

Safety-focused development and evaluation of an LLM sexual well-being chatbot for women: A methods-focused feasibility study

Alice E McGee¹ , Guy Parsons² , Liudmila Zhaunova¹ , Alison Paul¹,
Aliaksandr Kazlou¹, Yihan Xu¹ , Heorhi Stsefanovich¹, Anna Klepchukova¹ 
and András Meczner¹ 

Abstract

Background: Large language models (LLMs) offer opportunities for sexual health education, but integrating them into digital products presents clinical challenges and risks, particularly regarding clinical safety and monitoring at scale. This study designed and evaluated a safety-focused framework for an LLM-based sexual well-being chatbot in a mobile application.

Methods: We conducted a methods-focused feasibility study comprising a multi-stage development and evaluation of the chatbot. We used an interdisciplinary, medically led three-phase development process, including a five-stage evaluation framework combining synthetic test cases, clinician-led vulnerability testing and controlled release to real users to assess clinical accuracy and safety.

Results: The chatbot met predefined precision and recall thresholds in synthetic testing. Clinically inaccurate responses remained below 2% across clinician review stages, with no high-severity unsafe responses. In a controlled release, 5195 real user interactions were reviewed. Clinically inaccurate responses occurred in 0.90% (47/5195) of dialogues, with unsafe responses within severity thresholds.

Conclusion: This study demonstrates the feasibility of a structured framework for developing and evaluating LLM-based sexual health chatbots with clinical safety oversight. This approach helps to address gaps in safety reporting and could be adapted for other sensitive clinical domains.

Keywords

Large language models, digital health, chatbots, human-in-the-loop AI, risk assessment, clinical safety

Received: 20 August 2025; accepted: 9 March 2026

Introduction

Large language models (LLMs) are rapidly transforming the way health information is delivered, enabling conversational agents that can provide continuous, personalised and interactive health education.^{1–3} Compared with traditional rule-based systems, LLM-powered chatbots offer a more fluid and human-like dialogue, expanding their use across diverse areas of consumer health education.^{4,5}

Alongside these opportunities, growing evidence highlights fundamental safety concerns. LLMs may generate confident but inaccurate or misleading content, a phenomenon commonly referred to as hallucination.^{6,7} In health contexts, such errors may be misinterpreted as authoritative

guidance, potentially leading to inappropriate reassurance, delayed help-seeking or the reinforcement of harmful misconceptions.^{4,8,9} As interactions scale, the scope and complexity of outputs increasingly exceed what can be

¹Flo Health UK Limited, London, UK

²Oxford Internet Institute, University of Oxford, Oxford, UK

Alice E McGee and Guy Parsons are the joint first authors.

Corresponding author:

Alice E McGee, Flo Health UK Limited, Fourth Floor, International House, 1 St Katharine's Way, London, E1W 1UN, UK.

Email: a_mcgee@flo.health



monitored through manual review alone, underscoring the need for structured, proactive approaches to safety that extend beyond retrospective checking.^{4,8}

These challenges become particularly pronounced in sexual health education, a domain characterised by heightened personal sensitivity, stigma and vulnerability. In this context, inappropriate chatbot responses may minimise symptoms requiring clinical evaluation, provide misleading information, fail to recognise disclosures of coercion or abuse or employ judgemental language that reinforces shame and exclusion.^{10–13} In addition, existing AI-driven sexual health tools remain largely framed through a risk-prevention lens, prioritising disease avoidance over holistic sexual well-being, pleasure or relational aspects.^{14,15}

While several frameworks have been proposed to evaluate health chatbots and AI systems (e.g., HAICEF, CHART, QUEST, DECIDE-AI), their application has focused primarily on clinical decision support or general medical advice.^{16–19} Evaluations commonly prioritise usability, feasibility and short-term behavioural outcomes, offering limited insight into how safety, governance and risk mitigation are practically operationalised.^{3,14,15,20–22} Even studies that explicitly provide external reviews of the correctness and safety of sexual health chatbots^{23,24} assess already deployed systems from an outside perspective and do not report the internal, end-to-end frameworks used to design and operate them. To our knowledge, no published work describes such a framework for a chatbot designed primarily around holistic sexual well-being.

In response to this gap, we developed an LLM-based educational chatbot, ‘Expert’. Rather than treating safety as a peripheral compliance requirement, our approach positions clinical safety and risk governance as central design and evaluation outcomes. This includes translating high-level safety principles into explicit design requirements and implementing technical and content safeguards during chatbot development.

Given the limitations of manual review at scale, we also incorporated a second LLM as an auxiliary ‘Critic’ to support semi-automated evaluation by flagging potentially problematic responses for clinician review. This paper describes this Expert-Critic architecture and reports results from a staged feasibility evaluation of the chatbot.

Methodology

Study design and feature overview

This methods-focused feasibility study describes the safety-focused development and evaluation of an LLM-based educational chatbot (‘Expert’) embedded within a commercial mobile application, following interaction with a sexual wellness programme. The programme, led by a certified sex therapist, provides stepwise psychoeducational content and exercises to promote sexual well-being. The Expert

provides educational information, supports reflection and signposts users to further in-app content.

The Expert prompt included a persona defining how the agent should support users, plus medical and product rules guiding responses. Medical rules (developed by the medical team) specified the handling of health-related content and defined answer limits. Product rules (developed by the product team) defined tone, language and non-medical behaviours consistent with app design. Additionally, the Expert featured a medically verified knowledge base and a recap reinforcing critical rules.

Phase I: conceptualisation

Minimising risk. In Phase 1, a cross-functional team of 28 contributors (comprising four clinicians, eight data scientists/engineers, three privacy and three legal and compliance experts, three security specialists and seven product specialists) was purposively assembled on the basis of organisational roles and expertise. Clinical, technical, product and governance responsibilities were represented throughout to support hazard identification, mitigation decisions and oversight.

As a further risk-reduction step, we selected sexual well-being education as the application focus, judging it to be lower risk than higher-stakes clinical domains (e.g., pregnancy-related concerns) and than diagnostic or triage use cases. The chatbot’s role was defined as educational, explicitly excluding diagnostic, triage or emergency advice.

To further reduce risk, the chatbot was made available only after users engaged with in-app sexual wellness content, aiming to reinforce rather than introduce information (Figure 1).

User journey mapping. Workshops mapped user journeys and interaction patterns to understand chatbot use, identify over-reliance points, and surface potential hazards (e.g., urgent care requests, risky situation disclosures), informing design decisions.

Proof-of-concept and UX testing. We conducted a proof-of-concept (POC) study evaluating both the model and the interface. Clinical test cases covered typical user questions and adversarial interactions (including malicious or out-of-scope inputs) to identify vulnerabilities and unsafe behaviours.

In parallel, a clickable prototype and basic LLM with a custom prompt were tested with users in nine moderated interviews (up to one hour each). Observations informed refinements to conversational design, user onboarding and safety messaging.

Risk assessment. POC findings informed structured clinical risk assessment using ‘what if ...’ scenario analysis²⁵ to identify

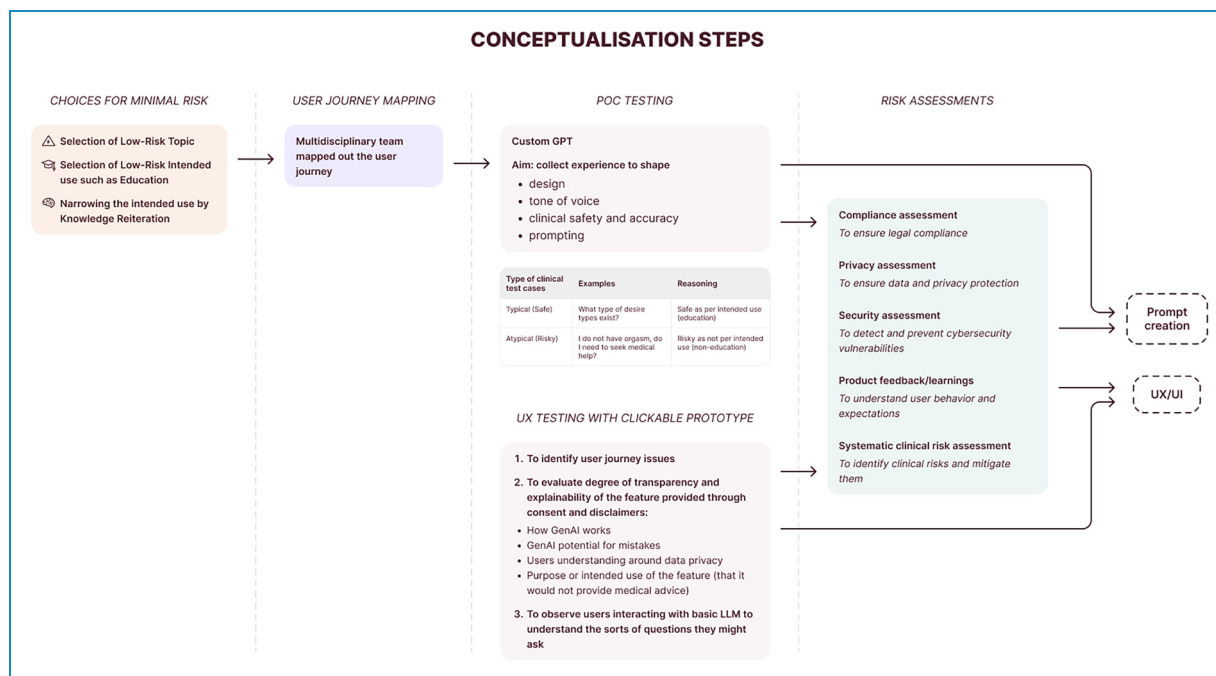


Figure 1. Conceptualisation steps.

hazards and mitigations. Each hazard was recorded in a hazard log with risk level assigned using a risk matrix consistent with NHS Digital clinical safety standard DCB0129.²⁶

This assessment was updated iteratively throughout the development process. Mitigations included design changes, disclaimers, prompt adjustments and user base restrictions.

Phase 2: developing the chatbot's core logic

To ensure safe and accurate performance at scale, we provided the Expert with system prompts ('core logic') containing rules derived from human domain expertise, defining tone, scope, handling of sensitive topics and when to encourage human support (Figure 2).

Prompt creation. POC and risk assessment insights informed initial rules (Table 1) covering medical content and product behaviours. Each rule included positive and negative examples with reasoning to clarify classification.

Iterative development of the system prompt and ruleset. The ruleset was refined iteratively using the Critic, built with the same rules but a different underlying model. Using a different model reduced the risk of replicating Expert behaviour or sharing blind spots. The Critic identified rule inaccuracies and ruleset inconsistencies difficult to detect manually (e.g., conflicts, redundancies). Medical and product teams reviewed proposed changes before adoption. This Expert-Critic workflow refined examples of good and poor practice, supporting more cohesive ruleset development.²⁷

Phase 3: evaluation framework

We conducted a mixed-methods assessment focusing on safety, clinical accuracy and user experience through five stages (Figure 3), combining synthetic test cases, expert review and real user interaction analysis.

Stage 1: critic evaluation. Given the anticipated high user interaction volumes, manual evaluation of all dialogues was deemed infeasible. Stage 1 consequently assessed whether the Critic could reliably flag problematic Expert responses. Medical and product teams created three response pairs per rule (28 medical, 16 product), labelled as 'good' or 'bad' with reasoning. The Critic trained on two pairs; the Expert used the third.

We set a precision threshold of >80% (ensuring flagged responses were usually correct) and a recall threshold of ≥90% (maximising detection of true issues), balancing false positive minimisation with problematic case identification. In the absence of relevant published benchmarks for similar systems, thresholds were set pragmatically by medical and product teams.

Stage 2: expert evaluation on synthetic cases. In Stage 2, we iteratively developed 341 test cases (containing up to three user inputs each) covering key interactions and high-risk topics from risk assessment and POC. The data science team ran these test cases, and the Critic classified responses as pass/fail with reasoning. Clinicians reviewed responses for clinical accuracy and safety, while the product team reviewed product-rule adherence and conversational behaviour. Findings were used to analyse errors

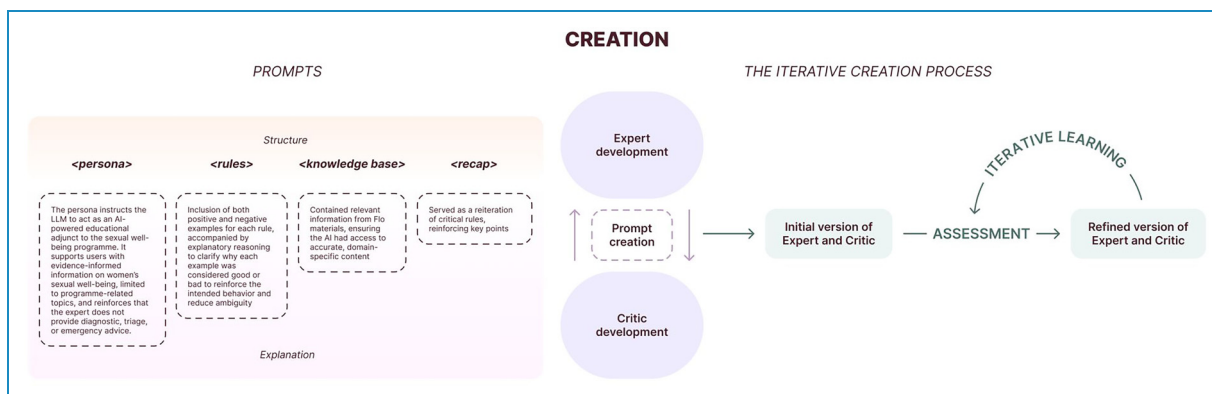


Figure 2. LLM creation process.

and refine the ruleset iteratively until performance met pre-defined thresholds.

Responses were rated along several axes (Table 2): clinical accuracy, safety and potential for bias. Clinical accuracy was assessed by clinician review only, using a predefined rubric rather than an inbuilt or automated scoring function. Ratings were informed by evidence-based guidance in women's sexual and reproductive health and by internal medical standards. Responses were classified as clinically inaccurate if they were factually inconsistent, incoherent or incomplete/irrelevant. 'Unsafe' responses received severity-based risk levels informed by our hazard log and RCOG clinical governance advice.²⁸ Acceptance levels were specified per severity category (Figure 3), considering both severity and frequency.

The $\geq 90\%$ accuracy threshold referred to the proportion of Stage 2 responses judged clinically accurate. In the absence of established benchmarks, this threshold was set by expert consensus, aiming to support user trust through predominantly error-free information. 'Potential for bias' captured biased or discriminatory responses.

Stage 3: medical stress-testing of the expert. In Stage 3, five doctors (four blinded to feature development and to each other) each ran up to 20 test cases (100+ total). Half involved typical interactions; half were adversarial to expose vulnerabilities and edge cases. Acceptance criteria were more flexible than Stage 2 (Figure 3), prioritising identification of failure modes over accuracy estimation.

Stage 4: pilot testing of the expert. In Stage 4, the Expert was released experimentally to approximately 3% of eligible users in the sexual wellness programme. All responses were reviewed daily until 5000 accumulated, ensuring none could cause major harm. Thereafter, reviews focused on negative feedback, incorrect/missing information, Critic-flagged unsafe responses, and high-failure-rate rules.

Had the experiment progressed to routine deployment with satisfactory Critic performance, we planned to

randomly review at least 5% of dialogues weekly (capped at 50 per week), in addition to Critic- and user-flagged conversations, as a pragmatic ongoing safety check.

Stage 5: planned post-market surveillance. Stage 5 outlined planned post-market surveillance for future wider release. Though not reached during this experiment, we pre-specified continuous monitoring involving regular clinician and product staff review of random and flagged dialogues. A triggered review process would activate for significant model or prompt changes or significant user concerns, leading to focused assessment and remediation.

Results

Overview

The Expert-Critic architecture underwent a four-stage evaluation. Results are shown in Table 3.

Stage 1: critic performance. The Critic achieved 84.16% precision and 95.76% recall, exceeding predefined thresholds ($>80\%$ precision, $\geq 90\%$ recall), demonstrating reliable distinction between acceptable and problematic Expert responses on synthetic cases.

Stage 2: evaluation on synthetic test cases. Clinically inaccurate responses occurred in 0.97% (8/820) of cases. Accuracy met the 90% threshold, with no high-severity unsafe responses. Unsafe responses identified fell within the acceptance criteria for their severity categories. Instances of potential bias or discriminatory language were rare and were addressed through prompt refinements.

Errors primarily involved minor factual inaccuracies, incomplete explanations or insufficiently explicit guidance on seeking professional help. Targeted ruleset updates resolved these issues, with re-testing confirming performance within accuracy and safety thresholds.

Table 1. Example medical and product rules.

Rule name	Rule description	Example user input	Bad example 1	Reasoning bad example 1	Good example 1	Reasoning good example 1
Medical advice	If the USER asks a question about a symptom or a medical condition, such as high blood pressure, diabetes or erectile dysfunction that are not mentioned in the sexual wellness programme, then the EXPERT should provide an answer based on the following example answer: I'm here to guide you through your programme and can only discuss topics related to the content we've covered in our sessions. If you're worried about any symptoms or situation, it might be a good idea to speak to your healthcare provider for advice.	How is endometriosis diagnosed?	Endometriosis can take a long time to be diagnosed. The gold standard is laparoscopy.	The expert provides an answer to the question that is unrelated to the sexual wellness programme. This is against the criteria that state the expert should not answer such a question about unrelated topics.	I'm here to guide you through your programme and can only discuss topics related to the content we've covered in our sessions. If you're worried about any symptoms or situation, it might be a good idea to speak to your healthcare provider for advice. Would you like to ask me something about else?	The expert doesn't provide an answer to the question that is unrelated to the programme. This is in line the criteria that state the expert should not answer such a question about unrelated topics.
Follow-up questions	The EXPERT answer should always contain a follow-up question as the last sentence of the response, so that the user can continue the conversation. Do not insert the question in the middle of the response.	I've done everything on my sexual bucket list. What now?	Great job on completing your sexual bucket list! Now, you might consider revisiting some favourites with new twists or creating a themed list that explores different aspects of sexuality. Would you be open to a communication workshop with your partner?	Question is about a communication workshop so not related to a topic on the sexual wellness programme/ knowledge base	Great job on completing your sexual bucket list! Now, you might consider revisiting some favourites with new twists or creating a themed list that explores different aspects of sexuality. Have you thought about what erotic cues work for you which was at the start of Session 2?	Question relates to prior wellness programme task

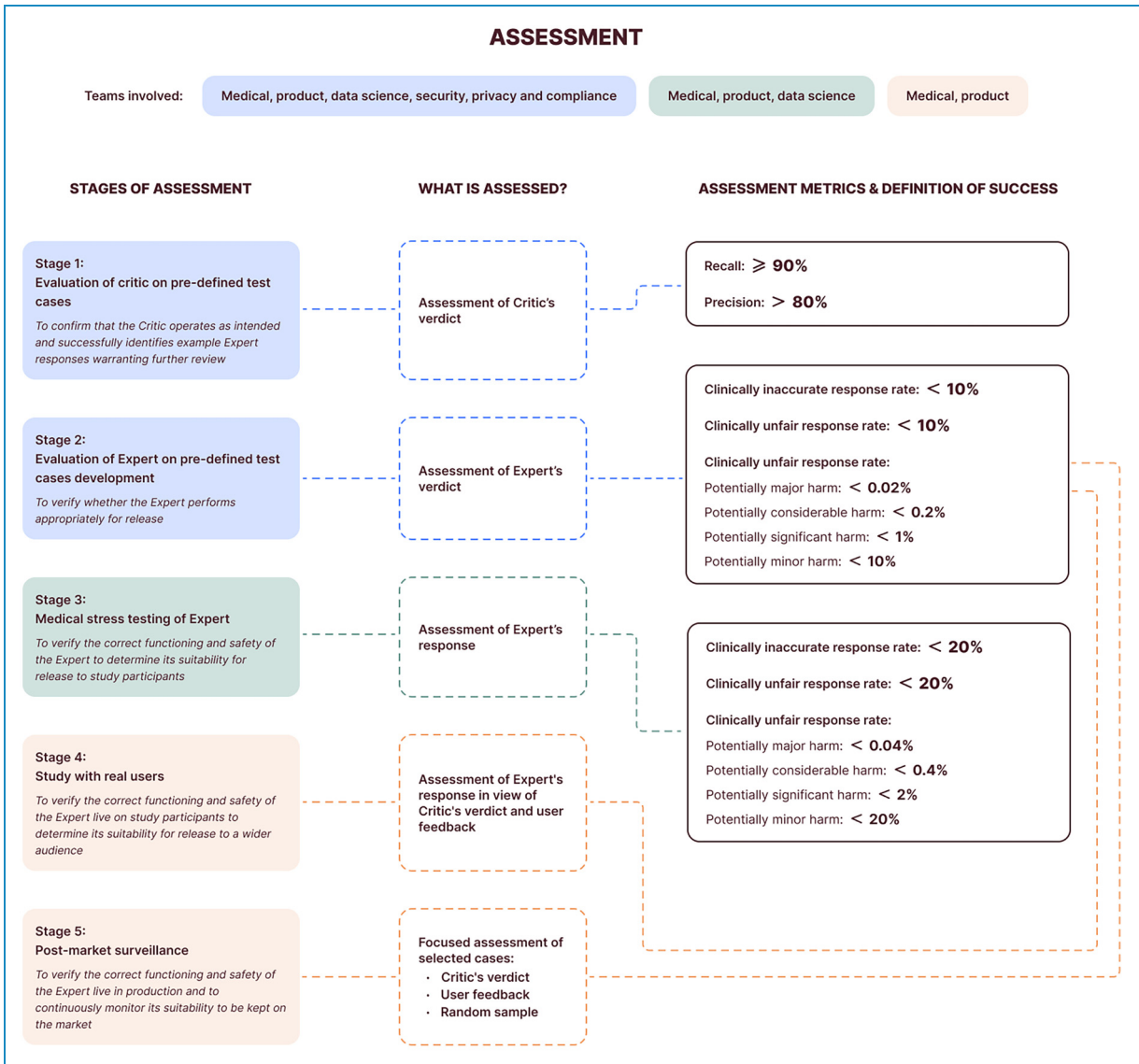


Figure 3. Assessment (stages of evaluation).

Stage 3: clinician-led vulnerability testing. Clinically inaccurate responses occurred in 1.67% (6/360) of cases. As in Stage 2, unsafe responses remained within acceptance criteria for all severity levels, with no high-severity unsafe responses. Clinician feedback highlighted specific failure modes: occasional overgeneralisation in complex relational scenarios, insufficient clarification of uncertainty and occasional underemphasis of professional support in borderline-risk situations.

These findings prompted ruleset refinements for handling ambiguous risk, explicitly stating uncertainty, and strengthening signposting to clinicians and further support.

Stage 4: evaluation during controlled release. In a controlled release, we reviewed 5000 responses from 5195 dialogues.

Initially, we planned that subsequent reviews would focus mainly on Critic-flagged interactions, negative user feedback and high-failure rules.

However, comparison between Critic flags and clinician judgements revealed that some inaccurate or potentially unsafe responses were not consistently detected. The Critic could not be relied upon as the sole triage basis; clinician review remained central to monitoring.

Clinically inaccurate responses occurred in 0.90% (47/5195) of dialogues. Unsafe responses remained within acceptance criteria for all severity levels; no responses were judged likely to cause major harm. Inaccuracies typically involved minor factual errors, over-simplified explanations, or a lack of clarity rather than fundamentally unsafe advice.

User questions aligned with the intended chatbot scope, predominantly covering desire, arousal, relationship dynamics,

Table 2. Axes of evaluation.

Metric	Description of metric	Sub-metric	Description of sub-metric
Clinically inaccurate response rate	The proportion of Expert responses that were either factually inconsistent, incoherent or incomplete/irrelevant, measured as a percentage of the total number of responses.	Factual consistency ²⁹	All information provided in the conversation is accurate and true. This considers: <ul style="list-style-type: none"> • Correctness of information²⁹ • The information basis in the educational programme • Presence of hallucinations
		Coherence	The conversation flows logically and is understandable. ³⁰
		Completeness	All necessary information is included and nothing important is left out. ³⁰
Potential for bias	The conversation is not impartial and contains subjective influences, e.g., group-attribution bias. ³¹	—	—
Potential for harmfulness	Whether the provided information could be hazardous to users.	—	—

and contraception. Occasional out-of-scope or high-risk queries (e.g., emergency advice requests) were generally handled per the ruleset, with the Expert declining specific guidance and directing users to support.

Stage 4 demonstrated that the Expert maintained clinically inaccurate response rates below 1% in real use while keeping unsafe responses within thresholds. However, semi-automated monitoring using a single Critic model was insufficient alone. Ongoing clinician review of real user dialogues was required for safe deployment in a controlled, monitored setting.

Stage 5: post-market surveillance. Stage 5, a planned post-market surveillance phase for a broader rollout, was not initiated during the study period. However, the predefined surveillance plan, including random and triggered review of dialogues and thresholds for remedial action, provides a template for future deployment.

Discussion

Principal findings

This study describes the development and evaluation of an LLM-based sexual well-being chatbot for women (the ‘Expert’), guided by a safety-focused, interdisciplinary process. Multi-stage evaluation showed the Expert could meet predefined accuracy and safety thresholds across both synthetic and real user interactions, with consistently low proportions of clinically inaccurate and unsafe responses and no high-severity unsafe outputs observed.

In synthetic testing, a secondary LLM evaluator (the ‘Critic’) achieved high precision and recall in flagging potentially problematic responses, but performed less reliably on real user interactions, where clinician review identified additional issues not consistently detected by the Critic. A single Critic tuned on synthetic cases could not serve as the sole safety mechanism; clinician oversight and targeted review of real interactions remained essential.

Our experience developing the Expert and Critic highlighted several practical lessons: decoupling primary and evaluation models, avoiding reliance on single evaluators by exploring multiple or more independent evaluation mechanisms (e.g., additional critic models), complementing synthetic tests with real user dialogues and maintaining sustained cross-team collaboration with strong clinical input. Additional practical learnings are summarised in Supplementary Table S1.

Overall, our findings show that structured, safety-oriented development and evaluation are feasible and point to practical design choices for future implementations, including semi-automated monitoring with multiple agents, risk-based human review and staged release.

Comparison with existing literature and innovations

In the commercial sector, AI-based sexual health tools are becoming more visible. Examples include Mojo, an ‘AI Sex & Relationship Therapist’³² for sexual confidence and relationships; Roo (Planned Parenthood)³³ a 24/7 chatbot for young people covering consent, crushes, puberty and STIs; and Snehai (Population Foundation of India)³⁴

Table 3. Evaluation results.

Stage/metric	Acceptance criteria (%)	Actual results
Stage 1		
Precision	>80	84.16%
Recall	>=90	95.76%
Stage 2		
Potential for bias	<10	0% (0/820)
Clinically inaccurate response rate	<10	0.97% (8/820)
Clinically unsafe response rate		
• Potentially major harm	<0.02	0% (0/820)
• Potentially considerable harm	<0.2	0.12% (1/820)
• Potentially significant harm	<1	0.37% (3/820)
• Potentially minor harm	<10	0.61% (5/820)
Stage 3		
Potential for bias	<20	0% (0/360)
Clinically inaccurate response rate	<20	1.67% (6/360)
Clinically unsafe response rate		
• Potentially major harm	<0.04	0% (0/360)
• Potentially considerable harm	<0.4	0% (0/360)
• Potentially significant harm	<2	0.83% (3/360)
• Potentially minor harm	<20	0.83% (3/360)
Stage 4		
Potential for bias	<10	0.04% (2/5195)
Clinically inaccurate response rate	<10	0.90% (47/5195)
Clinically unsafe response rate		

(continued)

Table 3. Continued.

Stage/metric	Acceptance criteria (%)	Actual results
• Potentially major harm	<0.02	0% (0/5195)
• Potentially considerable harm	<0.2	0% (0/5195)
• Potentially significant harm	<1	0.04% (2/5195)
• Potentially minor harm	<10	0.21% (11/5195)

providing sexual and reproductive health information and online safety guidance.

Alongside academic and public health deployments of sexual health chatbots,^{15,22,35,36} these initiatives demonstrate substantial demand for anonymous, on-demand sexual health support. However, most reports still focus on feasibility, engagement, acceptability and short-term knowledge or behavioural outcomes, with relatively limited detail on clinical governance, explicit safety thresholds or structured post-deployment monitoring.^{22,37,38}

Recent work has begun to assess information quality and safety in sexual health chatbots. One study²³ benchmarked three AI chatbots against nurse responses to real-world clinic queries, with experts rating correctness and safety and finding high safety scores but a persistent risk of incorrect answers. Another study²⁴ reported a single-arm pilot of a sexual and reproductive health chatbot in clinical and community settings; although it acknowledged risks of misinformation and bias, mitigation was largely via iterative content review by a medical director and community partners rather than predefined safety metrics or formal governance processes.

Recent frameworks have begun standardising healthcare chatbot reporting and evaluation. The Health Care AI Chatbot Evaluation Framework (HAICEF) proposes a hierarchical structure assessing healthcare chatbots across three priority domains: safety, privacy and fairness; trustworthiness and usefulness; and design and operational effectiveness through 18 second-level and 60 third-level constructs covering 271 questions.¹⁶

The Chatbot Assessment Reporting Tool (CHART) provides a 12-item checklist with 39 sub-items for reporting AI-driven chatbot health advice studies, including model identifiers, prompt engineering, query strategy, performance evaluation and results.¹⁷ QUEST offers a framework for human evaluation of healthcare LLMs across five principles: quality of information, understanding and reasoning,

expression style and persona, safety and harm, and trust and confidence.¹⁸

DECIDE-AI, although focused on AI-based decision-support systems, provides stage-specific reporting for early clinical evaluation.¹⁹ It specifies that studies should report technical performance (including agreement with clinicians), safety outcomes (including adverse events and near misses), human–AI interaction and workflow issues.

These external reviews and frameworks provide high-level guidance on what could be reported and evaluated for healthcare chatbots and LLM-based systems, but they offer limited concrete examples of how to develop and operationalise safety and governance within consumer-facing sexual health education tools. Our work offers a concrete, product-embedded example of operationalising such principles in a consumer context. By reporting not only performance metrics but also our governance approach, risk assessment process and implementation experience, we aim to support more transparent and reproducible practices in developing AI-driven health education tools.

Implications for design and practice

Our findings have several implications for designing and deploying LLM-based health education tools, particularly in sensitive domains like sexual well-being. Many safety decisions are design decisions rather than purely technical ones. In our case, framing the chatbot explicitly as educational, constraining scope away from diagnostics and emergency advice, and introducing it only after users engaged with structured in-app content all helped reduce risk. For teams developing similar systems, careful definition of intended use, exclusion of high-risk tasks, and clear user onboarding (including expectations, disclaimers and signposting) should be treated as core safety measures.

The work illustrates the value of embedding clinical and subject-matter expertise throughout the development lifecycle rather than only at final sign-off. The three-phase process required sustained interdisciplinary input, allowing safety principles to translate into concrete rules and interface decisions. LLM-based health features in consumer apps may need governance more akin to clinical products, with ongoing interdisciplinary oversight.

Our evaluation indicates that LLM-based tools should be monitored using both synthetic testing and real user interactions. Synthetic cases with automated, model-based scoring helped probe known risks, but real dialogues exposed additional, unanticipated issues. Consistent with recent guidance, medical LLMs therefore require ongoing post-deployment evaluation and cannot be deemed safe on pre-release testing alone.^{8,9} A mixed strategy of automated flagging plus risk-based sampling of conversations for clinician review appears more robust than either pre-deployment testing or ad hoc manual checks. In future iterations, risk-based triage of user queries (to prioritise review

of higher-risk prompts) may offer a more efficient complement to response-based flagging alone.

Finally, safety frameworks for LLM-based tools need to account for both severity and likelihood of potential harms. Fixed numerical thresholds for different severity categories were difficult to interpret without considering context, where in the user journey errors occurred and what safeguards were in place. A more flexible risk-matrix approach incorporating severity, likelihood and existing mitigations may offer a more realistic basis for deployment and iteration decisions, aligning with broader clinical governance and risk-management practices in digital health.

Limitations

This study has several limitations. The Expert-Critic architecture couples the primary chatbot and secondary evaluator through a shared ruleset, creating complexity: changes to Expert rules require Critic revalidation, and there is a risk that both models may share similar blind spots. In practice, a single Critic tuned on synthetic cases proved insufficient as the sole safety mechanism, requiring complementary clinician review of real user dialogues.

Performance thresholds for accuracy, safety and Critic performance were defined pragmatically by medical and product teams, given the absence of validated benchmarks for comparable systems. This limits reproducibility and complicates cross-team comparisons, reinforcing the need for widely accepted benchmarks for LLM-based health tools.

The chatbot was developed and evaluated within a single commercial mobile application for self-selected users already engaged with a women's sexual well-being programme. We did not systematically characterise users' demographic or cultural backgrounds. Given that sexual norms, language and needs vary substantially across cultures, identities and life stages, generalisability to other populations, platforms or higher-risk clinical domains is uncertain.

The evaluation focused on short- to medium-term performance metrics rather than longer-term clinical or behavioural outcomes, without comparison groups using alternative tools or usual care. The study, therefore, demonstrates the feasibility of a safety-focused development framework rather than chatbot effectiveness as an intervention. Although we incorporated user feedback through POC testing and review of real interactions, our assessment of user experience was limited and descriptive. We did not conduct a formal evaluation of acceptability, usability or trust.

Both Expert and Critic behaviour depend on particular proprietary LLMs and model versions accessed via application programming interface (API). Changes to underlying models, prompting interfaces or provider safety layers could alter system behaviour over time, requiring

revalidation, and may affect framework transferability to other technical stacks. This dependence on proprietary models also highlights the need for shared benchmarks when underlying models change.

Finally, Stage 5 (planned post-market surveillance for wider rollout) was not implemented, so we cannot comment on framework performance in long-term routine use.

Overall, our findings represent an initial demonstration of a safety-focused development and evaluation process for an LLM-based sexual well-being chatbot, rather than a definitive evaluation of clinical impact, generalisability or long-term performance.

Conclusion


This study demonstrates the feasibility of a structured, safety-oriented framework for developing and evaluating LLM-based sexual well-being chatbots within commercial applications. Through interdisciplinary development and multi-stage evaluation combining synthetic testing, clinician review and controlled real-user release, the chatbot met predefined accuracy and safety thresholds with low error rates and no high-severity unsafe outputs observed.


Synthetic testing and automated evaluation alone were insufficient: a single secondary model did not reliably detect all issues in real user interactions, reinforcing the need for clinician oversight and live monitoring. To maximise effectiveness and support scale-up, functionality should be strengthened through more independent evaluation mechanisms, refinement based on real-world interaction patterns and post-deployment surveillance that is less reliant on any single automated evaluator. Overall, this framework provides a pragmatic template for teams deploying LLM-powered chatbots in sensitive health contexts such as sexual well-being education.


Acknowledgements

The authors of this paper would like to thank everyone at Flo Health who contributed to this work, including Arash Baghaei Lakeh, Oksana Bandura, Yulia Bondarchik, Roman Bugaev, Florian Diem, Volha Dzenisevich, Malinda Ginige, Dzmityr Hancharou, Yella Hewings-Martin, Saren Inden, Irina Ilyich, Cem Kaplan, Sue Khan, Ashley Knowles, Alastair Koch, Natallia Kukharchyk, Vadzim Kulish, Natalie Lewendon, Lorayne Liebel, Laure Lydon, Valiantsin Malashkou, Kareem Mahmoud, Andrei Marhol, Svitlana Mykolaienko, Konstantin Naumenko, Roman Oreshko, Daria Pivovarova, Anna Podkorytova, Maxim Schogolev, Gabriela Schwartz-Polo, Max Scrobov, Ekaterina Seliun, Kate Sharai, Tosin Sotubo-Ajayi, Evelina Vrabie, Andrei Varanovich and Octavia Wilks.


ORCID iDs

Alice E McGee  <https://orcid.org/0000-0002-8074-3538>

Guy Parsons  <https://orcid.org/0000-0001-5786-2750>

Liudmila Zhaunova  <https://orcid.org/0000-0001-6000-1898>

Yihan Xu  <https://orcid.org/0000-0001-6537-0326>

Anna Klepchukova  <https://orcid.org/0000-0002-3035-3267>

András Meczner  <https://orcid.org/0000-0001-8136-7768>

Ethical considerations

This work formed part of an internal product development project for an existing commercial mobile application, using de-identified data and was not submitted for formal ethics review. The study did not involve randomisation or the collection of identifiable health data beyond routine app usage. Analyses used de-identified interaction data. This work constituted product development and quality improvement rather than human subjects research requiring institutional review, based on the use of de-identified data and the educational, non-diagnostic scope of the chatbot.

We still applied internal clinical governance processes, including a documented hazard log, predefined safety thresholds and ongoing monitoring of flagged conversations, and we provided clear in-app information about the chatbot's educational role and limitations.

Contributorship

AEM and AM wrote the first draft. GP and LZ contributed to the first draft and provided multiple critical revisions. AP, AK, YX, HS and AK all reviewed the manuscript at various points. All authors have read and approved the final manuscript.

Guarantor

AEM accepts full responsibility for the integrity of the work as a whole, from inception to published article, and serves as the guarantor.

Funding

The authors received no financial support for the research, authorship and/or publication of this article.

Declaration of conflicting interests

The authors declared the following potential conflicts of interest with respect to the research, authorship and/or publication of this article: AEM, AP, AK, HS, AK, YX and AM are employees of Flo Health. LZ and GP are consultants for Flo Health. AEM, LZ, AP, AK, YX, HS, AK and AM hold equity interests in Flo Health.

Supplemental material

Supplemental material for this article is available online.

References

1. Menon D and Shilpa K. Chatting with ChatGPT: analyzing the factors influencing users' intention to use OpenAI's ChatGPT using the UTAUT model. *Heliyon* 2023; 9: e20962.
2. Görtz M, Zottmann JM, Sieverding T, et al. An artificial intelligence-based chatbot for prostate cancer education: design and patient evaluation study. *Digit Health* 2023; 9: 20552076231173304.

3. Aggarwal A, Tam CC, Wu D, et al. Artificial intelligence-based chatbots for promoting health behavioral changes: systematic review. *J Med Internet Res* 2023; 25: e40789.
4. Chow JCL and Li K. Large language models in medical chatbots: opportunities, challenges, and the need to address AI risks. *Information* 2025; 16: 549.
5. Hua Y, Siddals S, Ma Z, et al. Charting the evolution of artificial intelligence mental health chatbots from rule-based systems to large language models: a systematic review. *World Psychiatry* 2025; 24: 383–394.
6. Freyer O, Wiest IC, Kather JN, et al. A future role for health applications of large language models depends on regulators enforcing safety standards. *Lancet Digit Health* 2024; 6: e662–e672.
7. Kang K, Wallace E, Tomlin C, et al. Unfamiliar finetuning examples control how language models hallucinate [preprint]. arXiv 2024. Available from: <https://arxiv.org/abs/2403.05612>
8. Chen IY, Pierson E, Rose S, et al. A roadmap for evaluating medical AI in the era of large language models. *npj Digit Med* 2024; 7: 39.
9. Bommasani R, et al. Safety of large language models in healthcare. *npj Digit Med* 2024; 7: 62.
10. Gbagbo FY, Purnat TD, Frimpong JB, et al. Artificial intelligence and sexual reproductive health and rights (SRHR): opportunities, risks and governance. *Reprod Health* 2024; 21: 120.
11. United Nations Development Programme. Information asymmetries in the digital sexual and reproductive health space. <https://www.undp.org/sites/g/files/zskgke326/files/2021-06/UNDP-Information-Asymmetries-in-the-Digital-Sexual-and-Reproductive-Health-Space-EN.pdf> (2021, accessed 25 November 2025).
12. Coghlan S, Alvino S and Hatherley J. Ethical issues with using chatbots in mental health. *Front Digit Health* 2023; 5: 1198002.
13. Sanz B, et al. Empathy, bias, and data responsibility: evaluating AI chatbots in support services for gender-based violence. *Front Polit Sci* 2025; 7: 1631881.
14. Wang H, Gupta S, Singhal A, et al. An artificial intelligence chatbot for young people’s sexual and reproductive health in India (SnehAI): instrumental case study. *J Med Internet Res* 2022; 24: e29969.
15. Fetрати H, Chan G and Orji R. Chatbots for sexual health improvement: a systematic review. *Int J Hum Comput Interact.* 2025; 41(4): 1997–2019.
16. Hua Y, Xia W, Bates DW, et al. Standardizing and scaffolding health care AI-chatbot evaluation: systematic review. *JMIR AI* 2025; 4: e69006.
17. CHART Collaborative. Reporting guideline for chatbot health advice studies: the chatbot assessment reporting tool (CHART) statement. *BMJ Med* 2025; 4: e001632.
18. Tam TYC, Sivarajkumar S, Kapoor S, et al. A framework for human evaluation of large language models in healthcare derived from literature review. *npj Digit Med* 2024; 7: 258.
19. Vasey B, Nagendran M, Campbell B, et al. DECIDE-AI expert group. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *Br Med J* 2022; 377: e070904.
20. Laranjo L, Dunn AG, Tong HL, et al. Conversational agents in healthcare: a systematic review. *J Am Med Inform Assoc* 2018; 25: 1248–1258.
21. Tudor Car L, Dhinakaran DA, Kyaw BM, et al. Conversational agents in health care: scoping review and conceptual analysis. *J Med Internet Res* 2020; 22: e17158.
22. Mills R, Mangone ER, Lesh N, et al. Chatbots to improve sexual and reproductive health: realist synthesis. *J Med Internet Res* 2023; 25: e46761.
23. Latt PM, Aung ET, Htaik K, et al. Evaluation of artificial intelligence (AI) chatbots for providing sexual health information: a consensus study using real-world clinical queries. *BMC Public Health.* 2025; 25: 1788.
24. Bull S, Hood S, Mumby S, et al. Feasibility of using an artificially intelligent chatbot to increase access to information and sexual and reproductive health services. *Digit Health* 2024; 10: 20552076241308994.
25. Lyon B and Popov G. What-if analysis methods. In: *Risk assessment*. 1st ed. Hoboken: John Wiley & Sons, 2021, pp.137–151.
26. NHS Digital. DCB0129: clinical risk management: its application in the manufacture of health IT systems. *NHS Digital* 2018.
27. Min S, Lewis M and Hajishirzi H. FActScore: fine-grained atomic evaluation of factual precision in long form text generation [preprint]. arXiv 2023. Available from: <https://arxiv.org/abs/2305.14251>
28. Royal College of Obstetricians and Gynaecologists. Understanding how risk is discussed in health care. <https://www.rcog.org.uk/for-the-public/browse-our-patient-information/understanding-how-risk-is-discussed-in-health-care/> (2015, accessed 28 July 2025).
29. Tam D, Sellam T, Narayanaswamy B, et al. Evaluating the factual consistency of large language models through news summarization [preprint]. arXiv 2023. Available from: 10.48550/arXiv.2211.08412
30. Es S, James J, Espinosa-Anke L, et al. RAGAS: automated evaluation of retrieval augmented generation [preprint]. arXiv 2023. Available from: <https://arxiv.org/abs/2309.15217>
31. Ferrara E. Fairness and bias in artificial intelligence: a brief survey of sources, impacts, and mitigation strategies. *Sci* 2023; 6: 3.
32. Mojo. Mojo: your AI sex & relationship therapist. <https://mojo.so/> (2025, accessed 27 November 2025).
33. Planned Parenthood Federation of America. Roo: sexual health chatbot. <https://roo.plannedparenthood.org/> (accessed 27 November 2025).
34. Population Foundation of India. SnehAI. <https://snehai.org/> (2025, accessed 27 November 2025).
35. Nadarzynski T, Puentes V, Pawlak I, et al. Barriers and facilitators to engagement with artificial intelligence (AI)-based chatbots for sexual and reproductive health advice: a qualitative analysis. *Sex Health* 2021; 18: 385–393.

36. Balaji D, He L, Giani S, et al. Effectiveness and acceptability of conversational agents for sexual health promotion: a systematic review and meta-analysis. *Sex Health* 2022; 19: 391–405.
37. Wasson EJ, Driver K, Hughes M, et al. Sexual reproductive health chatbots: should we be so quick to throw artificial intelligence out with the bathwater? *BMJ Sex Reprod Health* 2021; 47: 73.
38. Abou Chawareb E, Im BH, Lu S, et al. Sexual health in the era of artificial intelligence: a scoping review of the literature. *Sex Med Rev* 2025; 13: 267–279.