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 design 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

Article

Keywords:

Posted Date: October 15th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11222550/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 design 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*

**Correspondence: desmondobinna.obe@my.jcu.edu.au*

Abstract

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 responsible development remains poorly characterised. This scoping review examined how perinatal CAs are built, governed and evaluated, whom they serve, and what safeguards they incorporate. Nine interdisciplinary databases were searched for relevant studies (January 2015–September 2025). Of 6,118 records, 35 studies describing 30 unique CAs across 14 countries were included. Rule-based architectures predominated; large language model (LLM) and hybrid systems emerged rapidly after 2022. Twenty-two studies described screening, treatment, or crisis-related functions, yet therapeutic foundations were inconsistently reported. No study reported advanced privacy-preserving computation, although 14 relied on third-party commercial APIs. Nine systems reported automated risk detection, but none reported crisis-language sensitivity or false-negative performance; only one linked crisis detection to human follow-up within a defined timeframe. Fathers, couples, non-birthing parents, linguistically diverse populations, and users in low-connectivity settings remained poorly represented. Several operational safeguards were less consistently documented in newer architectures, a pattern reflected in the six-dimension Governance Maturity Score (0–24): median scores were 10 for rule-based and 5 for LLM-hybrid studies ($p = 0.003$), persisting across four sensitivity analyses. We describe this architecture-governance contrast as the *AI perinatal paradox.* Eight minimum standards are proposed to support governance-first development, in which developers, clinicians, and policymakers embed safety, equity, privacy, and human oversight from design to deployment of perinatal CAs.

Keywords: perinatal mental health; conversational agents; large language models; AI governance; privacy-preserving; federated learning; clinical readiness; 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 beyond the affected parent to impaired parent-infant bonding[8], disrupted family functioning[9], and adverse neurodevelopmental outcomes in offspring[10,11]. In severe cases, perinatal mental disorders contribute to maternal suicide[12], which remains among the leading causes of mortality during 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, financial 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 highly sensitive information concerning 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]. Data practices are central to responsible CA design. Key considerations include the provenance of data used to develop or evaluate models, where data are stored and processed, the involvement of third-party services, how systems are updated, and the transparency of these processes. 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, secondary use, and re-identification risk[32].

Beyond data governance, important questions remain about who these systems are designed for and how they function in practice. Paternal perinatal mental health is linked to maternal wellbeing[4] and child outcomes[33], while parents may also influence each other as an interdependent dyad[34]. Yet whether perinatal CAs meaningfully include fathers/non-birthing partners or model parents dyadically has not been systematically 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[28], and safety[40]. 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.

To address these gaps, 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 compare heterogeneous systems and assess their governance practices, we developed four complementary outputs: a Governance Maturity Score, a governance-by-model-class visualisation, a taxonomy of six recurrent design tensions, and eight proposed 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.

Results

Screening results

Our search identified 6,118 records (6,045 from nine databases, 73 from manual and grey-literature searching). After removing 2,404 duplicates, 3,714 records underwent title-and-abstract screening, of which 65 proceeded to full-text review. One report could not be retrieved, leaving 64; of these, 32 met the inclusion criteria. Forward and backward citation tracking and a supplementary Google Scholar search identified three more eligible studies, bringing the total to 35 (Fig. 1).

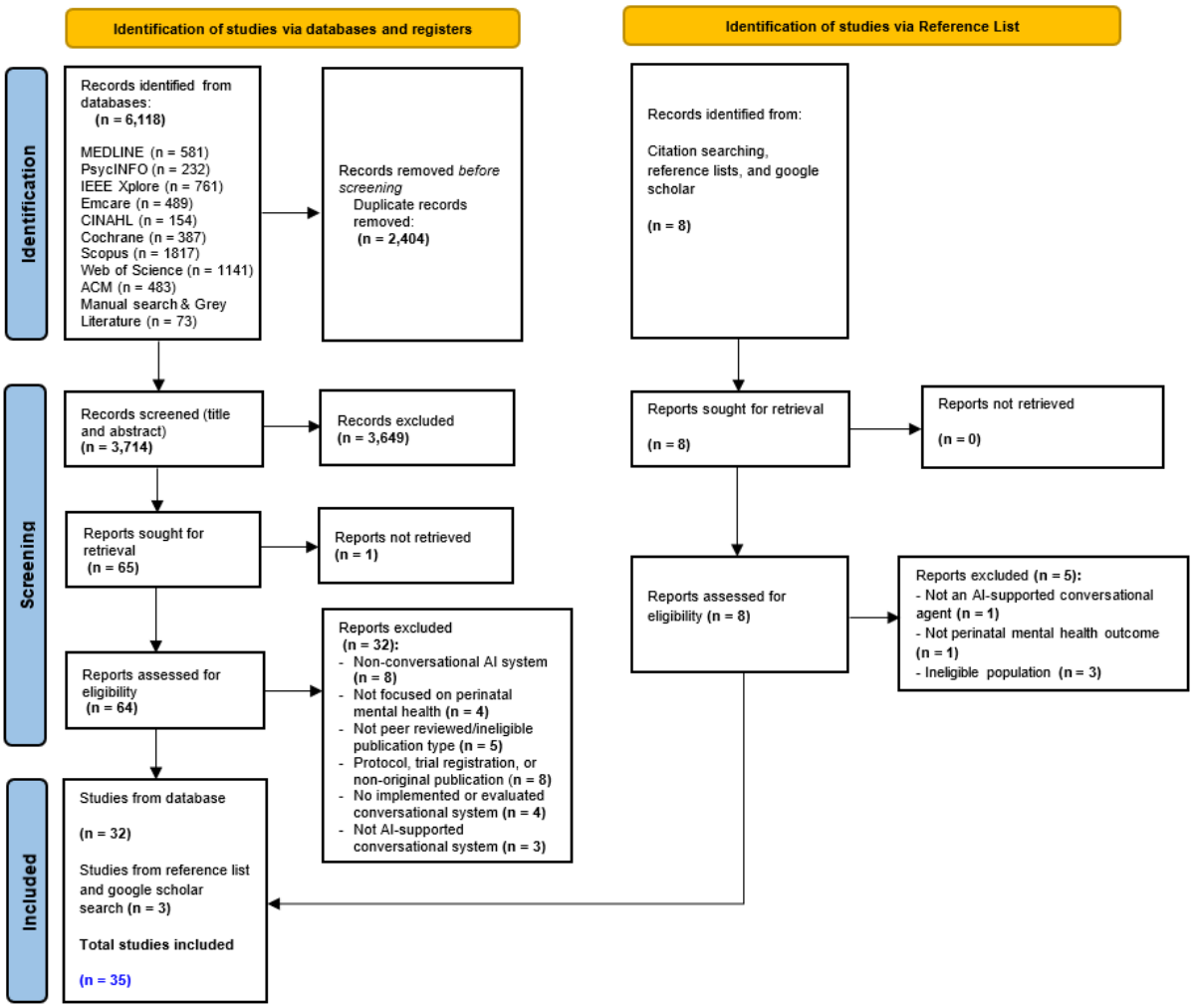

Fig. 1 | PRISMA-ScR flow diagram of the study selection process.

Descriptive overview

The 35 included studies described 30 unique perinatal CAs. Four systems were evaluated in more than one study: *Parentbot* (Studies 1, 17, and 34), *Woebot* (Studies 5 and 30), *ALBA* (Studies 14 and 22), and *Rover* (studies 27 and 35). The remaining 26 appeared once.

Publication activity accelerated across the review period: one study (3%) was published in 2015-2018, six (17%) in 2019-2021, and 28 (80%) in 2022-2025. Twenty-four studies (69%) were published in peer-reviewed journals and 11 (31%) in conference proceedings. Journal publications generally represented more mature stages of evaluation: 18 of 24 were clinical pilots or later-stage evaluations, whereas all 11 conference papers were conceptual or laboratory-stage studies.

Two of the 30 unique CAs were general-purpose commercial systems evaluated for perinatal use, leaving 28 purpose-built systems. Of these, 12 (43%) were developed by computing or engineering teams, seven (25%) by interdisciplinary teams, five (18%) by commercial companies, three (11%) by health teams, and one (4%) by a team from another academic discipline. Computing teams accounted for 10 of the 11 conference studies; the remaining study was developed by an interdisciplinary computing and clinical team. Clinical involvement in system development or validation was reported in 23 of 35 studies (66%), while four (11%) reported no clinical involvement and eight (23%) provided insufficient information to determine whether clinicians were involved. The 30 unique systems were developed across 14 countries, although development was

concentrated in high-income settings. Twenty systems (67%) were developed in high-income countries, four (13%) in upper-middle-income countries, five (17%) in lower-middle-income countries, and one (3%) USA-Kenya collaboration (Fig. 2). The USA contributed seven systems independently and one collaboratively, followed by Italy with four and the United Kingdom with three. All other countries contributed one or two systems. Evaluation followed the same pattern: 19 of the 35 studies (53%) were conducted in high-income countries, four (11%) in upper-middle-income and four (11%) in lower-middle-income settings, and eight (23%) did not report an evaluation country.

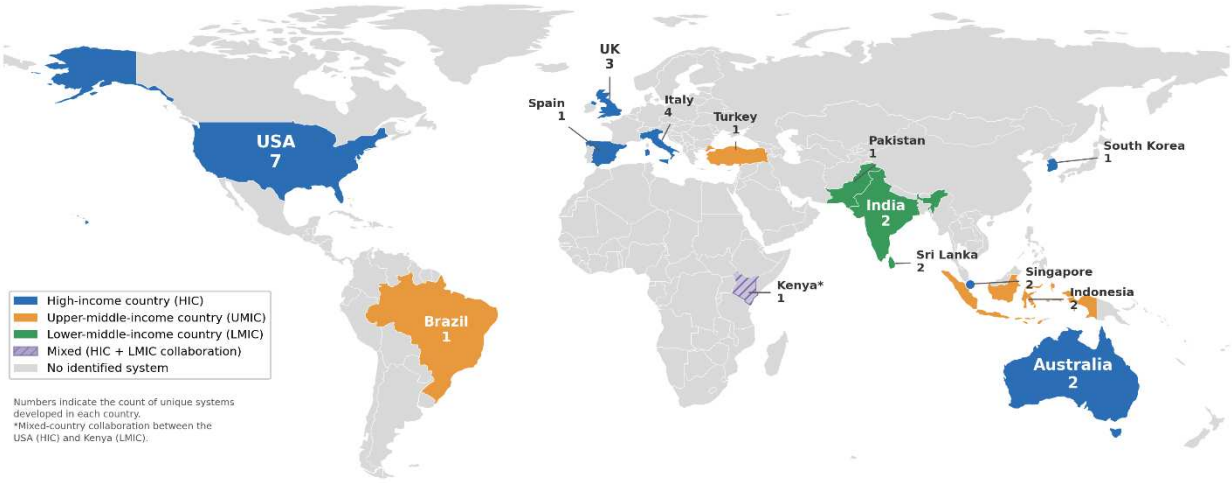

Fig. 2* | *Geographic distribution of perinatal CAs. The 30 unique systems were developed across 14 countries, classified by World Bank income group. *Zuri* was jointly developed in the USA and Kenya.

At study level, text was the reported dialogue modality in 32 studies (91%); two (6%) used voice-based systems and one (3%) did not report the dialogue modality. Mobile applications were the most reported delivery platform, used in 25 studies (71%). Two studies (6%) reported both app and web interfaces, one (3%) used a web interface only and one (3%) used SMS and a messaging platform; six studies (18%) did not report the delivery platform. Overall, 27 studies (77%) evaluated CAs delivered through a mobile application or an app-containing platform.

Mixed-methods evaluations were the most common design (n = 11, 31%), followed by simulation studies (n = 8, 23%), and randomised controlled trials (n = 5, 14%); the remainder comprised feasibility, pilot, quasi-experimental, observational, cross-sectional and protocol reports (Supplementary Data S1, sheet 1). Study samples ranged from two participants to an analytical dataset of 380,500 records; the median human-user sample size was 50 participants. Eleven studies (31%) enrolled no human end-users at all, relying instead on simulation, secondary dataset analysis, or expert review.

Computational architectures and model classes

Across the unique systems (n = 30), rule-based architecture predominated (n = 10, 33%), followed by hybrid systems (n = 7, 23%) and LLM systems (n = 6, 20%). Two systems (7%) used classical machine learning alone, three (10%) relied on commercial AI services of undisclosed architecture, and two (7%) claimed AI but reported no specific model. At study level (N = 35), the same classification yields 15 rule-based, 7 hybrid, 6 LLM, 3 unspecified-AI, 2 classical machine learning, and 2 not reported studies (Table 1; Supplementary Data 1, sheets 1 and 2).

Table 1 | 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%)	15 (43%)	*Parentbot*(Studies 1, 17 and 34)[52-54]; *Woebot*(Studies 5 and 30)[55,56]; *ALBA*(Studies 14 and 22)[57,58]; *iPosyandu chatbot*(Study 8)[59]; *ClaimIt chatbot*(Study 23)[60]; *Juno*(Study 16)[61]; *Maia*(Study 26)[62]; *Rover*(Studies 27 and 35)[63,64]; *Moment for Parents*(Study 28)[65]; *Pregnancy Companion*(Study 32)[66]
Hybrid	7 (23%)	7 (20%)	*PPD screener*(Study 4)[67]; *Baby Bump*(Study 13)[68]; *e-Midwife*(Study 15)[69]; *Dr. Joy*(Study 18)[70]; *Maria*(Study 19)[71]; *Rosie*(Study 29)[72]; *MamaBot*(Study 33)[73]
Large Language Model (LLM)	6 (20%)	6 (17%)	*MOMCare*(Study 6)[74]; *NurtureBot*(Study 7)[75]; *Breastfeeding chatbot*(Study 11)[76]; *DIAN*(Study 12)[77]; *ChatGPT intervention*(Study 20)[78]; *Storytelling LLM*(Study 24)[79]
Unspecific AI	3 (10%)	3 (9%)	*Zuri*(Study 2)[80]; *Voice assistants*(Study 3)[81]; *Wysa*(Study 31)[82]
Classical Machine Learning (ML)	2 (7%)	2 (6%)	*Chat robot*(Study 9)[83]; *SVM chatbot*(Study 10)[84]
Not reported	2 (7%)	2 (6%)	*AI chatbots*(Study 21)[85]; *HiBaby*(Study 25)[86]

NB: *Studies evaluating the same system are grouped. Classification follows the response-generation rule described in the Methods. Commercial Voice assistants in Study 3 were classified as unspecified AI because their response-generating architectures were not disclosed. Woebot's conversational responses are rule-based, although its input processing uses machine-learning intent classification; re-assignment rationales are provided in Supplementary Data S1, sheet 1.*

Model composition shifted over time. All six LLM systems and five of the seven hybrid systems appeared in 2022–2025, so 11 of the 28 studies published in that period (39%) used LLM or hybrid architectures (Fig. 3). Fourteen studies (40%) depended on third-party commercial application programming interfaces (APIs), including OpenAI, AWS, Kakao, Google/Dialogflow, Telegram and Twilio, routing conversational content, derived representations or service requests through external servers. Two systems (*MOMCare* and *DIAN*) used retrieval-augmented generation.

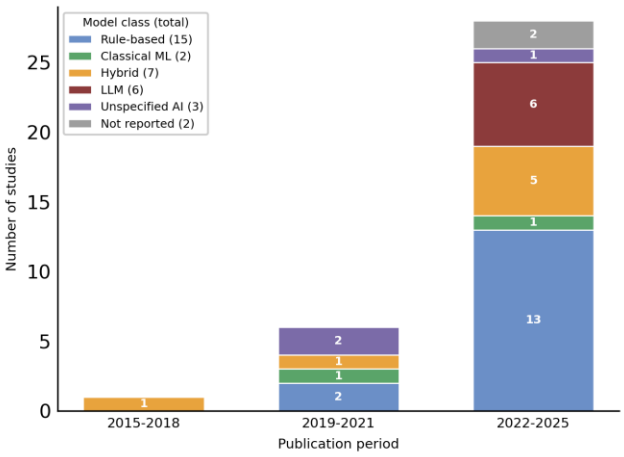

Fig. 3 | Publications over time by model class (N = 35 studies)

Clinical scope, functions, and therapeutic foundations

Depression was the most frequently targeted mental health condition, identified in 21 studies (60%), followed by wellbeing (16, 46%), anxiety (15, 43%) and stress (10, 29%); categories overlapped because individual systems could address more than one condition. When mental-

health targets were classified into mutually exclusive categories, multi-condition support was most common (16/35 studies, 46%), followed by depression (10/35, 29%), wellbeing (5/35, 14%), anxiety (2/35, 6%), stress (1/35, 3%) and other mental-health targets (1/35, 3%) (Fig. 4a). No included system explicitly targeted perinatal psychosis, obsessive-compulsive disorder, bipolar disorder, birth-related trauma/PTSD, substance use disorders, eating disorder, or perinatal loss and bereavement. Hypomania, however, was included among the mental-health targets in two studies. Stage coverage was relatively balanced across the 30 unique systems: 15 (50%) spanned both antenatal and postnatal periods, seven (23%) were antenatal-only and eight (27%) were postnatal-only. None of the eight postnatal-only systems reported a human-involving mental-health crisis pathway; however, such pathways were also not universal among systems covering the antenatal period or both stages.

Systems varied widely in their intended mental health applications, with many performing more than one clinical function. Informational support was most common (27/35 studies, 77%), followed by monitoring (22/35, 63%), treatment (16/35, 46%), screening (8/35, 23%), crisis detection or response (4/35, 11%), and decision support (1/35, 3%). No system performed a diagnostic function. Overall, 22 studies (63%) described systems performing at least one of screening, treatment, or crisis detection or response (Fig. 4b).

Therapeutic foundations also varied across the 30 unique systems. Hybrid approaches and psychoeducation were most common, each identified in nine systems (30%), followed by no explicit therapeutic foundation in seven (23%). Two systems (7%) were based primarily on cognitive behavioural therapy (CBT), two (7%) were scale-based, and one (3%) was mindfulness-based. No system used interpersonal psychotherapy (IPT) or acceptance and commitment therapy (ACT) as its sole therapeutic foundation, although these approaches were incorporated within hybrid systems (Fig. 4c).

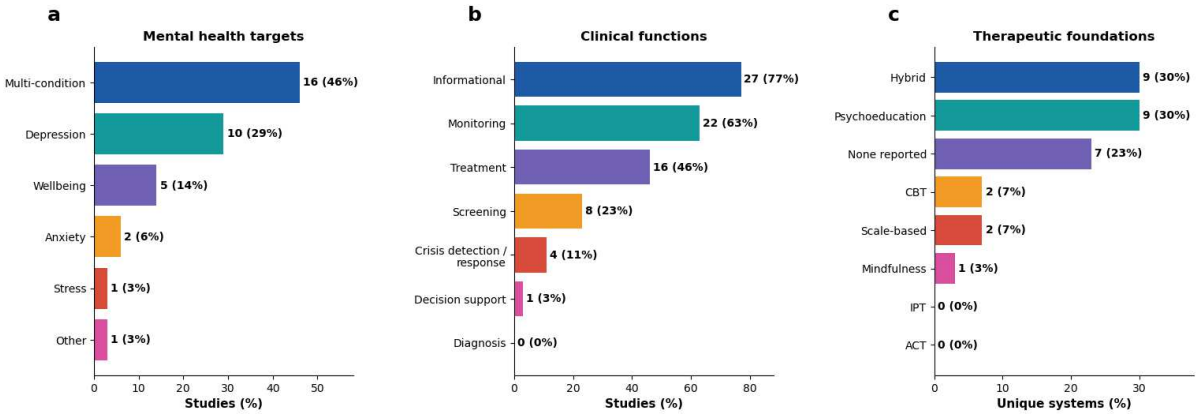

Fig. 4 | Clinical scope, functions, and therapeutic foundations of perinatal CAs. a*, Mental health targets across 35 studies.* ***b****, Clinical functions across 35 studies.* ***c****, Therapeutic foundations across 30 unique systems*

Privacy-preserving methods and data practices

No study reported advanced privacy-preserving computation, such as, federated learning, on-device inference, differential privacy, secure aggregation or homomorphic encryption (Table 2; Supplementary Data 1, sheet 3). Study 4 (the *PPD screener*) proposed on-device processing and stronger encryption as future safeguards but, as implemented, transmitted user input to a commercial cloud API. Anonymisation or pseudonymisation at the database level was reported in 13 studies (37%) and encryption in two (6%).

The 14 API-dependent studies (40%) processed conversational content on external servers, and the count rises to 15 (43%) when an authentication-only dependency on a national digital-identity

service (Study 26) is included. No system described on-device response generation. A data-storage location was stated in nine studies (26%), predominantly cloud-based; the remaining 26 did not report where data were stored.

Source code was openly available for one system (3%). The response-generating architecture could not be independently identified in five studies (three undisclosed commercial services, two with no reported model), and no study provided a model card or datasheet. Training-data or knowledge-base provenance was reported in 30 studies (86%), predominantly user-generated, synthetic, scraped or expert-authored content, with formal clinical datasets underpinning four systems.

Post-deployment change was assessed at two levels: model-level updating, involving retraining or learning from new data, and content or configuration updating, involving changes to prompts, rules, content or knowledge bases without modifying the underlying model. Model-level updating was reported in three studies (9%): two used continuous online learning and one used iterative model refinement. Two studies (6%) reported content or knowledge-base updating without model-level change. Eight studies (23%) reported a static model or decision mechanism, although two of these allowed content or configuration changes. Model-level updating was not applicable to one rule-based system with no trained model. The remaining 21 studies (60%) did not report an updating strategy.

Table 2 | Privacy-preserving measures, data practices, and transparency across included studies.

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

Population and representation

Four of the 30 systems addressed fathers (*Parentbot*, *NurtureBot*, *Dr. Joy*, and *Rover*, evaluated across six studies), leaving 26 systems (87%) focused exclusively on mothers (Supplementary Data 1, sheet 4). The depth of paternal inclusion varied widely. *Rover* was designed specifically for fathers, with father-specific content and male-voiced guidance; *Dr. Joy* enrolled two male participants. *Parentbot* recruited couples but adjusted for clustering with a couple identification number instead of modelling reciprocal partner effects. *NurtureBot* included both mothers and fathers but analysed participants independently. Representation of diverse parent and family types was also limited: 29 of 35 studies made no reference to non-birthing partners, sexual orientation or gender identity. *Parentbot* restricted participation to heterosexual couples, whereas *Rover*

explicitly included fathers regardless of biological parenthood, including fathers in same-sex relationships, but did not address non-birthing parents who did not identify as fathers. Across the included studies, no study modelled dyadic interdependence or estimated the effect of one partner's mental health on the other.

Nineteen systems (63%) operated exclusively in English. Infrastructure accessibility was similarly limited: *Zuri* was the only system that supported access without mobile broadband through SMS, allowing use on basic mobile phones. These representation and access constraints frequently overlapped. Across all the included studies, sixteen of the 35 studies (46%) simultaneously focused only on mothers, operated only in English, and offered no offline capability.

Development context also varied by architecture. Rule-based studies were almost exclusively from high-income settings (14 of 15), compared with 3 of 6 LLM and 4 of 7 hybrid studies; every system developed in a low-middle-income country used a hybrid or LLM architecture. Cultural adaptation was reported in 15 studies (43%) and did not differ substantially by model class. Two LLM systems delivered content in Turkish and Indonesian while building on base models trained predominantly on English-language data, and neither reported evaluating that base model for linguistic or cultural bias.

Crisis escalation

The crisis escalation was classified at the study level according to the highest documented tier (Table 3). Ten studies (29%) had no escalation mechanism among them, and in several cases the authors stated the absence explicitly; a further seven (20%) did not address crisis provision at all, so its presence could not be determined form the publication. The remaining tiers, from static disclaimers to automated detection without human follow-up.

Three studies (9%) paired automated detection with a human-response pathway; they differed in whether human follow-up was guaranteed and whether a response timeframe was specified. *Rosie* specified follow-up by a clinical social worker within 24 hours. *Zuri* alerted a live team member without an assured response time. The *Pregnancy Companion* notified a doctor and ambulance for general medical emergencies; mental health crisis language did not trigger escalation. Of the three, two were specific to mental health crises and one specified a defined response timeframe.

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 and sum to 35 studies. Human-in-the-loop refers to referral or transfer to a clinician or midwife without automated crisis detection. General-health referral pathways not specific to mental-health crises are identified in Supplementary Data 1, sheet 5.

Nine unique systems, evaluated across ten studies, reported automated risk or emergency detection. None reported the sensitivity, specificity, or false-negative rate of that detection for crisis language. Reported figures described related but distinct tasks: *e-Midwife*, for example, reported 97.9% accuracy for risk-keyword classification but no crisis-detection sensitivity or false-negative performance.

Governance maturity

Governance Maturity Scores ranged from 2 to 14 of 24 (median 7; interquartile range, IQR, 5–10.5; mean 7.5). Four studies (11%) fell within the descriptive None range (0–3), 16 (46%) within Basic (4–7), 10 (29%) within Moderate (8–11), and five (14%) within Advanced (12–14). No study reached the Comprehensive range (15–24). The highest score of 14 was shared by *Zuri*, the *ALBA* feasibility study and *Rosie*; *Maia* scored 13 and the *Parentbot* effectiveness trial 12. These five Advanced scores represented five distinct systems, none of which reported a privacy-preserving computational architecture or an empirical fairness evaluation (Fig. 5).

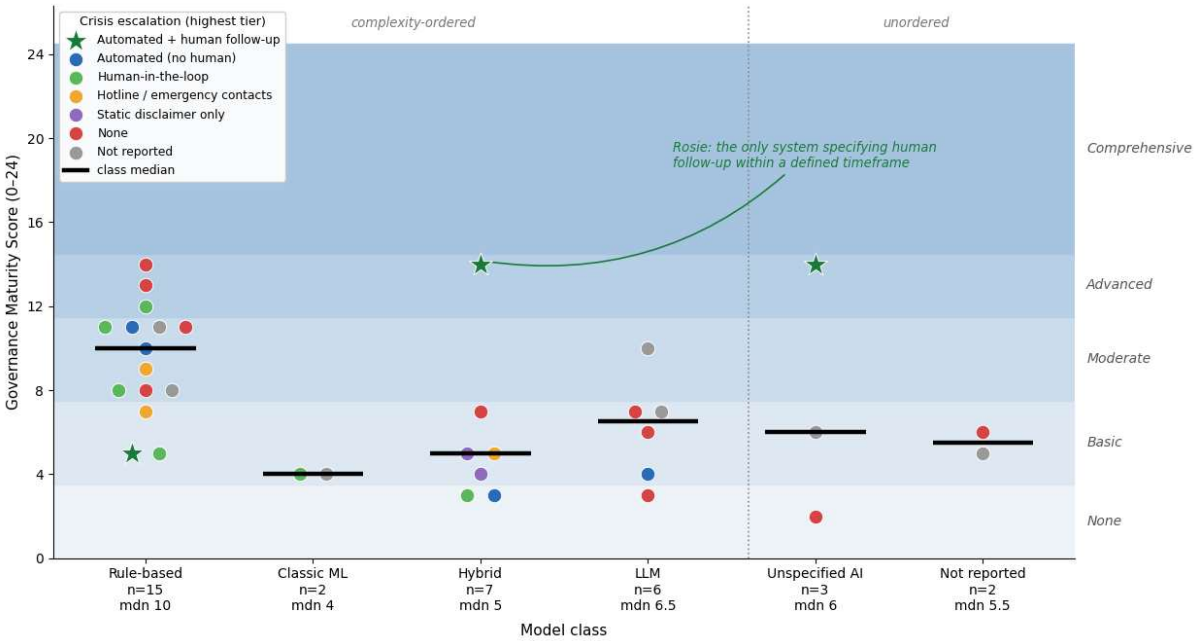

Fig. 5 | Governance maturity by model class and crisis-escalation tier. Each marker represents one study, coloured by its highest documented crisis-escalation tier.

Across the six dimensions (Fig. 6), mean scores were highest for bias and fairness (2.00 of 4) and safety (1.57), followed by consent (1.34), regulation (1.23), privacy (0.89) and monitoring (0.46). Bias and fairness were the only dimension with a mean reaching 2.0; all other dimension means remained below this level. The full per-class breakdown is provided in Supplementary Data 1, sheets 6 and 7.

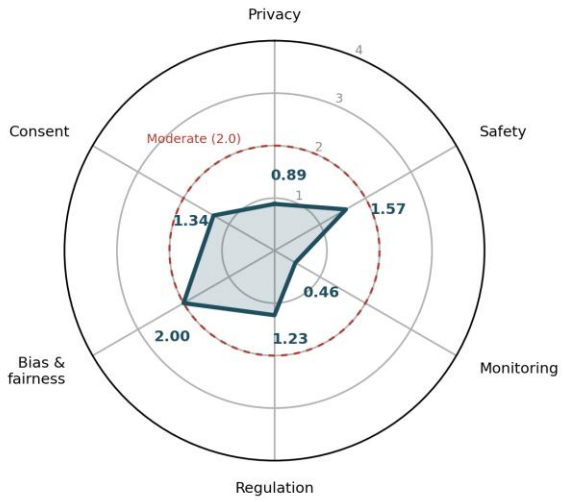

***Fig. 6 | Governance Maturity Profile across the six GMS dimensions*.** Mean scores across 35 studies, with the dashed ring denoting a score of 2.0 on the 0–4 scale.

Explicit informed 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 (17%) cited a named data-protection regulation (the GDPR in five, the Australian Privacy Act 1988 in one). Six systems disclosed their AI identity in-chat, and one stated explicitly that the system was not human. Clinical supervision was absent in 23 studies (66%), consultative in nine and not reported in three; no study reported continuous clinical review of conversational outputs during deployment.

Governance maturity and reporting by model class.

Among the 30 studies in the four defined model classes, GMS distributions differed across classes (Kruskal–Wallis H = 11.60, p = 0.009; Fig. 5). Rule-based studies recorded a median score of 10 of 24 (IQR 8–11), compared with 6.5 for LLM studies (IQR 4.5–7), 5 for hybrid studies (IQR 3.5–6) and 4 for the two classical-machine-learning studies; the combined LLM-and-hybrid group had a median of 5 (IQR 4–7). The primary comparison of rule-based studies and the combined LLM-and-hybrid group was significant (Mann–Whitney U = 161.5, p = 0.003; rank-biserial r = 0.66).

The model-class contrast remained across four sensitivity analyses. After collapsing systems evaluated in more than one study to single value per system, governance still differed across classes (n = 25; Kruskal–Wallis H = 9.86, p = 0.020; rule-based versus LLM-and-hybrid U = 109.0, p = 0.007). When restricted to studies published in 2022–2025, the medians were 11.0 versus 6.0 of 24 (U = 119.5, p = 0.006). Using the raw extracted classes without the response-generation rule, the medians were again 10.0 versus 6.0 (U = 141.0, p = 0.013; rank-biserial r = 0.57). When conference publications were excluded, the corresponding medians were 10.5 and 7.0 among peer-reviewed journal studies (U = 79.5, p = 0.024; rank-biserial r = 0.62). Across all four sensitivity analyses, rule-based studies retained higher GMS scores than the combined LLM-and-hybrid group.

Governance differences were also visible in the underlying indicators (Fig. 7; Supplementary Data 1, sheet 8). Rule-based studies more often documented explicit consent than LLM studies (11/15 versus 3/6), a privacy measure at rest (10/15 versus 1/6), clinical supervision (6/15 versus 0/6) and a human-involving crisis pathway (5/15 versus 0/6). The pattern was not uniform: LLM studies more often addressed bias or fairness (4/6, versus 9/15 rule-based and 3/7 hybrid studies) and reported a data-storage location (3/6 versus 4/15), while AI-identity disclosure was more frequent in hybrid (3/7) and LLM (2/6) studies than in rule-based studies (4/15). Reporting 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.

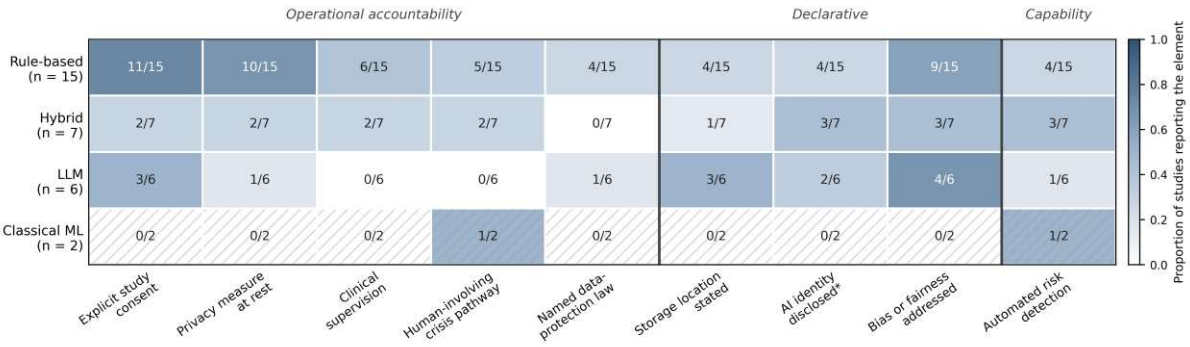

Fig. 7 | Governance reporting by model class. *AI-identity disclosure counts partial disclosure, where the system signals non-human status without an explicit statement that it is AI.

Evaluation and evidence

Eight studies (23%) evaluated systems through simulation without real users and two (6%) in laboratory settings. Fifteen studies (43%) evaluated systems in the home or another real-world context, one in a clinical setting and seven in mixed settings; two did not report the evaluation setting (Supplementary Data 1, sheet 9). Ten studies (29%), describing seven unique systems, reported clinical outcomes in actual users. The remainder reported usability, feasibility, acceptability, engagement or technical performance.

Five studies were primary reports of randomised controlled trials, and two further studies (Studies 1 and 30) reported secondary or summary analyses of completed trials. Where between-group effects were reported, they were small and inconsistent. The largest trial reported a small between-group increase in parenting self-efficacy at one month (mean difference 1.22, 95% confidence interval, CI, 0.06 to 2.39; d = 0.14). It also found a small reduction in maternal state anxiety at three months (mean difference −2.21, 95% CI −4.18 to −0.24; d = −0.22), with no significant effect on depression or stress. No study reported a health-economic or cost-effectiveness outcome, and none of the simulation-only systems had corresponding real-user evaluation.

Cross-cutting design tension

Six recurrent design tensions were identified across 28 of the 35 studies (80%) (Table 4; Supplementary Table 1). Safety versus autonomy was the most frequently observed. 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 4 | 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
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

This review shows that the rapid evolution of AI-supported CAs for PMH has not been consistently matched by equivalent development in governance, privacy, safety, and inclusive design. Across 35 studies describing 30 unique systems, rule-based architectures remained common, while LLM and hybrid systems emerged rapidly after 2022 and expanded the possibilities for flexible, personalised interaction. Twenty-two studies described screening, treatment, or crisis-related functions, showing that many systems are already operating in clinically consequential spaces. Yet no included study reported advanced privacy-preserving computation, and only one linked automated mental-health crisis detection to human follow-up within a defined timeframe. Fathers, couples, non-birthing partners, linguistically diverse populations, and users in low-connectivity settings remained poorly represented. These findings indicate a persistent mismatch between increasing technical capability and the safeguards, oversight and evidence required for responsible use.

The clearest expression of this mismatch emerged in the relationship between model class and documented governance. Rule-based studies more often reported explicit consent, privacy measures, clinical supervision, and human-involving crisis pathways than LLM studies, while newer architectures showed greater or comparable attention to selected indicators of bias, fairness, data-storage transparency, and AI-identity disclosure. The review-specific GMS reflected the same pattern, with a median score of 10 of 24 for rule-based studies compared with 5 for the combined LLM-and-hybrid group. This contrast persisted across four sensitivity analyses, including after collapsing multiple studies of the same system, restricting the publication period, using the raw model classifications, and excluding conference publications. We describe this architecture-governance contrast as the *AI perinatal paradox*: conversational systems are becoming more flexible and technically sophisticated, yet several operational safeguards central to responsible use remain less consistently implemented or reported[46,87].

The paradox describes an observed association and does not imply that rule-based systems are intrinsically safer or that LLMs weaken governance. Several factors may help explain the pattern. Rule-based systems were often described within clinical or research programmes in which ethics approval, consent procedures, and referral pathways were already embedded, while LLM-and-hybrid studies more often emphasised conversational performance, empathy, and personalisation. Variation within model classes also shows that architecture alone does not determine governance maturity. *Rosie* was among the highest-scoring hybrid systems, while both classical machine-learning prototypes scored poorly despite their lower technical complexity. Translational maturity may contribute, as all LLM studies were published after 2022 and many remained at formative or technical stages of evaluation. Yet the contrast persisted when analyses were restricted to studies published from 2022 to 2025, suggesting that publication timing alone does not explain the pattern. Hua et al. similarly found that only about 16% of LLM-based studies across 160 general mental-

health chatbot studies had reached clinical-efficacy testing[24]. These findings indicate that technical sophistication and governance maturity are distinct dimensions of progress[43,46,87].

The clinical functions assigned to these systems make this distinction consequential. Although many systems performed screening, treatment or crisis-related functions, seven of 30 unique systems reported no explicit therapeutic foundation. None performed diagnosis, but diagnostic labelling is not required for a system to influence care. Screening outputs may shape symptom recognition and help-seeking, therapeutic content may influence self-management, and responses to distress may affect subsequent care. Governance, evidence and oversight should reflect what a system is designed to do and the consequences if it fails, not the sophistication of its conversational interface alone.

Data practices revealed another important governance gap. Fourteen studies (40%) relied on third-party commercial APIs for conversational processing, while data-storage location was reported in only nine studies and encryption in two. No included system reported federated learning, on-device response generation, differential privacy, secure aggregation or homomorphic encryption. Source code was openly available for one system, no study provided a model card or datasheet, and most did not clearly report whether the response-generating model or knowledge base changed after deployment. Training-data and knowledge-source provenance were also inconsistently documented, limiting assessment of what data and evidence informed system behaviour and whether those sources were appropriate for the populations served. Regulatory reporting was similarly sparse, with only six studies citing a named data-protection regulation. These gaps do not establish inappropriate data handling because the review assessed published reporting, not deployed infrastructure, but they limit independent assessment of privacy, reproducibility, and accountability. Such omissions also raise questions about whether these considerations are routinely addressed in system design and reporting.

The consequences are particularly important in PMH, where conversations may involve suicidal ideation, intimate-partner violence, substance use, pregnancy or thoughts of infant harm[30]. Perinatal women already withhold such disclosures from clinicians for fear of being judged unable to cope or of child-protection involvement[88]. Exposure of this content, or the perception that it may be exposed, therefore carries two clinical risks: direct harm if disclosures reach third parties, including an abusive partner with access to a shared device or account[89], and indirect harm if parents censor what they disclose, so that the risk signals a system depends on to trigger escalation are never expressed. Anonymisation of stored records does not remove the risks created when user inputs, conversational content or derived representations cross a network boundary to an external provider, as occurred in 14 of the 35 included studies. Published privacy statements are also an unreliable guide to actual practice: in an audit of top-ranked depression and smoking-cessation apps, 33 of 36 transmitted data to third parties, yet fewer than half of those sending data to Google disclosed it[32]. The current evidence does not support recommending perinatal CAs to parents on the basis of stated confidentiality alone, since the information clinicians would need to judge data handling are usually unavailable. Until data flows are independently verified, clinicians recommending these systems should discuss the limits of confidentiality with parents beforehand, 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, how access is controlled, and whether information is reused[32]. These arrangements should be reassessed when changes to the model, content, provider or processing infrastructure alter the underlying data flow. Systems built on third-party foundation models may also change through provider-initiated updates outside the deployer's control. Behavioural variation between versions

of the same model has been documented[90], raising the possibility that safety or performance established during evaluation may not persist after an update. Transparency should also extend directly to users. Where sensitive conversational data are processed by external providers, users should be clearly informed about these data flows and, where feasible, given meaningful control over retention, secondary use, and deletion of their data. Such information should be presented at the point of use instead of being confined to general privacy policies or terms of service.

The absence of privacy-preserving computation reflects limited adoption, not a lack of technical alternatives. On-device processing, federated learning and differential privacy have demonstrated feasibility in related mental-health applications[41,91]. Their suitability will vary by architecture and use case, but systems handling sensitive disclosures should minimise unnecessary data movement and define data flows, retention, access controls, and external dependencies before deployment.

The crisis findings provide a high-stakes example of the gap between technical safeguards and effective crisis response. Nine unique systems reported automated risk or emergency detection, yet none reported sensitivity, specificity or false-negative performance for crisis language. Three studies paired automated detection with a human-response pathway, but only one specified mental-health follow-up by a human within a defined timeframe. None of the postnatal-only systems reported a human-involving mental-health crisis pathway. Recent studies have similarly questioned the reliability of LLM responses to suicide-related and other high-risk disclosures[28,92,93]. This creates an important trade-off: in lower-risk or subclinical populations, automated detection and escalation may strengthen safety where no crisis pathway currently exists, provided appropriate safeguards are in place. For higher-risk populations, however, fully automated responses are harder to justify. An automated detector with an unknown false-negative rate may create an appearance of safety without showing that crisis disclosures will be recognised and acted upon reliably. Guardrails and emergency links alone do not establish an end-to-end safety pathway. Consistent with recent editorial guidance[29], meaningful human involvement requires a defined escalation threshold, clarity about what is monitored and by whom, a qualified responder, a response timeframe and a contingency when contact fails. 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.

Population inclusion revealed a governance concern, not only a coverage gap. Bias and fairness cannot be addressed through model performance alone when systems are not designed and evaluated across the populations, languages, and cultural contexts they are intended to serve. This concern was evident across several dimensions. Fathers remained peripheral to system development: only 4 of 30 systems addressed them, and only *Rover* was designed specifically around paternal PMH. *Rover* extended beyond biological fatherhood to fathers in same sex relationships, but not other non-birthing parents. The absence of dyadic modelling is important given the interdependence of maternal and paternal mental health across the perinatal period[4]. Recruiting both partners does not capture this interdependence unless reciprocal effects are represented in the system design or analytic model. Linguistic and infrastructural reach were similarly constrained. Nineteen systems (63%) operated exclusively in English, none supported open-ended multilingual conversation, and only one used an SMS-based channel suited to lower-bandwidth settings; LLM and hybrid systems commonly depended on networked cloud infrastructure. Cultural adaptation, where reported, also did not coincide with stronger safety infrastructure, as these systems were generally rule-based and lacked crisis escalation (Supplementary Data 1, sheets 4 and 5). Higher-risk and specialised clinical needs were largely absent from system development, including perinatal psychosis, perinatal obsessive-compulsive disorder, birth-related trauma, and perinatal loss and bereavement. These patterns show that scalability does not ensure inclusion. This concern is particularly consequential for complex and adaptive systems, where poor representation may affect the relevance, interpretation, and safety

of responses across settings. Inclusion should be treated as a core governance and design consideration shaping language, delivery, therapeutic content, escalation and evaluation.

The six cross-cutting design tensions show that these challenges extend beyond a simple choice between rule-based and generative architectures. Greater conversational flexibility, personalisation and automation can increase demands on monitoring, data use and human oversight. These tensions are design choices, not technological inevitabilities: privacy-preserving approaches, human escalation and explicit AI-identity disclosure can mitigate some of these trade-offs. The central design question is whether architecture, data flow and oversight are proportionate to intended function, population, infrastructure and clinical risk.

The limited evaluation evidence makes these decisions more important. Eleven studies enrolled no human end-users, and only ten studies describing seven unique systems reported clinical outcomes in real-world users. Where comparative effects were reported, benefits were small or inconsistent. 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 increasing use of LLMs and simulation-based evaluation makes this distinction especially important because conversational competence can be demonstrated before clinical benefit has been established. Evidence requirements should scale with the claim being made. Informational systems may primarily require comprehension, usability and safety testing, while systems claiming treatment benefit, triage capability or crisis support require prospective human-user evaluation with clinically meaningful outcomes and explicit harm monitoring.

Viewed across all six governance dimensions, maturity was low and uneven. GMS values ranged from 2 to 14 of 24, with a median of 7 in the descriptive Basic range, and no study reached the Comprehensive range. Even the highest-scoring studies did not report a privacy-preserving computational architecture or empirical fairness evaluation, showing that higher relative maturity did not imply complete governance. The highest scores reflected clearer implementation and reporting of practices such as consent, crisis escalation and regulatory engagement, not greater model complexity. The GMS is a review-specific descriptive framework and should not be interpreted as a validated clinical, regulatory, safety or risk-assessment instrument. Its value lies in making governance conditions visible alongside conventional technical and clinical performance measures. The same broad architecture-related pattern was evident in the underlying governance indicators, indicating that the observed contrast was not created by the composite score alone. External evaluation is needed before the GMS is used beyond this review.

These findings have direct implications for developers, clinicians, and policymakers. 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 evidence proportionate to the system's clinical consequences, with stronger expectations for privacy, transparency, monitoring and human oversight for treatment, triage and crisis-related roles. Certification can support this, but only if it independently verifies how systems handle data. An assessment of apps accredited by the NHS Health Apps Library found widespread unencrypted transmission of identifying information that developer self-declaration had not detected[94]. Certification of perinatal CAs, for example as an extension of Australia's National Safety and Quality Digital Mental Health Standards[95], should include technical verification of data flows and recertification when the underlying model or provider changes. These priorities align with emerging regulatory expectations for AI in healthcare[96,87,97],

and support broader health-system governance of responsible AI, including accountability, transparency, human oversight, and equitable implementation in public health services[98].

Box 1 translates the recurrent gaps into eight proposed minimum safety and governance standards. No included study documented fulfilment of all eight. The standards require prospective evaluation and should be applied proportionately to intended function, clinical risk, and deployment setting.

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.

Future research should move beyond demonstrating that perinatal conversational agents can produce acceptable or empathetic responses. Crisis-related studies should report sensitivity, false-negative performance, completed escalation and time to human response in real users. Privacy-preserving approaches should be implemented and evaluated alongside usability, model performance and clinical outcomes, with data flows, retention practices and model updates reported transparently. The architecture-governance contrast should be independently replicated using prospectively specified governance criteria and comparable reporting standards. Future systems should involve fathers, non-birthing parents, couples, linguistically diverse populations and low-connectivity settings from design through evaluation. The proposed standards also require prospective testing to determine their feasibility and whether their adoption is associated with safer, more equitable, and clinically effective systems.

Several limitations should inform interpretation. The review assessed governance as documented in publications and cannot establish that an unreported safeguard was absent from the deployed system. Source code was unavailable for almost all systems, preventing independent verification through code audit, penetration testing or network inspection. Restriction to English-language publications and exclusion of commercial white papers and unreviewed preprints may underrepresent 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. Formal critical appraisal was not undertaken because the review aimed to map a heterogeneous evidence base, which limits conclusions about the methodological quality of individual effectiveness studies. Data extraction was completed by one reviewer with independent verification of a subset, departing from the registered plan for full dual extraction. The GMS was developed after protocol registration; its equal weighting and maturity bands are pragmatic analytical choices, and the inferential comparisons are exploratory. Independent

rescoring of a subset supports within-review reproducibility but does not establish external validity. The design-tension synthesis was interpretive, and cost-effectiveness, business models and disability-specific accessibility were not systematically assessed because reporting was sparse. Two conference papers analysed the same 223-record dataset; collapsing repeated systems in sensitivity analysis did not alter the overall conclusions.

CAs may extend access to perinatal mental-health information, screening, and support, but technical capability alone does not establish responsible use. Across this evidence base, advances in conversational architecture coexisted with limited privacy protection, weak crisis validation, narrow population coverage, limited clinical oversight and sparse evidence of patient benefit. The opportunity to introduce safeguards becomes more constrained as technologies become embedded in practice, reflecting the dilemma described by Collingridge four decades ago[99]; that window is narrowing as conversational AI moves closer to routine care. Governance-first development offers a useful organising principle for the field. Safety, equity, privacy and clinical validation should be specified from the outset and evaluated throughout the system lifecycle. Whether a person disclosing a crisis reaches qualified human support depends on choices made before deployment. By making those choices visible and comparable, this review provides a basis for more accountable development. Perinatal conversational AI should be safe before deployment, inclusive by design, privacy-preserving by default, and clinically validated in proportion to its intended role before it is trusted at scale

Methods

We conducted a scoping review following JBI methodology[100] and reported it in accordance with PRISMA-ScR checklist[51] (Supplementary Table 2). The protocol was prospectively registered on the Open Science Framework (https://osf.io/63mpb)[101].

Search strategy

We searched nine databases spanning health, psychology, computing, and engineering: MEDLINE, PsycINFO, Emcare, IEEE Xplore, ACM Digital Library, CINAHL, Scopus, Web of Science and Cochrane CENTRAL. Searches were conducted on 5 November 2025 and covered January 2015 to September 2025. Key studies identified by the review team informed the initial MEDLINE strategy, which combined three concept blocks: the perinatal period, mental health, and AI-supported conversational agents. The strategy was refined with an expert librarian (B.L.) and reviewed by our multidisciplinary team, with expertise in applied AI, digital mental health, and psychology, and adapted to each database using controlled vocabulary and free-text terms. Database searches were supplemented by backward and forward citation tracking and a grey-literature search using the same core concepts. Full search strategies are provided in Supplementary Table 3.

Eligibility criteria

Eligibility was defined using the PCC (Population, Concept, Context) framework[50]. The population comprised individuals during pregnancy and up 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 conversational agents enabling interactive natural-language dialogue and explicitly addressing perinatal mental health or wellbeing. The context included clinical, real-world, applied 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. Evaluations without human

participants were eligible where they reported technical or architectural characteristics. We excluded non-conversational systems, static screening or tracking applications, studies without an explicit perinatal mental-health focus, registered protocols without methodological content or results. We also excluded reviews, editorials, commentaries, unreviewed preprints, commercial white papers, non-English publications, abstracts, posters, and reports without accessible full text. Full eligibility criteria are provided in Supplementary Table 4.

Study selection

All retrieved records were imported into Rayyan software for deduplication and screening. Two reviewers (D.O.O. and A.B.R.S.) independently screened titles, abstracts and full texts against the eligibility criteria, with agreement of 87.3% and 89.5%, respectively. Disagreements at both stages were resolved by consensus. One report remained unavailable despite thorough searches and author contact via ResearchGate. 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. Reasons for full-text exclusion are summarised in Fig. 1.

Data extraction and verification

A standardised extraction form was developed from the review objectives and preliminary reading, piloted on five randomly selected studies and refined by D.O.O. and A.B.R.S. The final form comprised 92 variables across eight domains: (1) study context; (2) system and clinical scope; (3) population representation; (4) safety and risk; (5) conversational design and architecture; (6) data governance, privacy and accessibility; (7) evaluation evidence; and (8) design tensions, limitations, and summary; the full codebook is provided in Supplementary Data 1, sheet 10.

D.O.O. extracted data from all included studies, with A.B.R.S. independently verifying a randomly selected studies 11%; no discrepancies were identified. Ambiguities arising during extraction were resolved in weekly meetings. This departed from the registered plan of full dual extraction owing to resource constraints. Prespecified coding rules for clinical outcomes, language, and other variables are provided in Supplementary Data 1, sheet 10. Crisis responses were classified by the highest reported level of escalation, ranging from no reported mechanism to automated detection with human follow-up within a defined timeframe (Supplementary Data 1, sheet 5).

Unit of analysis and model classification

The 35 studies described 30 unique CAs. System-level characteristics were reported per unique agent (n = 30), whereas evaluation-specific characteristics, such as governance, crisis escalation, privacy practices, and study design, were reported per study (N = 35). Systems were classified by the architecture generating the user-facing response as rule-based, classical machine learning, LLM, hybrid, unspecified AI or not reported; input parsing and post hoc transcript analysis did not determine model class. Where multiple publications described the same system, classification used the fullest technical description (Supplementary Data 1, sheet 1).

Governance Maturity Score (GMS)

The GMS assessed six dimensions of reported governance: privacy, safety, monitoring, regulation, bias and fairness, and consent. Each dimension was scored from 0 (absent) to 4 (comprehensive) using predefined criteria informed by health-AI governance frameworks[44,43], WHO guidance[98], the perinatal mental-health literature[4,12], and requirements for special-category health data[96]. Dimension scores were summed to 0–24 and descriptively classified as None (0–3), Basic (4–7), Moderate (8–11), Advanced (12–14), and Comprehensive (15–24) (Table 5). Inferential analyses used the continuous total score. The GMS was designed to enable structured comparison of reported governance practices across studies.

Table 5 | Governance Maturity Score: 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
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

Scores were assigned at the study level, reflecting the governance documented within a given publication. A.B.R.S. independently re-scored 12 of the 33 studies eligible for blinded re-scoring (36%), selected at random, across all six dimensions. Studies 27 and 35 were excluded owing to A.B.R.S.'s co-authorship. Before consensus, reviewers agreed exactly on 51 of 72 dimension-level rating (71%), with 69 of 72 (96%) differing by no more than one point, and 9 of 12 studies (75%) were assigned to the same governance band. In each of the three studies with discordant bands, differences in dimension-level ratings produced totals one point apart at the Basic–Moderate threshold (7 versus 8). These cases were reviewed jointly to reach consensus, which did not change the band originally assigned by D.O.O. (Supplementary Data 1, sheets 11 and 12)

Minimum safety standards

Eight proposed minimum safety and governance standards were derived from the NIST AI Risk Management Framework[102], WHO guidance[98], FDA clinical decision support guidance[103], EU AI Act transparency requirement[104], GDPR requirements for special-category data[40,105] and practices identified across the included studies. Each unique system was assessed against predefined criteria (Box 1).

Critical appraisal

Formal critical appraisal was not conducted, consistent with PRISMA-ScR[51] and JBI guidance[100] for scoping reviews.

Data synthesis and analysis

We summarised the charted data using descriptive statistics and structured narrative synthesis. Counts, proportions, and tabulations examined patterns by model class, translational stage, development-country income level, perinatal stage and paternal inclusion; medians and IQRs were computed with linear interpolation of quartiles. Cross-cutting design tensions were synthesised inductively from reported design choices, implementation constraints, and study limitations using constant comparison. D.O.O. developed and iteratively refined the categories against the source publications, merging overlapping concepts and retaining tensions that recurred across multiple studies.

GMS totals were compared across rule-based, classical machine-learning, LLM, and hybrid studies using the Kruskal–Wallis test. The primary contrast between rule-based and combined

LLM-and-hybrid studies was assessed using the Mann–Whitney U test with rank-biserial correlation as the effect-size measure. Non-parametric tests were used because GMS totals are bounded scores derived from ordinal dimension ratings and group sizes were small and unequal. Robustness was examined in four sensitivity analyses: collapsing multiply-evaluated systems to one value per system, restricting to studies published 2022–2025, using raw extracted classes without the response-generation rule, and excluding conference publications. Given tied observations, Mann–Whitney *p* values used the asymptotic normal approximation and Kruskal–Wallis p values used the asymptotic chi-squared reference distribution because of tied observations. Analyses were conducted in Python (v3.11) using SciPy (v1.11.4)[106], with two-sided $p < 0.05$ considered statistically significant (Supplementary Data 1, sheets 13 and 14). These comparisons are exploratory and provide supporting evidence for the observed patterns.

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)

Human Ethics and Consent to Participate declaration

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

Acknowledgements

This work forms part of the first author's PhD research at James Cook University. We thank Benton Lewis (B.L.), Liaison Librarian, James Cook University, for expert advice in developing and refining the search strategy. 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.

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.

Competing interests

S.J.T. and A.B.R.S. are authors of two included *Rover* study (Studies 27 and 35). To manage this potential conflict, D.O.O. independently assessed eligibility, extracted the data, and assigned the GMS using the pre-defined criteria. Both studies were excluded from the blinded inter-rater subset. The remaining authors declare no competing interests..

References

1 Howard, L. M. & Khalifeh, H. Perinatal mental health: a review of progress and challenges. *World Psychiatry* **19**, 313–327 (2020). https://doi.org/10.1002/wps.20769

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

3 Dennis, C.-L., Falah-Hassani, K. & Shiri, R. Prevalence of antenatal and postnatal anxiety: Systematic review and meta-analysis. *British Journal of Psychiatry* **210**, 315–323 (2017). https://doi.org/10.1192/bjp.bp.116.187179

4 Paulson, J. F. & Bazemore, S. D. Prenatal and postpartum depression in fathers and its association with maternal depression: a meta-analysis. *Jama* **303**, 1961–1969 (2010). https://doi.org/10.1001/jama.2010.605

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

6 Bertram, G., Becuzzi, N. & Wilson, E. Trends in Perinatal Mental Health Within a Tertiary Institution From 2013 to 2022. *The Australian & New Zealand Journal of Obstetrics & Gynaecology* **63**, 821–824 (2023). https://doi.org/10.1111/ajo.13720

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

8 O'Dea, G. A. *et al.* Associations between maternal psychological distress and mother-infant bonding: a systematic review and meta-analysis. *Arch Womens Ment Health* **26**, 441–452 (2023). https://doi.org/10.1007/s00737-023-01332-1

9 Slomian, J., Honvo, G., Emonts, P., Reginster, J. Y. & Bruyère, O. Consequences of maternal postpartum depression: A systematic review of maternal and infant outcomes. *Womens Health (Lond)* **15**, 1745506519844044 (2019). https://doi.org/10.1177/1745506519844044

10 Rogers, A. *et al.* Association Between Maternal Perinatal Depression and Anxiety and Child and Adolescent Development: A Meta-analysis. *JAMA Pediatr* **174**, 1082–1092 (2020). https://doi.org/10.1001/jamapediatrics.2020.2910

11 Stein, A. *et al.* Effects of perinatal mental disorders on the fetus and child. *The Lancet* **384**, 1800–1819 (2014). https://doi.org/10.1016/S0140-6736(14)61277-0

12 Khalifeh, H., Hunt, I. M., Appleby, L. & Howard, L. M. Suicide in perinatal and non-perinatal women in contact with psychiatric services: 15 year findings from a UK national inquiry. *The Lancet Psychiatry* **3**, 233–242 (2016). https://doi.org/10.1016/S2215-0366(16)00003-1

13 Felker, A., Patel, R., Kotnis, R., Kenyon, S. & Knight, M. 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 Report No. 978-1-0687913-8-3, (National Perinatal Epidemiology Unit, University of Oxford., Oxford, 2025).

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

15 Fisher, J. *et al.* Prevalence and determinants of common perinatal mental disorders in women in low- and lower-middle-income countries: a systematic review. *Bull World Health Organ* **90**, 139g–149g (2012). https://doi.org/10.2471/blt.11.091850

16 Sawyer, A., Ayers, S. & Smith, H. Pre- and postnatal psychological wellbeing in Africa: a systematic review. *J Affect Disord* **123**, 17–29 (2010). https://doi.org/10.1016/j.jad.2009.06.027

17 Cribb, E., Spring, B., Galbally, M., Gardiner, F. & Coleman, M. Perinatal psychiatric emergencies in rural Australia: In urgent need of attention. *Australian & New Zealand Journal of Psychiatry* **59**, 109–114 (2025). https://doi.org/10.1177/00048674241308689
18 Sambrook Smith, M., Lawrence, V., Sadler, E. & Easter, A. 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* **9** (2019). https://doi.org/10.1136/bmjopen-2018-024803
19 Dennis, C. L. & Chung-Lee, L. Postpartum depression help-seeking barriers and maternal treatment preferences: a qualitative systematic review. *Birth* **33**, 323–331 (2006). https://doi.org/10.1111/j.1523-536X.2006.00130.x
20 Webb, R. *et al.* Barriers and facilitators to implementing perinatal mental health care in health and social care settings: a systematic review. *The Lancet Psychiatry* **8**, 521–534 (2021). https://doi.org/10.1016/S2215-0366(20)30467-3
21 Zhong, S. *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. *JMIR Public Health Surveill* **9**, e43730 (2023). https://doi.org/10.2196/43730
22 Li, H., Zhang, R., Lee, Y.-C., Kraut, R. E. & Mohr, D. C. Systematic review and meta-analysis of AI-based conversational agents for promoting mental health and well-being. *npj Digital Medicine* **6**, 236 (2023). https://doi.org/10.1038/s41746-023-00979-5
23 He, Y. *et al.* Conversational Agent Interventions for Mental Health Problems: Systematic Review and Meta-analysis of Randomized Controlled Trials. *J Med Internet Res* **25**, e43862 (2023). https://doi.org/10.2196/43862
24 Hua, Y. *et al.* Charting the evolution of artificial intelligence mental health chatbots from rule-based systems to large language models: a systematic review. *World Psychiatry* **24**, 383–394 (2025). https://doi.org/10.1002/wps.21352
25 Weizenbaum, J. ELIZA—a computer program for the study of natural language communication between man and machine. *Commun. ACM* **9**, 36–45 (1966). https://doi.org/10.1145/365153.365168
26 Sanjeewa, R., Iyer, R., Apputhurai, P., Wickramasinghe, N. & Meyer, D. Empathic Conversational Agent Platform Designs and Their Evaluation in the Context of Mental Health: Systematic Review. *JMIR Ment Health* **11**, e58974 (2024). https://doi.org/10.2196/58974
27 Amil, S. *et al.* Interactive Conversational Agents for Perinatal Health: A Mixed Methods Systematic Review. *Healthcare* **13**, 363 (2025).
28 Olisaeloka, L., Richardson, C. G., Wang, A. Y., Munthali, R. J. & Vigo, D. V. Safety Mechanisms and Risk Mitigation in Generative AI Mental Health Chatbots: A Systematic Scoping Review. *Healthcare (Basel)* **14** (2026). https://doi.org/10.3390/healthcare14101395
29 Birk M, K. S., Myrick K, Schueller S, Torous J. Digital Mental Health Research Priorities, Revisited for the AI and Large Language Model Era. *JMIR Ment Health* **13** (2026). https://doi.org/10.2196/104118
30 Austin, M.-P. V. *et al.* Disclosure of sensitive material at routine antenatal psychosocial assessment: The role of psychosocial risk and mode of assessment. *Women and Birth* **35**, e125–e132 (2022). https://doi.org/10.1016/j.wombi.2021.04.005
31 Khalil, S. S., Tawfik, N. S. & Spruit, M. Federated learning for privacy-preserving depression detection with multilingual language models in social media posts. *Patterns* **5**, 100990 (2024). https://doi.org/10.1016/j.patter.2024.100990
32 Huckvale, K., Torous, J. & Larsen, M. E. Assessment of the Data Sharing and Privacy Practices of Smartphone Apps for Depression and Smoking Cessation. *JAMA Network Open* **2**, e192542–e192542 (2019). https://doi.org/10.1001/jamanetworkopen.2019.2542

33 Ramchandani, P., Stein, A., Evans, J. & O'Connor, T. G. Paternal depression in the postnatal period and child development: a prospective population study. *The Lancet* **365**, 2201–2205 (2005). https://doi.org/10.1016/S0140-6736(05)66778-5

34 Böhlmann, P. *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* **19**, e0311989 (2024). https://doi.org/10.1371/journal.pone.0311989

35 Jiyeong Kim, L. M. D. A., 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* **10**, e43066 (2023). https://doi.org/10.2196/43066

36 John A Naslund, K. A. A., 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* **4**, 14 (2017).

37 Abd-Alrazaq, A. A., Rababeh, A., Alajlani, M., Bewick, B. M. & Househ, M. Effectiveness and Safety of Using Chatbots to Improve Mental Health: Systematic Review and Meta-Analysis. *J Med Internet Res* **22**, e16021 (2020). https://doi.org/10.2196/16021

38 Casu, M., Triscari, S., Battiato, S., Guarnera, L. & Caponnetto, P. AI Chatbots for Mental Health: A Scoping Review of Effectiveness, Feasibility, and Applications. *Applied Sciences* **14**, 5889 (2024).

39 Balan, R. & Gumpel, T. P. ChatGPT Clinical Use in Mental Health Care: Scoping Review of Empirical Evidence. *JMIR Ment Health* **12**, e81204 (2025). https://doi.org/10.2196/81204

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

41 Samir Khalil, S., Tawfik, N. & Spruit, M. Exploring the potential of federated learning in mental health research: a systematic literature review. *Applied Intelligence* **54**, 1–18 (2024). https://doi.org/10.1007/s10489-023-05095-1

42 Gehder, S., Brückner, S., Gilbert, S. & Goeldner, M. A scoping review on using real-world data to evaluate the effectiveness of mHealth applications. *NPJ Digit Med* **9** (2026). https://doi.org/10.1038/s41746-026-02562-0

43 Rowan Hussein, A. Z., 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* **9** (2026). https://doi.org/10.1038/s41746-026-02418-7

44 Jee Young Kim, A. H., Suresh Balu & Mark Sendak People process technology and operations framework for establishing AI governance in healthcare organizations. *npj Digital Medicine* **9** (2026). https://doi.org/10.1038/s41746-026-02419-6

45 Anaduaka, U. S., Oladosu, A. O., Katsande, S., Frempong, C. S. & Awuku-Amador, S. 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* **28**, e301445 (2025). https://doi.org/10.1136/bmjment-2024-301445

46 Cunningham, K., Mărginean, V. & Hylock, R. Navigating promise and perils: applying artificial intelligence to the perinatal mental health care cascade. *npj Health Systems* **2**, 26 (2025). https://doi.org/10.1038/s44401-025-00030-7

47 Tameez, H. & Appleton, N. S. AI as Emotional Support in Pregnancy: A Review and Synthesis for Emerging Research Directions. *Social Sciences* **15**, 100 (2026).

48 Yang, H. & Lin, D.-M. Conversational AI for perinatal mental health: promise, limits, and a human-AI stepped-care framework. *Frontiers in Psychiatry* **Volume 17 - 2026** (2026). https://doi.org/10.3389/fpsyt.2026.1847854

49 Islam, S. M. S., Gemmill, A. W., Hirshler, Y., Pascoe, M. & Milgrom, J. Application of Artificial Intelligence in Perinatal Mental Health: A Review. *AI* **7**, 314 (2026).
50 Peters, M. D. J. *et al.* Updated methodological guidance for the conduct of scoping reviews. *JBI Evidence Synthesis* **18**, 2119–2126 (2020). https://doi.org/10.11124/jbies-20-00167
51 Tricco, A. C. *et al.* PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. *Ann Intern Med* **169**, 467–473 (2018). https://doi.org/10.7326/m18-0850
52 Chua, J. Y. X. *et al.* A Mobile App–Based Intervention (Parentbot–a Digital Healthcare Assistant) for Parents: Secondary Analysis of a Randomized Controlled Trial. *J Med Internet Res* **27**, e64882 (2025). https://doi.org/10.2196/64882
53 Chua, J. Y. X. *et al.* The effectiveness of Parentbot - a digital healthcare assistant - on parenting outcomes: A randomized controlled trial. *Int J Nurs Stud* **160**, 104906 (2024). https://doi.org/10.1016/j.ijnurstu.2024.104906
54 Chua, J. Y. X. *et al.* 'Parentbot - A Digital healthcare Assistant (PDA)': A mobile application-based perinatal intervention for parents: Development study. *Patient Educ Couns* **114**, 107805 (2023). https://doi.org/10.1016/j.pec.2023.107805
55 Suharwardy, S. *et al.* Feasibility and impact of a mental health chatbot on postpartum mental health: a randomized controlled trial. *AJOG Glob Rep* **3**, 100165 (2023). https://doi.org/10.1016/j.xagr.2023.100165
56 Darcy, A. *et al.* Anatomy of a Woebot® (WB001): agent guided CBT for women with postpartum depression. *Expert Rev Med Devices* **19**, 287–301 (2022). https://doi.org/10.1080/17434440.2022.2075726
57 Fietta, V. *et al.* World Health Organization Evidence-Based Self-Help Plus Intervention for Stress Management via Chatbot: Protocol for Adaptation to a Tech-Enabled Model. *JMIR Res Protoc* **14**, e69644 (2025). https://doi.org/10.2196/69644
58 Rizzi, S. *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. *JMIR Res Protoc* **14**, e53891 (2025). https://doi.org/10.2196/53891
59 Puspitasari, I. W., Rinawan, F. R., Purnama, W. G., Susiarno, H. & Susanti, A. I. Development of a Chatbot for Pregnant Women on a Posyandu Application in Indonesia: From Qualitative Approach to Decision Tree Method. *Informatics* **9**, 88 (2022).
60 Wong, J. *et al.* A Chatbot to Engage Parents of Preterm and Term Infants on Parental Stress, Parental Sleep, and Infant Feeding: Usability and Feasibility Study. *JMIR Pediatr Parent* **4**, e30169 (2021). https://doi.org/10.2196/30169
61 Mancinelli, E., Magnolini, S., Gabrielli, S. & Salcuni, S. A Chatbot (Juno) Prototype to Deploy a Behavioral Activation Intervention to Pregnant Women: Qualitative Evaluation Using a Multiple Case Study. *JMIR Form Res* **8**, e58653 (2024). https://doi.org/10.2196/58653
62 Rizzi, S. *et al.* A Mindfulness-Based App Intervention for Pregnant Women: Protocol for a Pilot Feasibility Study. *JMIR Res Protoc* **13**, e53890 (2024). https://doi.org/10.2196/53890
63 Teague, S. J., Shatte, A. B. R., Fuller-Tyszkiewicz, M. & Hutchinson, D. M. User-centred design and evaluation of an mHealth app for fathers' perinatal mental health: a feasibility, acceptability, and usability study. *Behav Inf Technol*, 1–15 (2025). https://doi.org/10.1080/0144929x.2025.2502474
64 Teague, S. J., Shatte, A. B. R., Fuller-Tyszkiewicz, M. & Hutchinson, D. M. Mobile app-based intervention for paternal perinatal depression, anxiety, and stress: A randomised controlled trial. *J Affect Disord* **382**, 325–335 (2025). https://doi.org/10.1016/j.jad.2025.04.075
65 McAlister, K., Baez, L., Huberty, J. & Kerppola, M. Chatbot to Support the Mental Health Needs of Pregnant and Postpartum Women (Moment for Parents): Design and Pilot Study. *JMIR Form Res* **9**, e72469 (2025). https://doi.org/10.2196/72469
66 Sadavarte, S. S. & Bodanese, E. in *IEEE Pune Section International Conference (PuneCon)* 1–5 (IEEE, Pune, India, 2019).

67 García-Méndez, S. & de Arriba-Pérez, F. Detecting and Explaining Postpartum Depression in Real-Time with Generative Artificial Intelligence. *Applied Artificial Intelligence* **39**, 2515063 (2025). https://doi.org/10.1080/08839514.2025.2515063
68 Balasooriya, D. *et al.* in *2024 IEEE 9th International Conference for Convergence in Technology (I2CT)* 1–6 (IEEE, Pune, India, 2024).
69 Edirisinghe, D. B. *et al.* in *Intelligent Sustainable Systems Intelligent Sustainable Systems* 95–107 (Springer Nature Singapore, Singapore, 2025).
70 Chung, K., Cho, H. Y. & Park, J. Y. A Chatbot for Perinatal Women's and Partners' Obstetric and Mental Health Care: Development and Usability Evaluation Study. *JMIR Med Inform* **9**, e18607 (2021). https://doi.org/10.2196/18607
71 Montenegro, J. L. Z., da Costa, C. A. & Janssen, L. P. Evaluating the use of chatbot during pregnancy: A usability study. *Healthcare Analytics* **2** (2022). https://doi.org/10.1016/j.health.2022.100072
72 Nguyen, Q. C. *et al.* Rosie, a Health Education Question-and-Answer Chatbot for New Mothers: Randomized Pilot Study. *JMIR Formative Research* **8** (2024). https://doi.org/10.2196/51361
73 Lucia Vaira, M. A. B., Matteo Conte, Francesco Margiotta Casaluci, Antonio Melpignano. in *Proceedings of the 22nd International Database Engineering & Applications Symposium* 273–277 (Association for Computing Machinery, New York, NY, USA, 2018).
74 Khan, Z. *et al.* in *ICT for Intelligent Systems* Vol. 126 (eds Jyoti Choudrie, Eva Tuba, Thinagaran Perumal, & Amit Joshi) 459–477 (Springer Nature Singapore, Singapore, 2025).
75 Viswanathan, S. *et al.* in *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* Article 310 (Association for Computing Machinery, 2025).
76 Kerimoglu Yildiz, G., Turk Delibalta, R. & Coktay, Z. Artificial intelligence-assisted chatbot: impact on breastfeeding outcomes and maternal anxiety. *BMC Pregnancy and Childbirth* **25**, 631 (2025). https://doi.org/10.1186/s12884-025-07753-3
77 Valan, P. & Venugopal, P. Evaluating a retrieval-augmented pregnancy chatbot: a comprehensibility–accuracy-readability study of the DIAN AI assistant. *Frontiers in Artificial Intelligence* **Volume 8 - 2025** (2025). https://doi.org/10.3389/frai.2025.1640994
78 Mayangsari, N. Utilizing Chat GPT as a Digital Health Intervention in Pregnancy Care to Improve Maternal Well-Being. *Social Medicine* **18**, 258–266 (2025). https://doi.org/10.71164/socialmedicine.v18i3.2025.2003
79 Progga, F. T., Khan, A. & Rubya, S. in *Companion Publication of the 2024 Conference on Computer-Supported Cooperative Work and Social Computing* 653–657 (Association for Computing Machinery, San Jose, Costa Rica, 2024).
80 Green, E. P. *et al.* Expanding Access to Perinatal Depression Treatment in Kenya Through Automated Psychological Support: Development and Usability Study. *JMIR Form Res* **4**, e17895 (2020). https://doi.org/10.2196/17895
81 Yang, S., Lee, J., Sezgin, E., Bridge, J. & Lin, S. Clinical Advice by Voice Assistants on Postpartum Depression: Cross-Sectional Investigation Using Apple Siri, Amazon Alexa, Google Assistant, and Microsoft Cortana. *JMIR Mhealth Uhealth* **9**, e24045 (2021). https://doi.org/10.2196/24045
82 Inkster, B., Kadaba, M. & Subramanian, V. 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. *Front Glob Womens Health* **4**, 1084302 (2023). https://doi.org/10.3389/fgwh.2023.1084302
83 Wang, R., Wang, J., Liao, Y. & Wang, J. in *2020 International Conference on Intelligent Computing and Human-Computer Interaction (ICHCI)* 378–383 (Sanya, China, 2020).
84 Kaur, P., Sharma, R. & Singh, M. in *2024 4th International Conference on Technological Advancements in Computational Sciences (ICTACS)* 1498–1503 (IEEE, Tashkent, Uzbekistan, 2024).

85 Bakhtawar Ahtisham, S. T., Zoha Hayat Bhatti, Ayesha Masood, and Suleman Shahid. in *Human-Computer Interaction – INTERACT 2023, LNCS 14143, Springer* Vol. 14143 *Lecture Notes in Computer Science* 319–328 (Abdelnour Nocera, 2023).
86 Sam Jiangyue Yan, G. W., Thanh Vo, Yutong Xiao, Huy Truong in *Proceedings of the 2022 International Conference on Human Machine Interaction* 53–59 (Association for Computing Machinery, New York, NY, USA, 2022).
87 Meskó, B. & Topol, E. J. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *npj Digital Medicine* **6**, 120 (2023). https://doi.org/10.1038/s41746-023-00873-0
88 Button, S., Thornton, A., Lee, S., Shakespeare, J. & Ayers, S. Seeking help for perinatal psychological distress: a meta-synthesis of women's experiences. *British Journal of General Practice* **67**, e692–e699 (2017). https://doi.org/10.3399/bjgp17X692549
89 Freed, D. *et al.* in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* Paper 667 (Association for Computing Machinery, Montreal QC, Canada, 2018).
90 Chen, L., Zaharia, M. A. & Zou, J. Y. How is ChatGPT's behavior changing over time? *ArXiv* **abs/2307.09009** (2023).
91 Fiaz, I., Kanwal, N. & Al-Said Ahmad, A. A Systematic Review of Federated and Cloud Computing Approaches for Predicting Mental Health Risks. *Sensors (Basel)* **26** (2025). https://doi.org/10.3390/s26010229
92 Wei, H.-s., Atwal, P., Humphrey, L. & Yu, Y.-a. Evaluating the safety of AI responses in suicide-related interactions: A clinically informed framework and simulation study. *Psychiatry Research Communications* **6**, 100263 (2026). https://doi.org/10.1016/j.psycom.2026.100263
93 Hassan, B. & Ahmed, A. Fluency without understanding: risks of large language models in mental healthcare. *BJPsych Bulletin*, 1–4 (2026). https://doi.org/10.1192/bjb.2026.10227
94 Huckvale, K., Prieto, J. T., Tilney, M., Benghozi, P. J. & Car, J. Unaddressed privacy risks in accredited health and wellness apps: a cross-sectional systematic assessment. *BMC Med* **13**, 214 (2015). https://doi.org/10.1186/s12916-015-0444-y
95 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).
96 Union, E. P. a. C. o. t. E. 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). Report No. Regulation (EU) 2016/679, 1–88 (European Union, 2016).
97 Ong, J. C. L. *et al.* Innovating global regulatory frameworks for generative AI in medical devices is an urgent priority. *npj Digital Medicine* **9**, 364 (2026). https://doi.org/10.1038/s41746-026-02552-2
98 Organization, W. H. Ethics and governance of artificial intelligence for health: WHO guidance Executive summary. Report No. 9789240037403, 11 (World Health Organization, Geneva, 2021).
99 Collingridge, D. *The Social Control of Technology*. (St. Martin's Press, 1980).
100 Peters, M. D. J. *et al.* Scoping reviews: reinforcing and advancing the methodology and application. *Systematic Reviews* **10**, 263 (2021). https://doi.org/10.1186/s13643-021-01821-3
101 Obe, D., Shatte, A., Teague, S. & Lee, I. 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).
102 Tabassi, E. Artificial Intelligence Risk Management Framework (AI RMF 1.0). Report No. NIST AI 100-1, (National Institute of Standards and Technology, Gaithersburg, MD, 2023).

103 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, 2022).
104 Union, E. P. a. C. o. t. E. 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). Report No. Regulation (EU) 2024/1689, (European Union, 2024).
105 Government, U. Regulation (EU) 2016/679 of the European Parliament and of the Council. (2016).
106 Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods* **17**, 261–272 (2020). https://doi.org/10.1038/s41592-019-0686-2

Supplementary Files

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

- SupplementaryInformationScopingReview.pdf
- SupplementaryData1.xlsx