in five and the Australian Privacy Act 1988 in one), and six systems disclosed their AI identity in-chat.

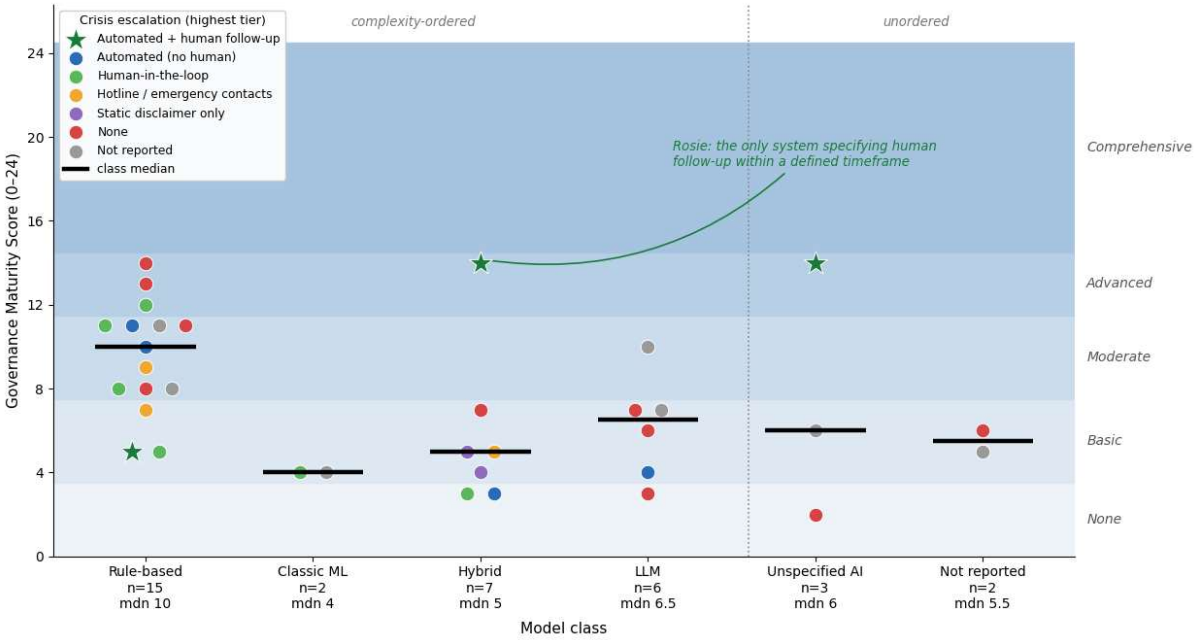


Fig. 6: Governance maturity by model class and crisis-escalation tier. Each marker represents one study, coloured by its highest documented crisis-escalation tier

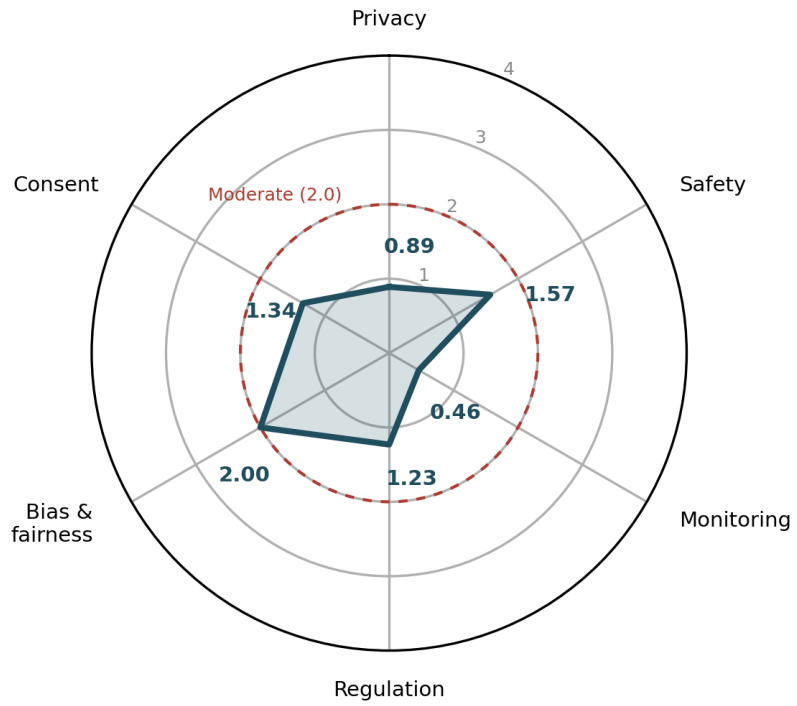


Fig. 7: Governance Maturity Profile across the six GMS dimensions. Mean scores across 35 studies; the dashed ring denotes a score of 2.0 on the 0–4 scale

By model class, rule-based studies recorded a median of 10 of 24 (IQR 8–11), compared with 6.5 for LLM studies, 5 for hybrid studies, and 4 for the two classical machine-learning studies, and the combined LLM-and-hybrid group had a median of 5 (IQR 4–7) (Fig. 6). These descriptive differences coincided with differences in study design: six of 13 LLM-and-hybrid studies were technical evaluations without end users, compared with two of 15 rule-based studies, and six were conference papers, compared with one. Within design strata, rule-based studies had higher medians among clinical pilots or later (11 versus 7; n = 9 and 6) and technical evaluations (6.5 versus 4; n = 2 and 6), but not among early-stage studies with end users, where the single LLM study scored 10 against a rule-based median of 9.5. At indicator level, rule-based studies more often documented a privacy measure at rest (10/15 versus 1/6 LLM studies), clinical supervision (6/15 versus 0/6), and a human-involving crisis pathway (5/15 versus 0/6), whereas LLM studies more often addressed bias or fairness (4/6 versus 9/15) and reported a data-storage location (3/6 versus 4/15), and AI-identity disclosure was more frequent in hybrid and LLM studies (Fig. 8). Reporting then differed by the kind of element rather than declining uniformly with newer architecture: declarative elements were documented at least as often in newer systems, while elements establishing operational accountability to a clinician were documented less often.

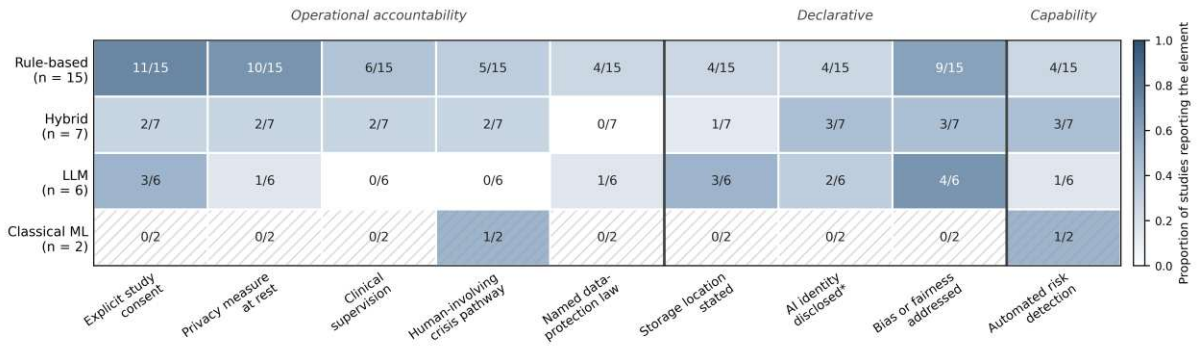


Fig. 8: Governance reporting by model class. AI-identity disclosure counts partial disclosure, where the system signals non-human status without an explicit statement that it is AI.

Evaluation evidence and design tensions (question iv)

Eight studies (23%) evaluated systems through simulation without real users and two in laboratory settings; 15 (43%) evaluated systems in the home or another real-world context, one in a clinical setting, and seven in mixed settings. Ten studies (29%), describing seven systems, reported clinical outcomes in real users; the remainder reported usability, feasibility, engagement, or technical performance. Five studies were primary reports of RCT. Where between-group effects were reported they were small, such as a reduction in maternal state anxiety at three months (mean difference −2.21, 95% CI −4.18 to −0.24; d = −0.22), and no study reported cost-effectiveness. The largest trial reported a small increase in parenting self-efficacy at one month (mean difference 1.22, 95% CI 0.06 to 2.39; d = 0.14), with no significant effect on depression or stress, and none of the simulation-only systems had a corresponding real-user evaluation. Six recurrent design tensions were identified across 28 studies (80%; Table 5; Online Resource 2). Safety versus autonomy was the most frequent, and the tensions spanned conversational architecture, data use, clinical oversight, accessibility, and population inclusion, indicating that the same design decision could create benefits in one dimension while introducing constraints in another.

Table 5 Six recurrent design tensions

Tension	**Recurring trade-off**
Safety versus autonomy	Constrained interaction is easier to control but less responsive; open-ended generation is more flexible but introduces greater safety uncertainty.
Scale versus personalisation	Broader reach may reduce tailoring; greater personalisation increases data, infrastructure, and computational demands
Clinical accuracy versus accessibility	Clinically detailed content may be harder to understand, while simplification can reduce clinical precision.
Privacy versus personalisation	Personalisation often requires more data, increasing the need for stronger privacy safeguards.
Automation versus clinical oversight	Automation supports scalability and reduces reliance on limited clinical capacity, but complicates human oversight and integration with health services when needs exceed the system's capabilities

Tension	Recurring trade-off
Father inclusion versus focused maternal care	Maternal-focused design excludes fathers and non-birthing partners; broader inclusion reduces father-specific tailoring and increases the support and escalation demand a system must meet.

Discussion

Critical synthesis

This review shows that the rapid evolution of AI-supported CAs for PMH has not been matched by equivalent progress in governance, privacy, safety, and inclusive design. Twenty-two studies described screening, treatment, or crisis-related functions, placing many systems in clinically consequential roles. Yet no study reported advanced privacy-preserving computation, none disclosed data jurisdiction or retention for conversations transmitted to commercial providers, and only one specified a timeframe for human follow-up after automated mental-health crisis detection. Fathers, couples, non-birthing partners, linguistically diverse populations, and users in low-connectivity settings remained poorly represented.

Despite these clinical functions, seven of 30 unique systems reported no explicit therapeutic foundation. Systems can influence care without making diagnoses. Screening may shape symptom recognition and help-seeking, therapeutic content may affect self-management, and responses to distress may influence subsequent care. Governance, evidence, and oversight should reflect intended function and consequences of failure, rather than conversational sophistication.

The crisis findings illustrate why oversight must reflect clinical risk. Maternal suicide is a leading cause of death in the year after birth[12, 13]. Perinatal women often withhold disclosure from clinicians due to fear of judgement or child-protection involvement[96], so a CA may receive disclosures that no clinician hears. Previous reviews recorded whether detection or referral features were present[28, 97], but tracing each step distinguishes feature availability from effective crisis response. A detector with an unknown false-negative rate may create an appearance of safety without establishing that crises are recognised. Guardrails or referral links do not establish that a person responds. Recent studies questioning the reliability of LLM responses to suicide-related disclosures reinforce this concern[98, 99].

Automated detection and escalation may strengthen safety in lower-risk or subclinical populations where no crisis pathway exists, provided appropriate safeguards are in place. For higher-risk populations, however, fully automated responses are harder to justify. Meaningful human involvement requires a defined escalation threshold, a qualified responder, a response timeframe, and a contingency when contact fails[29]. Crisis response should be evaluated as an end-to-end pathway in real users, assessing detection sensitivity, false-negative performance, completed escalation, and time to human response.

These safeguards require ongoing algorithmovigilance, or oversight of AI performance after deployment. Monitoring was the weakest governance dimension, no study reported continuous clinical review of outputs, and most did not report whether the model or knowledge base changed after deployment. Systems built on third-party

foundation models may change through provider updates outside the deployer's control, and behavioural variation between versions of the same model has been documented[100], so safety established during evaluation may not persist. Post-deployment monitoring of crisis handling, output safety, and data flows, with reassessment when the model, provider, or content changes, should be a core requirement.

Data practices revealed a further governance gap. Fourteen studies relied on third-party commercial APIs for conversational processing, while storage location was reported in only nine studies and encryption in two; training-data provenance was inconsistently documented, limiting assessment of whether data sources were appropriate for the populations served. These gaps do not establish inappropriate data handling, because the review assessed published reporting rather than deployed infrastructure, but they limit independent assessment of privacy, reproducibility, and accountability. Data practices raise particular risks in PMH, where conversations may involve suicidal ideation, intimate partner violence, or thoughts of infant harm[30]. Exposure of this content, or the perception that it may be exposed, carries two clinical risks: direct harm if disclosures reach third parties, including an abusive partner with access to a shared device[101], and indirect harm if parents censor what they disclose, so that the signals a system depends on to trigger escalation are never expressed. Anonymisation of stored records does not address content that crosses a network boundary to an external provider, as occurred in 14 studies. Published privacy statements are also an unreliable guide: in an audit of depression and smoking-cessation apps, 33 of 36 transmitted data to third parties, often without disclosure[32], and accreditation has failed to detect unencrypted transmission[102]. Privacy-preserving computation has demonstrated feasibility in related mental-health applications[41, 103]. Until data flows are verified, clinicians should discuss the limits of confidentiality with parents before recommending these tools, with particular caution where intimate partner violence or child-protection concerns are present. Reports should move beyond generic statements of confidentiality and specify where inputs are processed, which external providers receive data, what is retained and for how long, under which jurisdiction, and how access is controlled[32]. Where sensitive conversational data are processed by external providers, users should be informed at the point of use and, where feasible, given meaningful control over retention, secondary use, and deletion.

Population inclusion is a governance concern, not only a coverage gap. Assessments of bias and fairness are limited when intended populations, languages, and settings are absent from design and evaluation. Only one system was designed around fathers[72], none modelled the parental dyad despite the interdependence of maternal and paternal mental health[4]. None supported multilingual conversation or was developed in a low-income country. Higher-risk needs such as perinatal psychosis and birth-related trauma were not represented; reported cultural adaptation did not coincide with stronger safety infrastructure. Scalability alone does not ensure inclusion especially where underrepresentation may affect response relevance, interpretation, and safety. Inclusive design must shape language, delivery, therapeutic content, escalation, and evaluation from the outset.

The six design tensions show that governance challenges extend beyond the choice between rule-based and generative architectures. Conversational flexibility, personalisation, and automation can increase demands on monitoring, data use, and human oversight. Privacy-preserving methods, human escalation, and explicit AI-

identity disclosure may mitigate some of these trade-offs. The central question is whether architecture, data flows, and oversight are proportionate to intended function, population, infrastructure, and clinical risk.

The architecture comparison requires careful interpretation. Rule-based systems more often documented operational safeguards, but were also more often studied in clinical programmes with established ethics approval, consent, and referral pathways. LLM and hybrid systems were more often evaluated in technical studies without end users. Broader evidence indicates that only 16% of LLM-based mental-health chatbot studies have reached clinical-efficacy testing[24]. Architecture does not explain governance: Rosie, a hybrid system [81], was among the highest-scoring studies, while both classical machine-learning prototypes scored poorly despite their lower technical complexity. Technical sophistication and governance maturity are distinct dimensions of progress[43, 46, 104], and comparisons of governance across AI architectures should account for study design.

Across six dimensions, governance reporting remained low and uneven. No study reached the Comprehensive range, and even the highest-scoring studies did not report privacy-preserving architecture or empirical fairness evaluation. The GMS is a review-specific descriptive index, not a validated clinical, regulatory, or risk-assessment instrument. It makes governance reporting visible alongside technical and clinical outcomes but requires external validation before wider use.

Implications for practice and research

The limited evaluation evidence makes these decisions more important. Eleven studies enrolled no human end users, and only ten studies describing seven systems reported clinical outcomes in real users, with small or inconsistent benefits. Technical accuracy, usability, conversational fluency, and perceived empathy are useful development outcomes, but they do not establish reduced symptoms, prevention of harm, or improved parent-infant outcomes. The growing use of LLMs and simulation-based evaluation makes this distinction important, because conversational competence can be demonstrated before clinical benefit has been established.

Evidence and oversight should scale with intended function. Informational systems may primarily require comprehension, usability, and safety testing, whereas systems claiming treatment benefit, triage capability, or crisis support require prospective human-user evaluation with clinically meaningful outcomes and harm monitoring. Developers should document how sensitive data are handled, how model changes are managed, when users are escalated to human support, and where automated support reaches its limits. Clinicians should establish where data are processed, who monitors the system, how high-risk interactions are escalated, whether users know they are interacting with AI, and whether evaluation populations reflect intended users. Policymakers should require stronger expectations for privacy, transparency, monitoring, and human oversight for treatment, triage, and crisis-related roles. Certification, for example as an extension of Australia's National Safety and Quality Digital Mental Health Standards[105], should include technical verification of data flows and recertification when the underlying model or provider changes, consistent with emerging regulatory expectations [54, 55, 104-106].

Box 1 translates these gaps into eight proposed minimum standards; no included study documented all eight. The standards require prospective evaluation and proportionate application.

Box 1 | Proposed minimum safety and governance standards for perinatal CAs

#	Standard	Minimum operational requirement
1	Crisis recognition	Detection of suicidal ideation, self-harm, or crisis language, with tested sensitivity and false-negative rates reported.
2	Human escalation	A pathway to a qualified clinician or crisis service within a defined timeframe.
3	AI-identity disclosure	In-chat disclosure that the user is interacting with an AI.
4	Safety disclaimer	In-chat statement that the system is not a substitute for professional care.
5	Accessibility and inclusion	Language, cultural and family inclusivity, with delivery channels appropriate to the intended population and setting, including SMS or low-bandwidth access where needed.
6	Documented governance assessment	Transparent assessment across privacy, safety, monitoring, regulation, fairness and consent.
7	Privacy-preserving data governance	Documented data minimisation, encryption, access controls, and retention practices, with privacy-preserving computation where appropriate to the system's function and setting.
8	Real-world clinical validation	Evidence from human-user evaluation appropriate to the system's intended function, including clinically meaningful outcomes where clinical benefit is claimed.

Standards should be applied proportionately to intended function, clinical risk, and deployment setting, and require prospective evaluation.

Future research should move beyond demonstrating acceptable or empathetic CA responses. Crisis studies should report detection sensitivity, false-negative performance, completed escalation, and time to human response in real users, with continued monitoring after deployment. Advanced privacy-preserving approaches should be evaluated alongside usability, model performance, and clinical outcomes, with transparent reporting of data flows, retention, and model updates. Systems should involve fathers, non-birthing parents, couples, linguistically diverse populations, and low-connectivity communities throughout design and evaluation. The proposed standards also need prospective assessment of feasibility and effectiveness.

Limitations

Several limitations should inform interpretation. We assessed governance as documented in publications, so an unreported safeguard may have existed, and source code was unavailable for almost all systems. Restriction to English-language peer-reviewed source may underrepresent commercial commercially deployed systems, lower-resource settings, and systems with null or negative findings. This may also reflect a disconnect between commercial deployment and the peer-reviewed evidence base. Conversely, many systems described in the scientific literature are research prototypes that may not be available to consumers. Rapidly evolving systems may also have changed after publication. Data extraction was completed by one reviewer with independent verification of a subset. Formal critical appraisal was not undertaken. The GMS was developed after protocol registration and has not been validated; inter-rater reliability supports reproducibility within the review, not validity. Comparisons across model classes were descriptive only, rested on small groups, and are confounded by study design. Two conference papers analysed the same 223-record dataset, although collapsing repeated systems did not alter the conclusions. The design-tension synthesis was interpretive, and cost-effectiveness, business models, and disability-specific accessibility were not systematically assessed because reporting was sparse.

Conclusion

CAs may extend access to perinatal mental-health information, screening, and support, but technical capability alone does not establish responsible use. Advances in conversational architecture coexisted with limited privacy protection, unvalidated crisis detection, narrow population coverage, limited clinical oversight, and sparse evidence of patient benefit. The opportunity to introduce safeguards narrows as technologies become embedded in practice, reflecting the dilemma Collingridge described four decades ago[107]; that window is closing as conversational AI moves towards routine care. Governance-first development offers an organising principle for addressing these gaps. Whether a parent disclosing a mental health crisis receives timely, qualified human support depends on safeguards established before deployment and maintained thereafter. Perinatal conversational AI should be privacy-preserving by default, inclusive by design, clinically validated for its intended role and continuously monitored throughout its lifecycle.

Statements and Declarations

Funding: D.O.O. is supported by the James Cook University Postgraduate Research Scholarship (JCUPRS). S.J.T. is supported by an Australian Government National Health and Medical Research Council Investigator Grant (APP2025839). The funders had no role in the design or conduct of the review, study selection, data extraction, analysis, interpretation of the findings, or preparation of the manuscript.

Competing interests: S.J.T. and A.B.R.S. are authors of two included studies of Rover (Studies 27 and 35). To manage this potential conflict, D.O.O. independently assessed eligibility, extracted the data, and assigned GMS scores for these studies using the anchored scoring criteria, and both studies were excluded from the inter-rater subset. The remaining authors declare no competing interests.

Ethics approval: Not applicable. This scoping review analysed published studies and did not involve any human participants.

Consent to participate: Not applicable.

Consent to publish: Not applicable.

Data availability: All data supporting the findings of this review are provided with the article and its Supplementary Information. The review protocol and the verified dataset are available on the Open Science Framework (https://osf.io/63mpb).

Code availability: All code used to compute the statistical analyses and figures reported in this article is available at https://github.com/DesmondObe/perinatal-CA-governance-review and archived at Zenodo (https://doi.org/10.5281/zenodo.23060908)

Author contributions: Conceptualization: D.O.O., A.B.R.S.; Methodology: D.O.O., A.B.R.S.; Search strategy and protocol development: D.O.O., A.B.R.S., S.J.T.; Investigation: D.O.O., A.B.R.S.; Validation: A.B.R.S.; Formal analysis: D.O.O.; Data curation: D.O.O.; Software: D.O.O. A.B.R.S.; Visualization: D.O.O.; Writing – original draft: D.O.O.; Writing – review and editing: D.O.O., A.B.R.S., S.J.T., I.L.; Supervision: A.B.R.S., S.J.T., I.L.

Clinical trial number: not applicable.

References

[1] L. M. Howard and H. Khalifeh, "Perinatal mental health: a review of progress and challenges," (in eng), *World Psychiatry,* vol. 19, no. 3, pp. 313–327, Oct 2020, doi: 10.1002/wps.20769.

[2] J. Hahn-Holbrook, T. Cornwell-Hinrichs, and I. Anaya, "Economic and Health Predictors of National Postpartum Depression Prevalence: A Systematic Review, Meta-analysis, and Meta-Regression of 291 Studies from 56 Countries," (in eng), *Front Psychiatry,* vol. 8, p. 248, 2017, doi: 10.3389/fpsyt.2017.00248.

[3] C.-L. Dennis, K. Falah-Hassani, and R. Shiri, "Prevalence of antenatal and postnatal anxiety: Systematic review and meta-analysis," *British Journal of Psychiatry,* vol. 210, no. 5, pp. 315–323, 2017, doi: 10.1192/bjp.bp.116.187179.

[4] J. F. Paulson and S. D. Bazemore, "Prenatal and postpartum depression in fathers and its association with maternal depression: a meta-analysis," (in eng), *Jama,* vol. 303, no. 19, pp. 1961–9, May 19 2010, doi: 10.1001/jama.2010.605.

[5] W.-W. Rao *et al.*, "Prevalence of prenatal and postpartum depression in fathers: A comprehensive meta-analysis of observational surveys," *Journal of Affective Disorders,* vol. 263, pp. 491–499, 2020, doi: 10.1016/j.jad.2019.10.030.

[6] G. Bertram, N. Becuzzi, and E. Wilson, "Trends in Perinatal Mental Health Within a Tertiary Institution From 2013 to 2022," *The Australian & New Zealand Journal of Obstetrics & Gynaecology,* vol. 63, no. 6, pp. 821–824, 2023, doi: 10.1111/ajo.13720.

[7] C. A. Woody, A. J. Ferrari, D. J. Siskind, H. A. Whiteford, and M. G. Harris, "A systematic review and meta-regression of the prevalence and incidence of perinatal depression," (in eng), *J Affect Disord,* vol. 219, pp. 86–92, Sep 2017, doi: 10.1016/j.jad.2017.05.003.

[8] G. A. O'Dea *et al.*, "Associations between maternal psychological distress and mother-infant bonding: a systematic review and meta-analysis," (in eng), *Arch Womens Ment Health,* vol. 26, no. 4, pp. 441–452, Aug 2023, doi: 10.1007/s00737-023-01332-1.

[9] J. Slomian, G. Honvo, P. Emonts, J. Y. Reginster, and O. Bruyère, "Consequences of maternal postpartum depression: A systematic review of maternal and infant outcomes," (in eng), *Womens Health (Lond),* vol. 15, p. 1745506519844044, Jan–Dec 2019, doi: 10.1177/1745506519844044.

[10] A. Rogers *et al.*, "Association Between Maternal Perinatal Depression and Anxiety and Child and Adolescent Development: A Meta-analysis," (in eng), *JAMA Pediatr,* vol. 174, no. 11, pp. 1082–1092, Nov 1 2020, doi: 10.1001/jamapediatrics.2020.2910.

[11] A. Stein *et al.*, "Effects of perinatal mental disorders on the fetus and child," *The Lancet,* vol. 384, no. 9956, pp. 1800–1819, 2014/11/15/ 2014, doi: 10.1016/S0140-6736(14)61277-0.

[12] H. Khalifeh, I. M. Hunt, L. Appleby, and L. M. Howard, "Suicide in perinatal and non-perinatal women in contact with psychiatric services: 15 year findings from a UK national inquiry," *The Lancet Psychiatry* vol. 3, no. 3, pp. 233–242, 2016, doi: 10.1016/S2215-0366(16)00003-1.

[13] A. Felker, R. Patel, R. Kotnis, S. Kenyon, and M. Knight, "Saving Lives, Improving Mothers' Care: Lessons learned to inform maternity care from the UK and Ireland Confidential Enquiries into Maternal Deaths and Morbidity 2021-23 " in "MBRRACE-UK: Saving Lives, Improving Mothers' Care," National Perinatal Epidemiology Unit, University of Oxford., Oxford, 978-1-0687913-8-3, September 2025 2025. [Online]. Available: https://dx.doi.org/10.5287/ora-4javr692x

[14] M. C. Prom *et al.*, "A Systematic Review of Interventions That Integrate Perinatal Mental Health Care Into Routine Maternal Care in Low- and Middle-Income Countries," (in English), *Frontiers in Psychiatry,* Systematic Review vol. Volume 13 - 2022, 2022–March–14 2022, doi: 10.3389/fpsyt.2022.859341.

[15] J. Fisher *et al.*, "Prevalence and determinants of common perinatal mental disorders in women in low- and lower-middle-income countries: a systematic review," (in eng), *Bull World Health Organ,* vol. 90, no. 2, pp. 139g–149g, Feb 1 2012, doi: 10.2471/blt.11.091850.

[16] A. Sawyer, S. Ayers, and H. Smith, "Pre- and postnatal psychological wellbeing in Africa: a systematic review," (in eng), *J Affect Disord,* vol. 123, no. 1-3, pp. 17–29, Jun 2010, doi: 10.1016/j.jad.2009.06.027.

[17] E. Cribb, B. Spring, M. Galbally, F. Gardiner, and M. Coleman, "Perinatal psychiatric emergencies in rural Australia: In urgent need of attention," *Australian & New Zealand Journal of Psychiatry,* vol. 59, no. 2, pp. 109–114, 2025, doi: 10.1177/00048674241308689.

[18] M. Sambrook Smith, V. Lawrence, E. Sadler, and A. Easter, "Barriers to accessing mental health services for women with perinatal mental illness: systematic review and meta-synthesis of qualitative studies in the UK," *BMJ Open,* vol. 9, no. e024803, 2019, doi: 10.1136/bmjopen-2018-024803.

[19] C. L. Dennis and L. Chung-Lee, "Postpartum depression help-seeking barriers and maternal treatment preferences: a qualitative systematic review," (in eng), *Birth,* vol. 33, no. 4, pp. 323–31, Dec 2006, doi: 10.1111/j.1523-536X.2006.00130.x.

[20] R. Webb *et al.*, "Barriers and facilitators to implementing perinatal mental health care in health and social care settings: a systematic review," *The Lancet Psychiatry,* vol. 8, no. 6, pp. 521–534, 2021, doi: 10.1016/S2215-0366(20)30467-3.

[21] S. Zhong *et al.*, "The Usability, Feasibility, Acceptability, and Efficacy of Digital Mental Health Services in the COVID-19 Pandemic: Scoping Review, Systematic Review, and Meta-analysis," (in eng), *JMIR Public Health Surveill,* vol. 9, p. e43730, Feb 13 2023, doi: 10.2196/43730.

[22] H. Li, R. Zhang, Y.-C. Lee, R. E. Kraut, and D. C. Mohr, "Systematic review and meta-analysis of AI-based conversational agents for promoting mental health and well-being," *npj Digital Medicine,* vol. 6, no. 1, p. 236, 2023/12/19 2023, doi: 10.1038/s41746-023-00979-5.

[23] Y. He *et al.*, "Conversational Agent Interventions for Mental Health Problems: Systematic Review and Meta-analysis of Randomized Controlled Trials," (in English), *J Med Internet Res,* Review vol. 25, p. e43862, 2023, doi: 10.2196/43862.

[24] Y. Hua *et al.*, "Charting the evolution of artificial intelligence mental health chatbots from rule-based systems to large language models: a systematic review," (in eng), *World Psychiatry,* vol. 24, no. 3, pp. 383–394, Oct 2025, doi: 10.1002/wps.21352.

[25] J. Weizenbaum, "ELIZA—a computer program for the study of natural language communication between man and machine," *Commun. ACM,* vol. 9, no. 1, pp. 36–45, 1966, doi: 10.1145/365153.365168.

[26] R. Sanjeewa, R. Iyer, P. Apputhurai, N. Wickramasinghe, and D. Meyer, "Empathic Conversational Agent Platform Designs and Their Evaluation in the Context of Mental Health: Systematic Review," (in eng), *JMIR Ment Health,* vol. 11, p. e58974, Sep 9 2024, doi: 10.2196/58974.

[27] S. Amil *et al.*, "Interactive Conversational Agents for Perinatal Health: A Mixed Methods Systematic Review," *Healthcare,* vol. 13, no. 4, p. 363, 2025. [Online]. Available: https://www.mdpi.com/2227-9032/13/4/363.

[28] L. Olisaeloka, C. G. Richardson, A. Y. Wang, R. J. Munthali, and D. V. Vigo, "Safety Mechanisms and Risk Mitigation in Generative AI Mental Health Chatbots: A Systematic Scoping Review," (in eng), *Healthcare (Basel),* vol. 14, no. 10, May 20 2026, doi: 10.3390/healthcare14101395.

[29] K. S. Birk M, Myrick K, Schueller S, Torous J, "Digital Mental Health Research Priorities, Revisited for the AI and Large Language Model Era," *JMIR Ment Health,* vol. 13, 2026, doi: 10.2196/104118.

[30] M.-P. V. Austin, N. Reilly, V. Mule, D. Kingston, E. Black, and D. Hadzi-Pavlovic, "Disclosure of sensitive material at routine antenatal psychosocial assessment: The role of psychosocial risk and mode of assessment," *Women and Birth,* vol. 35, no. 2, pp. e125–e132, 2022, doi: 10.1016/j.wombi.2021.04.005.

[31] S. S. Khalil, N. S. Tawfik, and M. Spruit, "Federated learning for privacy-preserving depression detection with multilingual language models in social media posts," *Patterns,* vol. 5, no. 7, p. 100990, 2024, doi: 10.1016/j.patter.2024.100990.

[32] K. Huckvale, J. Torous, and M. E. Larsen, "Assessment of the Data Sharing and Privacy Practices of Smartphone Apps for Depression and Smoking Cessation," *JAMA Network Open,* vol. 2, no. 4, pp. e192542–e192542, 2019, doi: 10.1001/jamanetworkopen.2019.2542.

[33] P. Ramchandani, A. Stein, J. Evans, and T. G. O'Connor, "Paternal depression in the postnatal period and child development: a prospective population study," *The Lancet,* vol. 365, no. 9478, pp. 2201–2205, 2005/06/25/ 2005, doi: 10.1016/S0140-6736(05)66778-5.

[34] P. Böhlmann *et al.*, "The prospective impact of extradyadic stress on depressive symptoms and the mediating role of intradyadic stress in parents–an actor-partner interdependence mediation model," *PLOS ONE,* vol. 19, no. 11, p. e0311989, 2024, doi: 10.1371/journal.pone.0311989.

[35] L. M. D. A. Jiyeong Kim, Heejung Bang, Steffi Prajogo, Yong K Choi, Jeffrey S Hoch, Elizabeth L Prado, "Effectiveness of Digital Mental Health Tools to Reduce Depressive and Anxiety Symptoms in Low- and Middle-Income Countries: Systematic Review and Meta-analysis," *JMIR Ment Health,* vol. 10, p. e43066, 2023, doi: 10.2196/43066.

[36] K. A. A. John A Naslund, Ricardo Araya, Lisa A Marsch, Jürgen Unützer, Vikram Patel, Stephen J Bartels, "Digital technology for treating and preventing mental disorders in low-income and middle-income countries: a narrative review of the literature - ClinicalKey," *Lancet Psychiatry,* vol. 4, no. 6, p. 14, 2017. [Online]. Available: http://dx.doi.org/10.1016/S2215-0366(17)30096-2.

[37] A. A. Abd-Alrazaq, A. Rababeh, M. Alajlani, B. M. Bewick, and M. Househ, "Effectiveness and Safety of Using Chatbots to Improve Mental Health: Systematic Review and Meta-Analysis," (in eng), *J Med Internet Res,* vol. 22, no. 7, p. e16021, Jul 13 2020, doi: 10.2196/16021.

[38] M. Casu, S. Triscari, S. Battiato, L. Guarnera, and P. Caponnetto, "AI Chatbots for Mental Health: A Scoping Review of Effectiveness, Feasibility, and Applications," *Applied Sciences,* vol. 14, no. 13, p. 5889, 2024. [Online]. Available: https://www.mdpi.com/2076-3417/14/13/5889.

[39] R. Balan and T. P. Gumpel, "ChatGPT Clinical Use in Mental Health Care: Scoping Review of Empirical Evidence," (in eng), *JMIR Ment Health,* vol. 12, p. e81204, Dec 24 2025, doi: 10.2196/81204.

[40] L. Olisaeloka, C. G. Richardson, A. Y. Wang, R. J. Munthali, and D. V. Vigo, "Generative AI mental health chatbots: a scoping review of intervention design and user experience," *npj Digital Medicine,* 2026/07/23 2026, doi: 10.1038/s41746-026-02972-0.

[41] S. Samir Khalil, N. Tawfik, and M. Spruit, "Exploring the potential of federated learning in mental health research: a systematic literature review," *Applied Intelligence,* vol. 54, pp. 1–18, 01/12 2024, doi: 10.1007/s10489-023-05095-1.

[42] S. Gehder, S. Brückner, S. Gilbert, and M. Goeldner, "A scoping review on using real-world data to evaluate the effectiveness of mHealth applications," (in eng), *NPJ Digit Med,* vol. 9, no. 1, Apr 8 2026, doi: 10.1038/s41746-026-02562-0.

[43] A. Z. Rowan Hussein, Bashar Ramadan, Frederick M. Howard, Maia Hightower, Sachin Shah & Brett K. Beaulieu-Jones, "Advancing healthcare AI governance through a comprehensive maturity model based on systematic review," *npj Digital Medicine,* vol. 9, no. 236, 2026, doi: 10.1038/s41746-026-02418-7.

[44] A. H. Jee Young Kim, Suresh Balu & Mark Sendak "People process technology and operations framework for establishing AI governance in healthcare organizations," *npj Digital Medicine,* vol. 9, no. 224, 2026, doi: 10.1038/s41746-026-02419-6.

[45] U. S. Anaduaka, A. O. Oladosu, S. Katsande, C. S. Frempong, and S. Awuku-Amador, "Leveraging artificial intelligence in the prediction, diagnosis and treatment of depression and anxiety among perinatal women in low- and middle-income countries: a systematic review," *BMJ Mental Health,* vol. 28, no. 1, p. e301445, 2025, doi: 10.1136/bmjment-2024-301445.

[46] K. Cunningham, V. Mărginean, and R. Hylock, "Navigating promise and perils: applying artificial intelligence to the perinatal mental health care cascade," *npj Health Systems,* vol. 2, no. 1, p. 26, 2025/07/23 2025, doi: 10.1038/s44401-025-00030-7.

[47] H. Tameez and N. S. Appleton, "AI as Emotional Support in Pregnancy: A Review and Synthesis for Emerging Research Directions," *Social Sciences,* vol. 15, no. 2, p. 100, 2026. [Online]. Available: https://www.mdpi.com/2076-0760/15/2/100.

[48] H. Yang and D.-M. Lin, "Conversational AI for perinatal mental health: promise, limits, and a human-AI stepped-care framework," (in English), *Frontiers in Psychiatry,* Perspective vol. Volume 17 - 2026, 2026–May–13 2026, doi: 10.3389/fpsyt.2026.1847854.

[49] S. M. S. Islam, A. W. Gemmill, Y. Hirshler, M. Pascoe, and J. Milgrom, "Application of Artificial Intelligence in Perinatal Mental Health: A Review," *AI,* vol. 7, no. 8, p. 314, 2026. [Online]. Available: https://www.mdpi.com/2673-2688/7/8/314.

[50] M. D. J. Peters *et al.*, "Updated methodological guidance for the conduct of scoping reviews," *JBI Evidence Synthesis,* vol. 18, no. 10, pp. 2119–2126, 2020, doi: 10.11124/jbies-20-00167.

[51] A. C. Tricco *et al.*, "PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation," (in eng), *Ann Intern Med,* vol. 169, no. 7, pp. 467–473, Oct 2 2018, doi: 10.7326/m18-0850.

[52] M. D. J. Peters *et al.*, "Scoping reviews: reinforcing and advancing the methodology and application," *Systematic Reviews,* vol. 10, no. 1, p. 263, 2021/10/08 2021, doi: 10.1186/s13643-021-01821-3.

[53] D. Obe, A. Shatte, S. Teague, and I. Lee, "AI-Supported Conversational Agents for Perinatal Mental Health: A Scoping Review of Applications, Opportunities, and Challenges in Computational, Ethical, and Privacy-Preserving Design," OSF, 2025/11/03/ 2025. Accessed: 2026/03/10/09:09:28. [Online]. Available: https://osf.io/63mpb

[54] W. H. Organization, "Ethics and governance of artificial intelligence for health: WHO guidance Executive summary," World Health Organization, Geneva, 9789240037403, 2021. Accessed: **2026/07/06/03:50:31**. [Online]. Available: https://www.who.int/publications/i/item/9789240037403

[55] E. P. a. C. o. t. E. Union, "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, and repealing Directive 95/46/EC (General Data Protection Regulation)," European Union, Regulation (EU) 2016/679, 4 May 2016 2016, vol. L 119. [Online]. Available: http://data.europa.eu/eli/reg/2016/679/oj

[56] T. K. Koo and M. Y. Li, "A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research," (in English), *Journal of Chiropractic Medicine,* vol. 15, no. 2, pp. 155–163, 01 Jun 2016 2016, doi: 10.1016/j.jcm.2016.02.012.

[57] E. Tabassi, "Artificial Intelligence Risk Management Framework (AI RMF 1.0)," National Institute of Standards and Technology, Gaithersburg, MD, NIST AI 100-1, January 2023 2023. [Online]. Available: https://doi.org/10.6028/NIST.AI.100-1

[58] Food and Drug Administration, "Clinical Decision Support Software: Guidance for Industry and Food and Drug Administration Staff," U.S. Food and Drug Administration, Silver Spring, MD, September 2022 2022. [Online]. Available: https://www.federalregister.gov/d/2022-20993

[59] E. P. a. C. o. t. E. Union, "Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act)," in "Official Journal of the European Union," European Union, Regulation (EU) 2024/1689, 12 July 2024 2024. [Online]. Available: http://data.europa.eu/eli/reg/2024/1689/oj/eng

[60] U. Government, "Regulation (EU) 2016/679 of the European Parliament and of the Council," 2016. [Online]. Available: https://www.legislation.gov.uk/eur/2016/679/article/9

[61] J. Y. X. Chua *et al.*, "A Mobile App–Based Intervention (Parentbot–a Digital Healthcare Assistant) for Parents: Secondary Analysis of a Randomized Controlled Trial," (in English), *J Med Internet Res,* Original Paper vol. 27, p. e64882, 2025, doi: 10.2196/64882.

[62] J. Y. X. Chua *et al.*, "The effectiveness of Parentbot - a digital healthcare assistant - on parenting outcomes: A randomized controlled trial," (in eng), *Int J Nurs Stud,* vol. 160, p. 104906, Dec 2024, doi: 10.1016/j.ijnurstu.2024.104906.

[63] J. Y. X. Chua *et al.*, "'Parentbot - A Digital healthcare Assistant (PDA)': A mobile application-based perinatal intervention for parents: Development study," (in eng), *Patient Educ Couns,* vol. 114, p. 107805, Sep 2023, doi: 10.1016/j.pec.2023.107805.

[64] S. Suharwardy *et al.*, "Feasibility and impact of a mental health chatbot on postpartum mental health: a randomized controlled trial," (in eng), *AJOG Glob Rep,* vol. 3, no. 3, p. 100165, Aug 2023, doi: 10.1016/j.xagr.2023.100165.

[65] A. Darcy *et al.*, "Anatomy of a Woebot® (WB001): agent guided CBT for women with postpartum depression," (in eng), *Expert Rev Med Devices,* vol. 19, no. 4, pp. 287–301, Apr 2022, doi: 10.1080/17434440.2022.2075726.

[66] V. Fietta *et al.*, "World Health Organization Evidence-Based Self-Help Plus Intervention for Stress Management via Chatbot: Protocol for Adaptation to a Tech-Enabled Model," (in eng), *JMIR Res Protoc,* vol. 14, p. e69644, Jun 26 2025, doi: 10.2196/69644.

[67] S. Rizzi *et al.*, "A Chatbot-Based Version of a World Health Organization-Validated Intervention (Self-Help Plus) for Stress Management in Pregnant Women: Protocol for a Usability Study," (in eng), *JMIR Res Protoc,* vol. 14, p. e53891, Aug 1 2025, doi: 10.2196/53891.

[68] I. W. Puspitasari, F. R. Rinawan, W. G. Purnama, H. Susiarno, and A. I. Susanti, "Development of a Chatbot for Pregnant Women on a Posyandu Application in Indonesia: From Qualitative Approach to Decision Tree Method," *Informatics,* vol. 9, no. 4, p. 88, 2022. [Online]. Available: https://www.mdpi.com/2227-9709/9/4/88.

[69] J. Wong, A. C. Foussat, S. Ting, E. Acerbi, R. M. van Elburg, and C. Mei Chien, "A Chatbot to Engage Parents of Preterm and Term Infants on Parental Stress, Parental Sleep, and Infant Feeding: Usability and Feasibility Study," (in eng), *JMIR Pediatr Parent,* vol. 4, no. 4, p. e30169, Oct 26 2021, doi: 10.2196/30169.

[70] E. Mancinelli, S. Magnolini, S. Gabrielli, and S. Salcuni, "A Chatbot (Juno) Prototype to Deploy a Behavioral Activation Intervention to Pregnant Women: Qualitative Evaluation Using a Multiple Case Study," (in eng), *JMIR Form Res,* vol. 8, p. e58653, Aug 14 2024, doi: 10.2196/58653.

[71] S. Rizzi *et al.*, "A Mindfulness-Based App Intervention for Pregnant Women: Protocol for a Pilot Feasibility Study," (in eng), *JMIR Res Protoc,* vol. 13, p. e53890, May 10 2024, doi: 10.2196/53890.

[72] S. J. Teague, A. B. R. Shatte, M. Fuller-Tyszkiewicz, and D. M. Hutchinson, "User-centred design and evaluation of an mHealth app for fathers' perinatal mental health: a feasibility, acceptability, and usability study," (in eng), *Behav Inf Technol,* pp. 1–15, May 9 2025, doi: 10.1080/0144929x.2025.2502474.

[73] S. J. Teague, A. B. R. Shatte, M. Fuller-Tyszkiewicz, and D. M. Hutchinson, "Mobile app-based intervention for paternal perinatal depression, anxiety, and stress: A randomised controlled trial," (in eng), *J Affect Disord,* vol. 382, pp. 325–335, Aug 1 2025, doi: 10.1016/j.jad.2025.04.075.

[74] K. McAlister, L. Baez, J. Huberty, and M. Kerppola, "Chatbot to Support the Mental Health Needs of Pregnant and Postpartum Women (Moment for Parents): Design and Pilot Study," (in eng), *JMIR Form Res,* vol. 9, p. e72469, Apr 8 2025, doi: 10.2196/72469.

[75] S. S. Sadavarte and E. Bodanese, "Pregnancy Companion Chatbot Using Alexa and Amazon Web Services," presented at the IEEE Pune Section International Conference (PuneCon), Pune, India, 18–20 Dec. 2019, 2019. [Online]. Available: https://ieeexplore.ieee.org/document/9105762.

[76] S. García-Méndez and F. de Arriba-Pérez, "Detecting and Explaining Postpartum Depression in Real-Time with Generative Artificial Intelligence," *Applied Artificial Intelligence,* vol. 39, no. 1, p. 2515063, 2025/12/31 2025, doi: 10.1080/08839514.2025.2515063.

[77] D. Balasooriya *et al.*, "Baby Bump: A Monitoring System for Pregnant Mothers and Babies," presented at the 2024 IEEE 9th International Conference for Convergence in Technology (I2CT), Pune, India, 5–7 April 2024, 2024. [Online]. Available: https://ieeexplore.ieee.org/document/10544017.

[78] D. B. Edirisinghe, J. A. M. G. Jayakody, M. L. Sandeepani, I. E. Madiwilage, J. Wickramarathne, and T. Shyamalee, "'e-Midwife'—Empowering Mothers with Smart, Personalized Support," presented at the Intelligent Sustainable Systems, Singapore, 18 March 2025, 2025. [Online]. Available: https://doi.org/10.1007/978-981-97-8695-4_10.

[79] K. Chung, H. Y. Cho, and J. Y. Park, "A Chatbot for Perinatal Women's and Partners' Obstetric and Mental Health Care: Development and Usability Evaluation Study," (in eng), *JMIR Med Inform,* vol. 9, no. 3, p. e18607, Mar 3 2021, doi: 10.2196/18607.

[80] J. L. Z. Montenegro, C. A. da Costa, and L. P. Janssen, "Evaluating the use of chatbot during pregnancy: A usability study," *Healthcare Analytics,* vol. 2, no. 100072, 2022/11/01/ 2022, doi: 10.1016/j.health.2022.100072.

[81] Q. C. Nguyen *et al.*, "Rosie, a Health Education Question-and-Answer Chatbot for New Mothers: Randomized Pilot Study," *JMIR Formative Research,* vol. 8, 2024/01/01/ 2024, doi: 10.2196/51361.

[82] M. A. B. Lucia Vaira, Matteo Conte, Francesco Margiotta Casaluci, Antonio Melpignano, "MamaBot: a System based on ML and NLP for supporting Women and Families during Pregnancy," presented at the Proceedings of the 22nd International Database Engineering & Applications Symposium, New York, NY, USA, 18 June 2018, 2018. [Online]. Available: https://doi.org/10.1145/3216122.3216173.

[83] Z. Khan *et al.*, "MOMCare with AI: A Dual Embedding-Based RAG-LLM Chatbot for Postpartum Depression," presented at the ICT for Intelligent Systems, Singapore, 2025. [Online]. Available: https://doi.org/10.1007/978-981-95-1361-1_37.

[84] S. Viswanathan, S. Ibrahim, R. Shankar, R. Binns, M. V. Kleek, and P. Slovak, "The Interaction Layer: An Exploration for Co-Designing User-LLM Interactions in Parental Wellbeing Support Systems," presented at the Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, 2025. [Online]. Available: https://doi.org/10.1145/3706598.3714088.

[85] G. Kerimoglu Yildiz, R. Turk Delibalta, and Z. Coktay, "Artificial intelligence-assisted chatbot: impact on breastfeeding outcomes and maternal anxiety," *BMC Pregnancy and Childbirth,* vol. 25, no. 1, p. 631, 2025/05/30 2025, doi: 10.1186/s12884-025-07753-3.

[86] P. Valan and P. Venugopal, "Evaluating a retrieval-augmented pregnancy chatbot: a comprehensibility–accuracy-readability study of the DIAN AI assistant," (in English), *Frontiers in Artificial Intelligence,* Original Research vol. Volume 8 - 2025, 2025–September–22 2025, doi: 10.3389/frai.2025.1640994.

[87] N. Mayangsari, "Utilizing Chat GPT as a Digital Health Intervention in Pregnancy Care to Improve Maternal Well-Being," *Social Medicine,* vol. 18, no. 3, pp. 258–266, March 18, 2025 2025, doi: 10.71164/socialmedicine.v18i3.2025.2003.

[88] F. T. Progga, A. Khan, and S. Rubya, "Large Language Models and Personalized Storytelling for Postpartum Wellbeing," presented at the Companion Publication of the 2024 Conference on Computer-Supported Cooperative Work and Social Computing, San Jose, Costa Rica, 2024. [Online]. Available: https://doi.org/10.1145/3678884.3681921.

[89] E. P. Green *et al.*, "Expanding Access to Perinatal Depression Treatment in Kenya Through Automated Psychological Support: Development and Usability Study," (in eng), *JMIR Form Res,* vol. 4, no. 10, p. e17895, Oct 5 2020, doi: 10.2196/17895.

[90] S. Yang, J. Lee, E. Sezgin, J. Bridge, and S. Lin, "Clinical Advice by Voice Assistants on Postpartum Depression: Cross-Sectional Investigation Using Apple Siri, Amazon Alexa, Google Assistant, and Microsoft Cortana," (in eng), *JMIR Mhealth Uhealth,* vol. 9, no. 1, p. e24045, Jan 11 2021, doi: 10.2196/24045.

[91] B. Inkster, M. Kadaba, and V. Subramanian, "Understanding the impact of an AI-enabled conversational agent mobile app on users' mental health and wellbeing with a self-reported maternal event: a mixed method real-world data mHealth study," (in eng), *Front Glob Womens Health,* vol. 4, p. 1084302, 2023, doi: 10.3389/fgwh.2023.1084302.

[92] R. Wang, J. Wang, Y. Liao, and J. Wang, "Supervised Machine Learning Chatbots for Perinatal Mental Healthcare," presented at the 2020 International Conference on Intelligent Computing and Human-Computer Interaction (ICHCI), Sanya, China, 4–6 Dec. 2020, 2020. [Online]. Available: https://ieeexplore.ieee.org/document/9424857.

[93] P. Kaur, R. Sharma, and M. Singh, "Perinatal Mental Detection With Chatbots Using Machine Learning," presented at the 2024 4th International Conference on Technological Advancements in Computational Sciences (ICTACS), Tashkent, Uzbekistan, 13–15 Nov. 2024, 2024. [Online]. Available: https://ieeexplore.ieee.org/document/10840611.

[94] S. T. Bakhtawar Ahtisham, Zoha Hayat Bhatti, Ayesha Masood, and Suleman Shahid, "An AI Chat-Based Solution Aimed to Screen Postpartum Depression," presented at the Human-Computer Interaction – INTERACT 2023, LNCS 14143, Springer, Abdelnour Nocera, 2023. [Online]. Available: https://doi.org/10.1007/978-3-031-42283-6_18.

[95] G. W. Sam Jiangyue Yan, Thanh Vo, Yutong Xiao, Huy Truong "Design and Evaluation of a Prototype 'HiBaby' for Pregnancy Assistance," presented at the Proceedings of the 2022 International Conference on Human Machine Interaction, New York, NY, USA, 2022. [Online]. Available: https://doi.org/10.1145/3560470.3560477.

[96] S. Button, A. Thornton, S. Lee, J. Shakespeare, and S. Ayers, "Seeking help for perinatal psychological distress: a meta-synthesis of women's experiences," *British Journal of General Practice,* vol. 67, no. 663, pp. e692–e699, 2017, doi: 10.3399/bjgp17X692549.

[97] R. Huo *et al.*, "Role of Chatbots on Supporting Women in Perinatal Period: A Scoping Review," *Women's Health Reports,* vol. 7, p. 26884844261453447, 2026, doi: 10.1177/26884844261453447.

[98] H.-s. Wei, P. Atwal, L. Humphrey, and Y.-a. Yu, "Evaluating the safety of AI responses in suicide-related interactions: A clinically informed framework and simulation study," *Psychiatry Research Communications,* vol. 6, no. 2, p. 100263, 2026/06/01/ 2026, doi: 10.1016/j.psycom.2026.100263.

[99] B. Hassan and A. Ahmed, "Fluency without understanding: risks of large language models in mental healthcare," *BJPsych Bulletin,* pp. 1–4, 2026, doi: 10.1192/bjb.2026.10227.

[100] L. Chen, M. A. Zaharia, and J. Y. Zou, "How is ChatGPT's behavior changing over time?," *ArXiv,* vol. abs/2307.09009, 2023.

[101] D. Freed, J. Palmer, D. Minchala, K. Levy, T. Ristenpart, and N. Dell, "'A Stalker's Paradise': How Intimate Partner Abusers Exploit Technology," presented at the Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, Montreal QC, Canada, 2018. [Online]. Available: https://doi.org/10.1145/3173574.3174241.

[102] K. Huckvale, J. T. Prieto, M. Tilney, P. J. Benghozi, and J. Car, "Unaddressed privacy risks in accredited health and wellness apps: a cross-sectional systematic assessment," *BMC Med,* vol. 13, no. 214, p. 214, Sep 7 2015, doi: 10.1186/s12916-015-0444-y.

[103] I. Fiaz, N. Kanwal, and A. Al-Said Ahmad, "A Systematic Review of Federated and Cloud Computing Approaches for Predicting Mental Health Risks," (in eng), *Sensors (Basel),* vol. 26, no. 1, Dec 30 2025, doi: 10.3390/s26010229.

[104] B. Meskó and E. J. Topol, "The imperative for regulatory oversight of large language models (or generative AI) in healthcare," *npj Digital Medicine,* vol. 6, no. 1, p. 120, 2023/07/06 2023, doi: 10.1038/s41746-023-00873-0.

[105] Australian Commission on Safety and Quality in Health Care, "National Safety and Quality Digital Mental Health Standards," Australian Commission on Safety and Quality in Health Care, Sydney, 2020. [Online]. Available: https://www.safetyandquality.gov.au/national-standards/digital-mental-health-standards

[106] J. C. L. Ong *et al.*, "Innovating global regulatory frameworks for generative AI in medical devices is an urgent priority," *npj Digital Medicine,* vol. 9, no. 1, p. 364, 2026/03/19 2026, doi: 10.1038/s41746-026-02552-2.

[107] D. Collingridge, *The Social Control of Technology*. St. Martin's Press, 1980.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ESM1.pdf
- ESM2.pdf
- ESM3.xlsx