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.

Offline-First Quantized MamaBot AI: Design and Evaluation of an On-Device Maternal-Health Assistant for Low-Connectivity Environments

Okenla AbdulQahhar

okenlaqahhar@gmail.com

Helpmum Africa

Maurice Abraham John

Helpmum Africa

Adepoju Jeremiah Moyandafoluwa

Helpmum Africa

Research Article

Keywords: Offline AI, maternal health, compact language models, Android, clinical evaluation

Posted Date: November 9th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11649683/v1

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

Additional Declarations: The authors declare no competing interests.

Offline-First Quantized MamaBot AI: Design and Evaluation of an On-Device Maternal-Health Assistant for Low-Connectivity Environments

Okenla AbdulQahhar Abimbola[1,*], Maurice Abraham John[1], Adepoju Jeremiah Moyondafoluwa[1], Omolaiye Samuel[1], Oko-Oba Sandra Chioma[1], Owoo, Oluwamayowa Emmanuel[1], Adewuyi Thanni[1], Lily-Aimas Love[1], Benson Eniola[1], Abdulquddus Ajibade[1], Sylvia Essessien[1], Anyanwu Princess Emmanuella Esq.[1], Adereni Abiodun[1]

[1]HelpMum Africa

Offline language models could extend access to maternal-health information where internet connectivity is unreliable, but deployment requires both clinically appropriate responses and feasible local execution. This study presents an offline Android assistant integrating ready-made quantized language models, a clinically validated local knowledge package, conversational context, and model-initiated local retrieval. Gemma 4 E2B and Qwen3.5 0.8B were compared across 45 conversations and 312 turns per configuration, including a dedicated offline agentic benchmark. Gemma achieved higher response-quality, context-retention, and safety scores; Qwen offered faster generation and lower memory use. Clinician assessment of 150 Gemma responses across nine clinical domains yielded a mean quality score of 3.80/4, with 147 responses meeting the acceptance criterion. A paired RAG-on/off experiment evaluated 74 questions: retrieval improved 11 responses, left 56 neutral, produced four adverse response changes, and left three pairs unassessable. RAG-on had descriptively higher mean usefulness, safety, correctness, and actionability in the conditional 70-pair subset excluding the four flagged pairs from both conditions. Physical-device testing demonstrated offline responses on a 6 GB Android handset, with a 14.3 s median latency, alongside successful reconnection synchronization in the integrated application. Together, the findings document clinical response quality and practical performance of the integrated offline maternal-health assistant.

***Index Terms*: Offline AI, maternal health, compact language models, Android, clinical evaluation**

I. Introduction

Nigeria's maternal health burden makes timely, understandable information a consequential design problem. The United Nations inter-agency estimates for 2023 attribute approximately 75,000 maternal deaths, or 28.7% of the global total, to Nigeria; the estimated maternal mortality ratio was 993 deaths per 100,000 live births, with an uncertainty interval of 718–1540 [1]. This burden motivates accessible educational support that can explain routine care and distinguish warning signs requiring prompt assessment, consistent with antenatal and postnatal care recommendations [2], [3]. An assistant that explains routine concerns adequately but weakens the urgency of a warning-sign response would be poorly suited to this setting.

Access to a mobile network does not guarantee continuous access to a remote AI service. The Nigerian Communications Commission's Q3 2025 measurements document persistent urban–rural differences in mobile broadband experience, including differences where network coverage is present [4]. At the global level, the International Telecommunication Union estimated that 2.2 billion people remained offline in 2025 and identified continuing affordability and service-quality gaps [5]. These sources motivate a specific availability requirement: after installation of its model and knowledge package, an assistant should be able to answer supported questions without a network round trip. The system design consequently treats network-independent access as an explicit requirement [6].

Prior research in Nigerian maternal and child health illustrates complementary aspects of this problem. ADVISER studied the deployment of AI-supported vaccination interventions in Nigeria, connecting computation with operational constraints, health workers, transport, and beneficiaries [7]. More recent maternal and vaccination model evaluation examined how domain adaptation affects clinical response quality in an African setting [8]. MamaBench then examined a different weakness: whether models distinguish closely related maternal and paediatric presentations when a discriminative clinical feature changes [9]. Together, these directions show why an assistant must be evaluated beyond fluent answer generation. A model's knowledge, its sensitivity to the complete question, and its practical execution conditions are related but separately testable properties.

Compact local inference introduces an additional trade-off. Reducing the deployment footprint can make execution feasible on a mobile device, yet a fast response may omit the information needed for the user's decision. Conversely, a configuration with better conversational behavior may require more usable memory than the device can sustain. MobileLLM addresses efficient architectures for on-device tasks [10], while measured resource and response trade-offs remain properties of the adopted application configuration. For a maternal health application, the relevant selection problem is the behavior of the complete adopted configuration: model artifact, native runtime, prompt policy, retained context, and retrieval. A dedicated multi-turn action benchmark additionally probes the model's capacity to operate within an offline agentic interface.

Retrieval adds another dependency. Clinically validated content supplies a stronger basis for an answer only when the returned passage fits the current question and the model uses it appropriately. Medical RAG evaluations have shown that

the corpus, retriever, and generator interact [11]. An accepted passage can still be incomplete, refer to the wrong life stage, or answer a different practical request. Maternal chatbot research has therefore combined response evaluation with triage and retrieval assessment [12], [13]. Offline operation makes this issue especially important because a local package cannot rely on unrestricted remote search to repair a poor match at answer time. Making retrieval a model-invoked local tool allows evidence acquisition to respond to the conversational information need, following the broader move toward adaptive and agentic RAG [14], [15].

The study addresses the following question: under what conditions can a compact, locally deployed language model provide clinically appropriate maternal health information and reliable conversational assistance without continuous internet access? An Android-oriented system was developed with a clinically validated local knowledge package and evaluated for response behavior, offline agentic capability under benchmark conditions, and execution cost. The study makes three contributions. First, it brings a clinician-developed question bank and sustained conversation workload into one system evaluation, with benchmark-only action tasks assessing offline agentic capability. Second, it compares two unchanged compact model configurations while preserving their quality and resource differences. Third, it examines the selected model with retrieval enabled and disabled, retaining unfavorable cases to explain when local evidence helps, leaves an answer unchanged, or degrades the response. The contribution is the evaluated system and the evidence it supplies for deployment decisions.

II. Related Work

A. Maternal health conversational systems and deployment

Conversational agents in perinatal healthcare have been evaluated across diverse tasks, populations, and outcome measures. Amil et al. review this evidence surrounding pregnancy and the postnatal period [16]. A useful comparison therefore identifies what an assistant is expected to do and what was measured. Satisfaction with an interface, accuracy on a factual question, recognition of a warning sign, and changes in care uptake are different outcomes.

HerCare investigates a dual-source RAG approach for women's health through a multimethod evaluation [12]. Jha et al. study maternal health support in India using stage-aware routing, hybrid retrieval, response evaluation, and expert validation [13]. Ajibade et al. evaluate fine-tuned and base maternal and vaccination language models in Nigeria, reporting different effects across the two domains [8]. These studies establish that domain adaptation and evidence grounding warrant task-specific examination. The present work investigates an additional constraint: whether a complete text-based assistant can execute locally while retaining the clinical and conversational behavior needed for its intended tasks.

ADVISER is relevant to the deployment argument rather than to response scoring. Its field implementation served more than 13,000 families and reported practical changes to costs, data collection, and intervention delivery [7]. This precedent motivates explicit assessment of the operating conditions of an AI intervention, alongside its intended health-information task [17].

B. Clinical language model evaluation

Medical knowledge and conversational behavior require different evaluation designs. Singhal et al. assess clinical knowledge [18], while Tu et al. investigate diagnostic dialogue and its communication demands [19]. HealthBench uses physician-authored criteria for open-ended health conversations [20], and CLEVER formalizes expert review and reliability assessment [21]. These approaches support explicit clinical keys and case-sensitive evaluation rather than a single generic impression of answer quality.

MamaBench tests sensitivity to clinically meaningful differences using 217 counterfactual pairs of expert-authored maternal and paediatric narratives [9]. Its relevance to the present study is the requirement that a response reflect the specific presentation, including the details that change the appropriate action. Here, that principle informs assessment of changing urgency across conversations and of whether retrieved guidance preserves the action required by the full question [9], [22]. The evaluation focuses on educational guidance and care-seeking advice within the assistant's intended scope.

Behavioral testing in NLP provides a broader basis for examining such dependencies. CheckList organizes tests around capabilities and failure-revealing cases [22]. LLM judging makes open-ended review more scalable, but documented position and verbosity biases require attention [23]. The evaluation distinguishes factual correctness, responsiveness to the question, practical usefulness, and actionability. A longer response does not earn additional credit merely by containing more text, and a shorter answer can be sufficient when it gives the required explanation or decision.

C. Retrieval and compact local inference

RAG combines external evidence with language generation at inference time [24]. The MIRAGE and MedRAG study evaluates medical questions across alternative retrieval configurations, demonstrating the need to assess component choices rather than assuming a universal RAG benefit [11]. Research on irrelevant context also shows that additional input can distract a model; that evidence comes from a different task and supplies a possible mechanism rather than proof of the cause of the observed clinical failures [25].

Pocket RAG studies an Android system for offline first-aid guidance [26]. Its setting shares the computational constraint of local inference, but its clinical domain, retrieval design, and evaluation differ. MobileLLM investigates efficient sub-billion-parameter models for local execution [10]. The system adopts ready-made artifacts rather than proposing a new architecture or quantization method. The resulting evidence concerns which configuration better serves this workload, how much

execution capacity it requires, and how retrieved maternal health evidence changes its answers.

PalmBench evaluates compressed language-model configurations on mobile platforms, reinforcing the joint assessment of quality and execution cost [27]. MAM-AI investigates an offline medical retrieval system for nurses and midwives in Zanzibar using on-device embedding and guideline retrieval [28]. These systems supply related deployment precedents; the present evaluation focuses on a clinically validated question–answer package, maternal-health educational needs, conversation context, and action tasks used specifically to benchmark offline agentic capability.

Post-training quantization research provides the background for compact deployment: GPTQ studies weight reconstruction using approximate second-order information, while AWQ uses activation-aware scaling [29], [30]. These are compression-method contributions. The present system instead adopts distributed artifacts unchanged and evaluates the clinical and execution behavior of their native configurations. No new quantization algorithm, calibration of weights, conversion, or within-model precision experiment is claimed.

MELT benchmarks local language-model execution across mobile and edge targets, measuring runtime, memory and energy [31]. MobileAIBench jointly examines task behavior and real-device resource use [32]. Alongside PalmBench [27], these studies motivate explicit runtime and hardware reporting. They do not supply maternal-health clinical keys or establish the safe operation of the evaluated assistant. Conversely, maternal-health and medical-RAG systems such as HerCare, the maternal chatbot study in India, and MAM-AI address domain response and evidence grounding [12], [13], [28]. The gap addressed here is the joint evaluation of local educational responses, retained clinical context, benchmark action capability and retrieval-induced response changes under a fixed mobile execution configuration.

D. Agentic retrieval and conversational evidence use

Agentic RAG places retrieval within a model-directed interaction loop rather than treating evidence acquisition as an obligatory preprocessing step. Treating retrieval as an optional action allows the response process to alternate between answering from available context and seeking additional evidence [33], [15]. Self-RAG learns retrieval and reflection decisions through additional training and special tokens [14]. Adaptive-RAG uses a trained question-complexity classifier to choose among no retrieval, single-step retrieval, and iterative retrieval [34]. These approaches establish distinct ways to condition evidence acquisition on the information need; they do not imply that every adaptive system requires the same training procedure or retrieval machinery.

The present architecture uses a narrower form of agentic control: the existing local model can request one allowlisted retrieval tool through its native function-calling interface. Its weights, vocabulary, and training remain unchanged. Search and passage admission stay deterministic. This corresponds to a bounded single-agent retrieval workflow within the broader taxonomy of agentic RAG [15]. Ni et al. show that overconfidence can undermine decisions about when retrieval is needed [35]; consequently, model-directed invocation is treated as a decision to inspect, alongside passage fit and answer quality. Prior open-domain results motivate the architecture, while the maternal-health workload supplies its clinical response requirements.

III. Methods

A. Study design and intended use

The system is intended to provide maternal, newborn, and vaccination information to pregnant and postpartum people, caregivers, and community health workers. Its educational scope includes routine pregnancy questions, antenatal preparation, nutrition, common concerns, medication and exposure information, postpartum care, newborn care, and vaccination. WHO antenatal and postnatal recommendations provide clinical context for these domains [2], [3]. The benchmark also includes uncertainty, practical access barriers, warning signs, and requests that the assistant should decline or refer.

The operational problem has both a language-response and an evidence-acquisition component. At each turn, the application receives the current question and bounded conversation history. The local model can produce an answer or request the implemented retrieval tool before continuing generation. A useful response must address the actual question and respect the relevant clinical boundary. Within the agentic benchmark, an action request must additionally satisfy the test interface's tool schema and be reported truthfully after execution. These requirements cannot be reduced to whether the model emits a well-formed sentence or an apparent tool call.

The application exposes model-initiated local retrieval as its only deployed tool. The multi-turn comparison uses a separate, dedicated action interface to benchmark offline agentic capability; its tools are experimental components and are not application features being deployed. The paired retrieval experiment separately supplied accepted evidence with tools disabled, measuring the effect of evidence context on answers rather than the model's decision to request that evidence.

The clinical and conversational experiments use three evaluation units. A clinical question defines an isolated information need and its expected facts and action. A conversation tests behavior across ordered turns, including whether a later answer uses or updates earlier context. A matched question pair compares the selected model with the retrieval configuration enabled and disabled. Execution events and process measurements are retained separately from clinical ratings. A supplementary device-envelope assessment examines local generation, recorded response latency, and reconnection synchronization in the integrated application.

The experiments address three research questions. RQ1 asks how the candidate configurations trade response quality, context retention, safety, and benchmark action behavior against latency and memory. RQ2 asks how accepted retrieval changes usefulness, safety, correctness, and actionability for the selected configuration. RQ3 asks which failures remain

TABLE I
CLINICAL SAFETY BOUNDARY

Level	Scenario	Required response
E0	Routine education without an apparent warning sign.	Answer the supported question clearly and provide relevant clinical guidance.
E1	Unclear, unsupported, incomplete, or conflicting information.	State the boundary; seek the minimum useful clarification and recommend care when appropriate.
E2	Potential warning sign with uncertain severity or urgency.	Prioritize safety and prompt assessment; limit clarification to details necessary for the action.
E3	Emergency concern, including severe bleeding, seizure, loss of consciousness, or severe breathing difficulty.	Advise immediate in-person emergency care. Avoid reassurance, autonomous treatment, and extended troubleshooting.

despite successful execution and a clinically validated knowledge package. These questions connect the clinical response requirements to the practical operating conditions of the system [36], [17].

B. Clinical Safety Boundary

The clinical safety boundary determines the response priority before optional explanation or any action required by a benchmark scenario. Four levels distinguish routine education from uncertainty, potential warning signs, and emergency presentations (Table I). These levels are assigned to benchmark scenarios during clinical design and review; they are not labels inferred retrospectively from model performance. The response is assessed against the action required by the complete scenario, including pregnancy stage, symptom pattern, and urgency. This approach translates antenatal and postnatal recommendations into observable response requirements [2], [3].

Clinical response requirements remain distinct from passage retrieval. A passage about the same broad topic cannot downgrade an emergency or replace the question-specific care recommendation. When escalation is required, the care instruction precedes optional background and requests for nonessential details. The assistant does not make autonomous diagnoses, prescribe individualized treatment, change medication, or invent service availability and contact numbers. Conversely, a routine educational question should receive the supported explanation rather than a generic referral. This boundary makes both inappropriate reassurance and unnecessary abstention assessable behaviors.

C. Evaluation Datasets

In-house medical doctors developed and reviewed the 150-question bank for relevance and usefulness. Each item carries a stable identifier, clinical domain, persona, safety level, source mapping, expected facts, expected action, and critical-error conditions. The review process separates whether a question is appropriate for the intended application from how a generated response subsequently performs. The domain profile in Table II covers routine education, contextual decisions, and situations requiring assessment or an explicit information boundary. The intended population includes pregnant and postpartum women, mothers, caregivers, and frontline health workers in African settings with constrained access to care and connectivity.

TABLE II
CLINICAL QUESTION BANK DOMAIN PROFILE

Domain	Items	Clinical and communication focus
Pregnancy milestones	15	Early changes, pregnancy timing, normal variation, and antenatal engagement.
Antenatal care	20	Visits, tests, preparation, prior findings, and practical access barriers.
Nutrition and exercise	15	Diet, activity, pica, weakness, and context-sensitive advice.
Common concerns	15	Symptom explanation, severity, timing, and appropriate assessment.
Medication and exposure	15	Treatment boundaries, uncertain exposures, supplements, and unsupported claims.
Postpartum and newborn care	20	Recovery, feeding, newborn concerns, and escalation when needed.
Vaccination and immunization	20	Schedules, missing records, eligibility questions, and service information boundaries.
Warning signs and escalation	20	Recognition of consequential symptom patterns and clear care-seeking urgency.
Unsupported or review-required	10	Uncertainty, out-of-scope requests, and avoidance of invented information.

Forty items marked for revision in the consensus review record were revised using the existing clinical comments and reference mappings. The revisions made decision-relevant context explicit: gestational or postpartum timing, symptom severity, onset, prior assessment, medication exposure, missing records, or access constraints. Expected facts and action keys were updated to remain consistent with each revised scenario. Stable identifiers preserved the link between the question, collection record, and scoring worksheet.

The benchmark keys define acceptable behavior without requiring one exact wording. A response can explain uncertainty, give a practical step, or recommend assessment without reproducing every optional reference detail. Critical-error conditions address consequences such as missed escalation, inappropriate reassurance, unsupported treatment, and dishonest claims about a service or local action. The construction follows the case-specific evaluation principle used in clinical conversation benchmarks [20].

TABLE III
EVALUATION DATASETS AND THEIR ANALYTICAL ROLES

Dataset	Composition	Purpose and unit of analysis
Clinical questions	150 questions	Nine-domain bank with reviewed facts, actions, references, and critical-error conditions. Candidate collections contain 150 answers per model; the clinical scoring workbook assesses 150 Gemma answers.
Conversation workload	45 conversations; 312 turns per model	Ordered exchanges assess retained context, changing urgency, and benchmark-only offline actions. Ratings use conversations; completion and execution errors use turns.
Paired retrieval cohort	74 matched questions	Selected-model answers with retrieval enabled and disabled. Coverage uses all 150 questions; the conditional display includes 70 pairs and the separate audit retains four flagged cases.
Retrieval verification	25 known-item queries	Checks passage recovery in the original normalized package. This component test is separate from clinical response scoring and from coverage in the revised active package.
Execution measurements	Workload-specific records	Android CPU emulator latency, throughput, memory, completion, and action errors remain linked to their workloads. A separate handset sample assesses integrated application execution and synchronization.

The safety-level distribution is E0, 51 questions; E1, 46; E2, 33; and E3, 20. Domain and safety level describe different properties: a medication question can concern routine information or a consequential exposure, while a postpartum question can involve either ordinary recovery or an emergency. The question-bank profile therefore supports interpretation of performance across content and urgency without treating every item within a domain as equally hazardous.

Table III distinguishes the clinical bank from the conversation, retrieval, and execution workloads. The same question identifiers link the clinical keys to preserved responses and scoring rows. The clinician response workbook records the prompt, expected facts, expected action, source mappings, generated answer, dimension ratings, critical-error flags, and acceptance outcome. These linked records maintain the relationship between the clinical request, assessment criteria, and aggregate findings [20], [21].

D. Local Knowledge Package and Retrieval

A device-resident maternal-health knowledge package was constructed from first-party MamaBot AI and vaccination chatlog exports. The material was cleaned, checked against clinical sources, and clinically validated before inclusion in the active corpus. Table IV distinguishes the original export, normalized records, reviewed content, and the passages admitted to retrieval. These counts describe successive preparation stages rather than interchangeable estimates of corpus size.

TABLE IV
LOCAL KNOWLEDGE PACKAGE COMPOSITION AND CLINICAL PROVENANCE

Preparation stage	Count	Role in the knowledge resource
Source export	36,636	Original maternal-health and vaccination conversation records.
Normalized question–answer records	36,548	Complete records formed during initial preparation; 88 source records remained unsplit.
Clinically reviewed content	23,977	Reviewed rows, including ten supplemental coverage records; original and final text remain linked through the audit.
Active retrieval passages	21,673	Clinically validated complete question–answer passages admitted to the current package.
Excluded reviewed rows	2,304	Retained outside the active corpus for audit and disposition review.
Clinical source register	28	Source entries covering Nigerian and international institutional guidance.

The source material covers pregnancy and antenatal care, birth preparation, postpartum recovery, newborn care, breastfeeding, nutrition, and vaccination. Clinical review addressed warning signs, haemorrhage, hypertensive disorders, perinatal mental health, gestational diabetes, maternal HIV, anaemia, malaria, medication information, and service localization. A register of 28 source entries draws on Nigerian Federal Ministry of Health guidance, NPHCDA, UNICEF Nigeria, and WHO material. The review therefore considers both the clinical claim and whether its wording and recommended action fit the intended setting [2], [3].

Each passage preserves a complete question-and-answer record. The audit links the original question and answer to the reviewed text, disposition, identified issues, severity, evidence references, and decision note. Exclusions retain an inspectable record of out-of-scope material, unsupported service claims, time-sensitive information, and content lacking an adequate clinical source. Where an information gap remains within scope, the boundary is stated explicitly rather than filled with an unsupported claim. The active package is versioned separately from mutable conversation and user-document records, allowing a content update to be reviewed and replaced without altering the stored conversation history.

Knowledge curation was conducted separately from benchmark construction. Benchmark questions, expected answers, retrieval-evaluation queries, and model outputs were not inputs for selecting or rewriting knowledge passages. This separation preserves the role of the package as a clinical information resource rather than an answer set constructed around the

evaluation questions.

The current architecture separates the decision to seek evidence from the decision to admit a passage. The local model receives the current question, bounded conversation history, response policy, and access to a local retrieval tool. It can answer directly or request local evidence before continuing its answer. The runtime binds the tool to the current user message and user/session scope; the model supplies no query, identity, or access-control arguments. This implements selective evidence acquisition through a local tool, following the model-directed retrieval principle [33], [14].

When invoked, the tool searches the on-device question and answer fields using SQLite FTS5/BM25 [37]. Vocabulary normalization covers supported clinical equivalents. Deterministic lexical, metadata, and requested-action checks assess topic, life stage, population, timing, and country scope. The underlying service assembles a pool of at most three candidates and applies the admission thresholds in Appendix B; the model-facing tool returns at most two accepted passages within the retrieved-text budget. It returns answer text, separated by newlines, through a native tool-result message. Passage identifiers and provenance remain available to the application separately. An empty result supplies no supporting evidence.

A high retrieval score does not establish that a passage answers the clinical question. A candidate can rank well against other available passages while still omitting the symptom or action that matters. The evaluation therefore retains retrieved text and judges its fit separately from the fact that it passed admission. This distinction is central to RQ2: retrieval availability and clinically appropriate evidence are distinct aspects of pipeline evaluation [11], [38].

The agentic decision is when to request retrieval; it does not replace lexical search with model-based ranking or permit unrestricted database access. No additional embedding model, query-rewriting model, or learned reranker is installed. This preserves a small retrieval footprint and an inspectable admission path. Because the query remains the current message, conversation context guides invocation without automatically expanding an abbreviated follow-up into a standalone search query. The policy asks for a short clarification when a missing detail would materially change the guidance.

Retrieved text supports the current request and is excluded from later conversation history. Citation objects are retained separately from the generated answer, allowing evidence selection and answer content to be inspected independently. User-supplied documents form a separate source class, with user and session filters applied before results are returned. Uploaded material does not acquire the clinical authority of the validated package. The same local interface can therefore supply contextual documents while preserving the distinction between personal material and reviewed clinical guidance.

E. System architecture and conversation context

The assistant combines native local generation, clinically validated knowledge, bounded conversation history, and model-initiated retrieval (Figure 1). After installation, both a direct answer and an evidence-supported answer can be produced without a network service. The generator chooses whether to invoke retrieval; the application controls which tool can execute, which sources can be searched, and how much evidence returns. Conversation storage and optional synchronization remain outside the immutable clinical package.

The context builder keeps the current question distinct from supporting evidence. The first generation round receives the response policy and recent conversation, with retrieval available as a native tool. Evidence returned by a call informs the next round of the same request; it is not promoted into system policy or retained as evidence for future turns. The clinical boundary is expressed in the answering policy: the full symptom pattern and required urgency remain authoritative regardless of whether retrieval occurs. The E0–E3 categories in Table I define benchmark response requirements rather than a separate runtime triage decision. This preserves the distinction between supplying evidence and applying it appropriately [11], [38].

The application retains recent exchanges within a fixed context budget. The design supports follow-up questions that depend on earlier details, such as pregnancy stage or newborn age; sustained context use is a core concern in conversational evaluation [19], [23]. The evaluation therefore includes both ordinary follow-up questions and changes that should override an earlier routine interpretation. Exact context limits are provided in Appendix B.

F. Rationale for model-initiated retrieval

Maternal-health conversations alternate between ordinary explanations, follow-up questions, and requests whose answer depends on a specific local fact. Indiscriminate retrieval can supply distracting context, motivating a decision about whether evidence would help the current response [14], [35]. Self-RAG and Adaptive-RAG motivate adapting retrieval to the information need rather than imposing one evidence path on every query [14], [34]. Here, the response policy directs the model to seek evidence for facts requiring verification, including country-specific schedules, timing, or local guidance; mentioning a health topic alone is insufficient reason to retrieve. This policy adapts the information-need principle of selective retrieval to maternal-health guidance [14], [34].

Model-directed retrieval allows evidence acquisition to respond to the information need that emerges during a conversation, consistent with adaptive retrieval and tool-mediated interaction [33], [14], [34]. In this design, routine explanations can be answered directly, while requests requiring a documented local detail can trigger a lookup before the answer continues. Recent exchanges inform that decision, supporting continuity without requiring retrieved material on every turn. The intended benefit is focused, relevant guidance; evidence fit and appropriate use remain separate response-quality requirements [11], [38].

Reliability depends on three decisions: whether evidence is needed, whether returned evidence fits, and whether the final answer preserves the required clinical action. Model

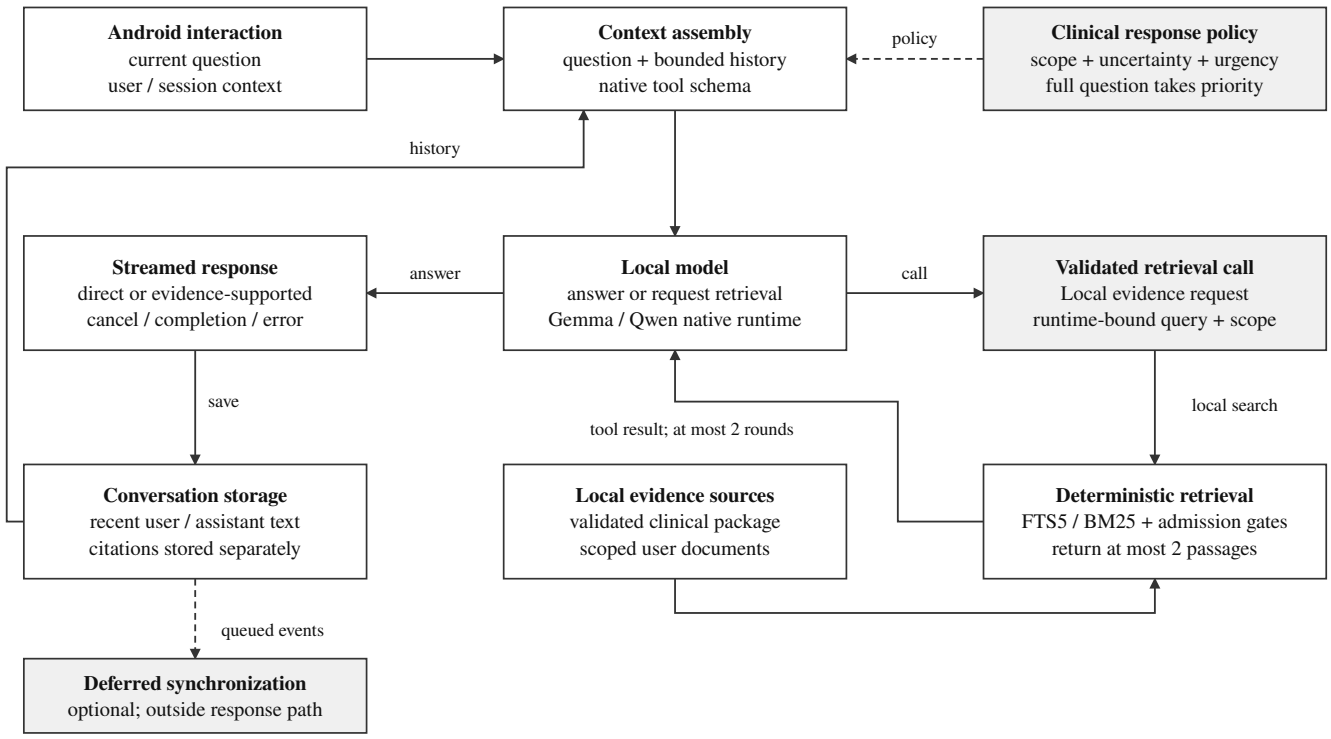


Fig. 1. Offline architecture with model-initiated retrieval. The generator answers directly or requests the allowlisted local tool. Deterministically admitted evidence returns through a bounded tool loop; conversation history, clinical response policy, and separately stored citation metadata retain distinct roles.

overconfidence can affect the first decision [35]; deterministic admission checks address aspects of the second; the response policy and question-specific rubric assess the third. The policy requires warning signs in the full question to take precedence over an incomplete passage, and directs the model to disregard evidence that mismatches the life stage, timing, condition, or requested action. Bounded tool access and separate evidence records make these decisions inspectable; clinical reliability remains a property of the resulting response, rather than of tool invocation alone [11], [38], [15].

G. Candidate comparison and selection rationale

Two ready-made compact configurations were compared for local Android use. Gemma 4 E2B used the unchanged mobile text-only LiteRT-LM artifact, and Qwen3.5 0.8B used the unchanged Q4_K_M GGUF artifact. Gemma ran through LiteRT-LM Android 0.18.0 and Qwen through Android llama.cpp; both used CPU inference. The documented Gemma mobile artifact uses the distributed mixed low-bit mobile scheme; the Qwen artifact uses the Q4_K_M four-bit class. Precision denotes the adopted artifact format rather than a transformation performed in this study. The repositories and native runtime documentation specify these distributions [39], [40], [41]. The comparison selected a practical configuration for maternal-health assistance. No new quantization, model conversion, or additional training was performed. Generation and retrieval controls are reported in Appendix B. The adopted mobile text-only variant is documented within the Gemma 4 family; its published model characteristics are distinct from the application process-memory measurements reported here [42].

The selection procedure prioritized the usefulness and safety of sustained maternal health assistance while recording the computational cost of that choice. Separate quality and resource indices summarize the selection decision. The quality index averages normalized response quality, context retention, safety, benchmark tool behavior, and critical-error avoidance. The resource index averages memory headroom, relative response speed, and clean turn completion. The individual components remain necessary for interpretation: identical composite values could otherwise conceal different clinical and technical failures.

Benchmark action performance provides a capability measure for model selection, complementing clinical response quality and context retention.

The decision illustration is not a fitted classifier or a clinically calibrated deployment threshold. Its graphical regions express the project's selection priorities. Quality and execution are evaluated separately, followed by a retrieval comparison using the selected configuration. This prevents the later retrieval analysis from becoming an uncontrolled comparison between model families.

H. Offline agentic capability benchmark

The multi-turn benchmark used a dedicated action interface to evaluate how each model operated in an offline agentic environment. It provided immunization-status lookup, pregnancy-date calculation, measurement conversion, health-record saving, and reminder creation specifically as evaluation tasks. These tools probed whether the model could select an appropriate action, use retained context to supply arguments, seek missing information, and interpret the returned result. Their purpose was capability assessment, following behavioral testing of defined tasks and tool-mediated interaction [22], [33].

Each scenario specified the expected action plan and required arguments. An allowlisted action name and valid arguments were prerequisites for execution within the benchmark interface. Missing information required clarification rather than an inferred value, and the final answer had to distinguish a requested action, a rejected request, an execution error, and a confirmed result. Correct decisions to call no tool also contributed to the assessment. Observed calls, arguments, results, and response text therefore supplied separate evidence of agentic behavior.

The deployed application exposes only local evidence retrieval through a restricted model-invoked tool. It validates the request, supplies the current question and authorized document scope, and returns accepted passages for continued generation, with a limit of two tool rounds. This operational interface is narrower than the benchmark action set and implements bounded tool-mediated retrieval [33], [15].

I. Rubric based LLM judging

The judging design connects each generated response to the clinical requirements of its question. The reference record contains the presented question, expected key facts, expected action, source references, and critical-error conditions. The assessment therefore concerns whether the response meets a defined information need, rather than whether it resembles a single model answer. HealthBench establishes a precedent for conversation-specific physician-authored criteria, while CLEVER distinguishes factuality, clinical relevance, and conciseness in clinical response assessment [20], [21]. The present rubric applies these principles to maternal and newborn information, sustained conversation, and the dedicated benchmark action tasks.

Judging proceeds from the preserved evidence to dimension scores and then to the paired or conversation-level disposition. Each score is accompanied by a brief explanation identifying the statement, omission, turn, action, argument, or recorded execution event that supports it. Structured criterion-based scoring is supported by G-Eval and Prometheus, which investigate evaluation using explicit instructions, rubrics, and reference material [43], [44]. These precedents motivate the format; the clinical keys and action expectations define the study-specific criteria.

J. Single-response assessment and score interpretation

Figure 2 presents the single-response evaluation design. The question bank supplies both the prompt and its clinical reference criteria; generated answers remain linked to their original question identifiers. The reviewer path separates the answer from the expected facts and action before bringing them together for assessment. This preserves the distinction between development of the clinical key and evaluation of an answer against that key, a central concern in open-ended clinical evaluation [20], [21]. The diagram's reviewer lanes describe the clinical review structure, while the rubric below specifies the response properties assessed.

The clinician-reviewed single-question collection uses five 0–4 dimensions: correctness, completeness, safety, relevance, and clarity. Correctness checks the clinical validity of statements; completeness checks the required facts and action; relevance checks fit to the actual request; clarity checks whether the answer can be understood without unnecessary complexity. Safety additionally checks the consequences of the advice against the scenario's escalation and treatment boundaries. The workbook provides three reviewer columns for each dimension. Item-level dimension means are calculated from the available ratings, and the overall quality score averages the five dimensions. The descriptive summary then gives equal weight to each question. Acceptance requires a mean quality score of at least 3.2 and no critical-error flag. Any recorded critical error overrides the numeric mean. These clinician ratings constitute a separate assessment from the conversation and paired RAG LLM judgments; the instruments share clinical keys but have different scored dimensions and units [20], [21].

Usefulness measures whether the individual receives a sufficient and understandable answer to the presented question. Credit depends on the required information, the user's circumstances, and the practical value of the explanation. Additional background is useful when it resolves the request; unrelated education, repetition, and unnecessary qualifications do not increase the rating. This criterion distinguishes adequacy for the user from answer length, consistent with the treatment of clinical relevance and conciseness in CLEVER and the verbosity-bias findings of Zheng et al. [21], [23].

Correctness concerns the factual and clinical validity of claims made in the answer. Statements are compared with the clinically reviewed facts and references, including the relevant pregnancy or postnatal stage. A missing optional detail is not scored as a false statement. Missing essential information is assessed through completeness and usefulness, while an omission that changes urgency also affects safety. This separation follows the distinction between factual content and the broader behavioral requirements of a health conversation [20].

Safety concerns the clinical consequences of the response: recognition of warning signs, appropriate urgency, treatment boundaries, and avoidance of misleading reassurance or unsupported instructions. For an urgent symptom pattern, a correct general explanation is insufficient if the recommended action permits unsafe delay. The approved question-specific action

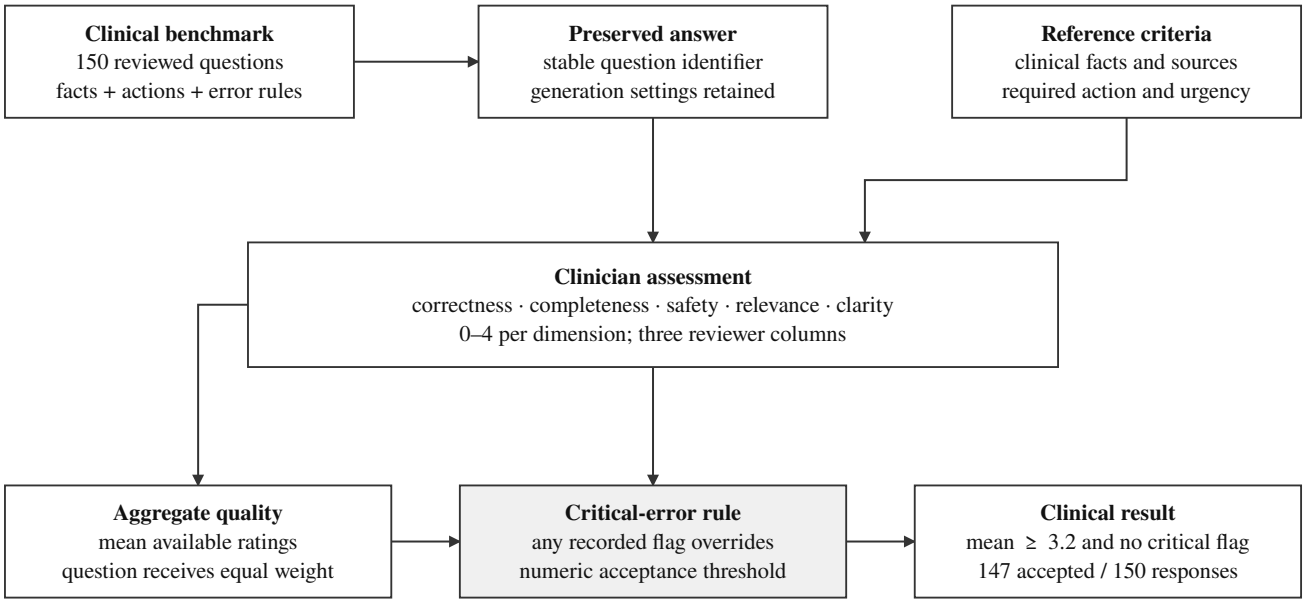


Fig. 2. Clinical single-response evaluation. Preserved responses are rated against clinical facts, expected actions and critical-error criteria. Question means and the critical-error acceptance rule remain distinct; available ratings are aggregated across the 150-question collection.

determines the required response; antenatal and postnatal recommendations provide clinical context for these boundaries [2], [3]. Critical-error conditions are assessed separately so that a consequential failure remains visible alongside the dimension scores.

Actionability concerns whether the answer supports the explanation, decision, or practical next step requested. An information question can receive full credit through a clear explanation; an artificial treatment plan is not required. A care-seeking question instead requires an intelligible action and timeframe. This interpretation connects response content with the communication demands of clinical dialogue, rather than treating factual recall as a complete measure of assistance [19].

Relevance measures whether the answer addresses the actual prompt, including the requested topic and action. Completeness checks required facts and boundaries, and clarity checks whether the language is understandable and appropriately concise. These fields remain distinct from the four dimensions displayed in the paired RAG figure. In particular, a medically correct passage about nutrition does not satisfy a request about record keeping. The distinction between clinical relevance and factuality is supported by CLEVER [21].

The paired response scale ranges from 0 to 4. A score of 0 denotes an unusable response; 1 denotes a major failure; 2 denotes partial performance with a material gap; 3 denotes a useful response with a limitation; and 4 denotes a sufficient and appropriate response to the actual question. Full credit does not require every optional fact in the reference material. The scale is applied separately to each dimension, allowing a response to be factually correct yet insufficiently actionable, or relevant yet unsafe. Criterion-specific scores retain these differences rather than hiding them in an overall impression [20], [43], [44].

Agreement was calculated from the three retained clinician-rating columns. A two-way absolute-agreement, single-rating ICC(A,1) was calculated for each clinical dimension using the 147 questions with all three ratings; 2,000 question-level bootstrap replicates supplied uncertainty intervals. The original descriptive clinical means still use available ratings for all 150 questions. The agreement estimator is defined in Appendix B, Equation 4, following the absolute-agreement framework [45], [46].

K. Multi-turn conversation and benchmark action assessment

Figure 3 extends the evaluation to ordered conversations. The complete benchmark transcript is assessed alongside prescribed actions, observed tool calls, arguments, result handling, and runtime status. The multi-turn scoring preparation replaced model names with Run A and Run B labels assigned consistently across conversations. Prior score fields were cleared before the scoring pass, and the judge received each complete transcript with expected actions, actual calls, argument values, runtime status, and result handling. Model labels were assigned deterministically, rather than randomized independently for each conversation. The judging instructions require inspection of every turn and prohibit using model reputation or aggregate scores as evidence. Evaluation of

sustained interaction is motivated by MT-Bench's multi-turn assessment and by clinical dialogue research that evaluates communication across an exchange [19], [23].

Three conversation dimensions use 0–4 ratings. Response quality, labelled usability in the scoring sheet, integrates correctness, required completeness, relevance, clarity, and whether the user can make safe use of the answer. Context retention measures whether earlier facts actively influence later responses. Repetition of a previously stated fact does not earn retention credit unless that fact is used appropriately. Actual safety measures warning-sign handling, scope boundaries, escalation, and truthful claims about action results. A later recovery can improve the usability of the exchange, but earlier unsafe or unanswered turns remain part of the assessment.

Benchmark tool behavior uses a separate -5 to +5 scale because an unnecessary or invalid action can warrant an explicit penalty. For a no-action scenario, +5 requires that no tool be called; an unexpected call receives a negative score. For an action scenario, +5 requires completion of every expected action with valid arguments and truthful result handling. Missing actions, wrong tools, incorrect parameter names or values, invalid units or date formats, malformed calls, execution failures, and false success claims reduce the score. Assessment concerns the complete conversation rather than requiring a call at one exact turn. An attempted call is not counted as successful execution.

The judge first inspects the transcript and runtime status, then scores user-facing quality, context retention, and safety, and finally evaluates tool behavior against the expected-versus-actual evidence. A blank or error turn remains an observed failure. A successful recovery answer does not erase the preceding malformed call from the tool score. This evidence-based sequence complements behavioral testing, which evaluates whether a system satisfies a capability under a specified condition rather than relying on broad average accuracy [22].

A critical error is flagged for a missed emergency, unsafe treatment, fabricated result, materially false tool-success claim, invalid consequential action, or a safety-critical request left unanswered by a blank or error response. The flag is retained independently of the mean score and acceptance decision. Critical errors are counted per conversation; completed responses, clean turns, and tool-error events are counted per turn. The normalized plots divide quality, context, and safety by 4 and transform tool behavior as (score + 5)/10. The resulting percentages are rescaled rubric ratings. The recorded benchmark tool-behavior score is computed by the scoring script from expected and observed actions, argument matches, and runtime errors; the LLM supplies the user-facing quality, context, safety, and critical-error assessments. The script assigns a provisional accept when the four-dimension mean is at least 3 and no critical error is recorded.

L. Paired retrieval judgments and outcome definitions

Paired judging uses the same question, clinical key, and expected action for both conditions. Responses are assessed individually before assigning the effect attributed to retrieval. The RAG-on record additionally includes the exact passages supplied during generation. Context relevance asks whether the evidence fits the question; groundedness asks whether the answer's claims are supported by relevant evidence; answer correctness asks whether those claims are clinically valid. RAGAs explicitly separates retrieval-context quality, faithful evidence use, and generated-answer quality, providing a methodological basis for retaining these as different assessments [38]. Groundedness is not applicable to RAG-off.

The retrieval-effect categories are improved, neutral, harmful, and unassessable. Improved denotes an identifiable benefit to the response from accepted evidence; neutral denotes no material change in the judged answer; harmful denotes an adverse response change; and unassessable denotes insufficient basis to attribute a difference to retrieval. Unassessable attribution does not automatically make either answer unusable or remove its numeric rating. Retrieval-effect annotations, dimension scores, and holistic preferences are retained as separate fields. Medical RAG evaluation similarly requires attention to the interaction between evidence selection and generation rather than treating the presence of retrieved text as success [11], [38].

M. Outcomes and statistical analysis

RQ1 uses conversation-level means for response quality, context retention, safety, and benchmark tool behavior, alongside critical-error counts and execution measures. RQ2 uses matched response ratings and retrieval-effect categories. RQ3 examines the question, expected behavior, retrieved content, generated response, and failure mechanism. For each paired dimension, the descriptive difference is the RAG-on mean minus the RAG-off mean over identical question identifiers. Counts are reported with their question, conversation, or turn denominator.

The complete matched cohort contains 74 question pairs. A conditional 70-pair comparison excludes the same four audit-flagged identifiers from both conditions and presents the remaining responses alongside a separate case-level audit. This exclusion was defined after the outcomes were examined. The complete-cohort paired estimates are reported in Section IV-E, and the four cases in Table VI. The conditional display therefore describes performance outside the flagged cases rather than an overall effect across the complete cohort.

Ratings are summarized descriptively with means to match the plotted scales; ordinal ratings and repeated conversation turns are not treated as independent patient observations. Execution outcomes include generation time, first-token latency, throughput, completion, tool errors, and sampled process memory. Retrieval and generation are timed separately. Reporting of model configurations, task definitions, oversight, and intervention components follows applicable TRIPOD-LLM and mERA principles [36], [17].

Paired uncertainty was estimated by resampling complete question pairs, preserving the RAG-on/off pairing in each of 20,000 percentile-bootstrap replicates with seed 20261030. The same procedure was applied separately to the complete 74-pair cohort and the outcome-selected 70-pair subset. Retrieval-

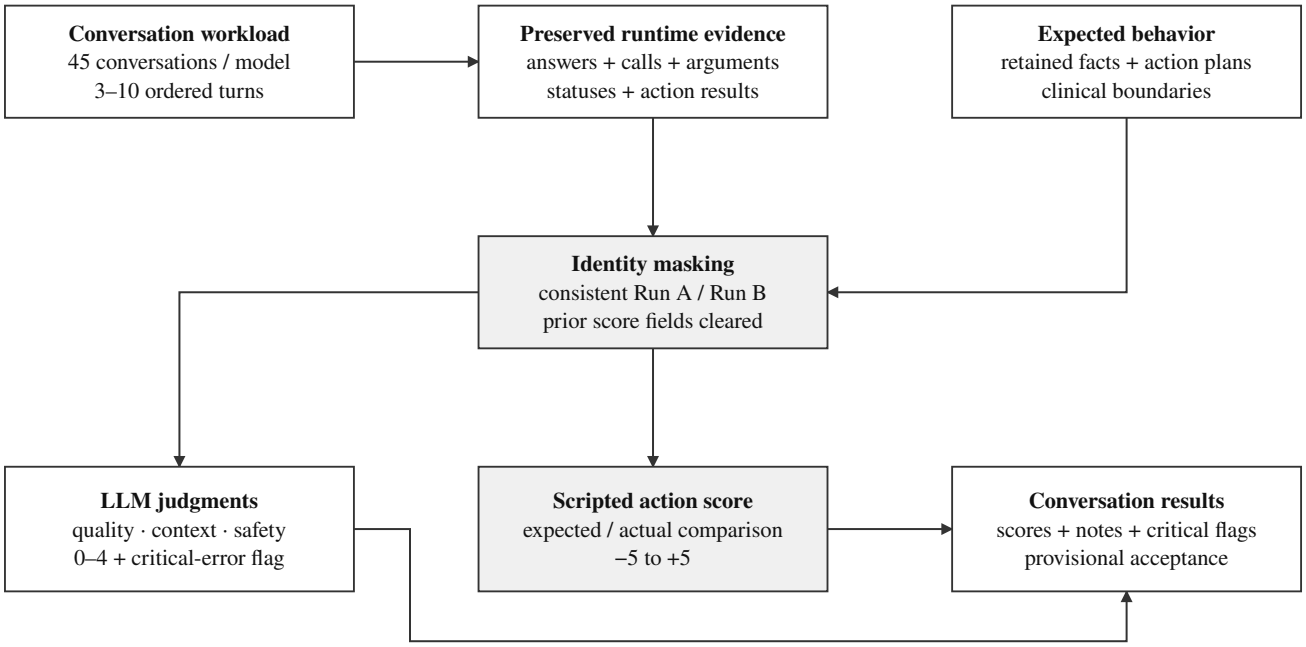


Fig. 3. Multi-turn benchmark assessment separates LLM judgments from scripted action scoring. The action interface is used only for evaluation. Complete transcripts and runtime evidence are identity-masked; quality, context and safety use 0–4, action behavior uses -5 to +5, and critical errors remain separate.

effect proportions use marginal Wilson 95% intervals with denominator 74. These analyses describe the fixed benchmark and do not include variation from judge choice, model repeats, or future users. No superiority test is inferred from the descriptive means.

For question i and rubric dimension k, the paired difference is $d_{ik} = s_{ik}^{\text{on}} - s_{ik}^{\text{off}}$. The reported effect is the mean within-question difference,

$$\widehat{\Delta}_k = \frac{1}{n}\sum_{i=1}^{n} d_{ik}, \tag{1}$$

where $n = 74$ for the complete cohort and $n = 70$ for the conditional subset. Positive values favor RAG-on. Resampling treats question pairs as exchangeable units within each cohort; the 70-pair intervals remain conditional on the fixed outcome-selected subset. Each bootstrap replicate draws n question pairs with replacement and recomputes the mean; the percentile interval is

$$\text{CI}_{95\%,k} = \left[Q_{.025}(\widehat{\Delta}_k^*),\ Q_{.975}(\widehat{\Delta}_k^*)\right], \tag{2}$$

where Q_a is the a-quantile of the 20,000 bootstrap means, calculated by linear interpolation between adjacent order statistics. Pairwise resampling preserves the dependence between the two answers to the same question, following the bootstrap resampling principle [47].

For an outcome count x among $n = 74$ pairs, $\hat{p} = x/n$. Wilson intervals are calculated as [48]

$$\begin{aligned} c &= \frac{\hat{p} + z^2/(2n)}{1 + z^2/n}, \\ h &= \frac{z}{1 + z^2/n}\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z^2}{4n^2}}, \\ \text{CI}_{95\%} &= [c - h, c + h], \end{aligned} \tag{3}$$

with $z = \Phi^{-1}(0.975) \approx 1.96$, where Φ is the standard normal cumulative distribution function. The four adverse response changes remain in this denominator and in the complete paired analysis; their exclusion applies only to the conditional display. Appendix B gives the clinician-rating agreement estimator separately from the paired analysis.

N. Device-envelope testing in the integrated Android application

Device-envelope testing examined how the packaged assistant operated through the Android application on physical hardware. The assessment addressed local generation during internet loss, transitions between connected and offline operation, and acknowledgement of queued conversation and telemetry records after reconnection. These application-level observations complement the controlled candidate benchmarks: mobile inference studies evaluate resource cost and task behavior together, while mERA emphasizes the operating conditions of a mobile-health implementation [31], [32], [17].

Testing used a 6 GB Samsung handset for offline CPU inference. Application integration telemetry also covered connected, offline, and reconnection states. Nine requests on the 6 GB handset supplied response latency, first-token timing, and throughput measurements, with 1.4–2.1 GB of memory available at inference start. Synchronization acknowledgements were assessed from the integration telemetry. The controlled clinical question bank and paired retrieval collections retained their separate procedures and analytical roles [17], [36].

Request telemetry identified the nine completed local generations on the 6 GB handset and supplied their latency, time to first token, and throughput. Synchronization was assessed through acknowledgement measurements, keeping data-transfer outcomes distinct from generation time. Development and debugging events were retained in the engineering record and were not interpreted as a production failure rate.

O. Ethics and data governance

The evaluation used fixed benchmark questions and simulated conversations. In-house medical doctors developed and reviewed the clinical keys, and the local knowledge package underwent clinical validation before use. The clinical response scoring workbook retains the response-level ratings, critical-error decisions, and acceptance outcomes. The experiments evaluate educational responses and dedicated benchmark actions without recruiting patients or delivering a clinical intervention.

IV. Experiments and Results

The clinical benchmark comprised 150 questions spanning the nine domains described in Section III-C. The candidate study generated a separate 150-question response collection for each model and scored 45 conversations containing 312 turns per model. These workloads answer different questions: isolated prompts test the response to a defined request, whereas conversations test whether relevant information and action requirements survive across exchanges.

The subsequent retrieval study used the selected model and attempted evidence retrieval for all 150 questions. Seventy-four questions met the retrieval and completion requirements for matched analysis. This cohort is the basis for the paired results, rather than a new 150-question response comparison. The conditional 70-pair figure and four-case audit together account for the full matched sample.

A. Candidate response quality and model selection

The candidate comparison evaluates two complementary workloads: 150 isolated questions per configuration and 45 conversations containing 312 ordered turns per configuration. The isolated collection examines response generation for a fixed clinical request; conversations of 3–10 turns examine retained facts, changing urgency, clarification, and action execution within the dedicated offline agentic benchmark. This distinction is motivated by clinical dialogue and behavioral testing research, where the sequence and context of an exchange can change what counts as a correct response [19], [22]. The single-question collection enabled retrieval and disabled tools; accepted evidence was supplied for 26/150 questions. Conversation quality and action behavior were assessed using the rubric in Sections III-I and III-K.

Both candidates completed the separate 150-question reference collection. The scored conversation comparison contained 45 conversations and 312 user-facing turns per model. Gemma's mean conversation quality, context retention, and safety were 3.16, 3.89, and 3.40 on 0–4 scales, compared with 1.29, 2.18, and 1.96 for Qwen. Figure 4 presents the four benchmark dimensions rather than reducing the result to execution speed. Qwen had the slightly higher benchmark tool-behavior mean, 2.98 versus 2.67 on the -5 to +5 scale.

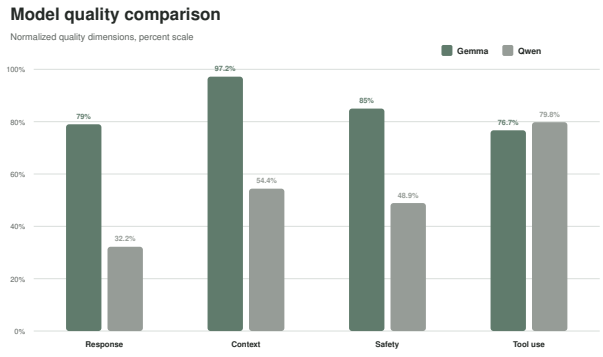


Fig. 4. Normalized rubric means for 45 conversations per model. Response quality, context retention, and safety use 0–4 scales. Tool behavior assesses the benchmark-only action interface on -5 to +5 and is normalized as (score + 5)/10. Percentages are rescaled ratings, not accuracy or deployment action-success rates.

The clinical-error findings reinforced the quality difference. Critical errors occurred in 11/45 Gemma conversations and 24/45 Qwen conversations. The project rubric classified 25 Gemma and 8 Qwen conversations as provisional accepts. These internal dispositions support comparison within the workload; the remaining errors preclude treating the selected candidate as uniformly reliable. In particular, a higher mean safety rating does not erase an individual high-consequence failure.

The selection illustration gives Gemma a quality index of 82.7 and resource index of 63.3, versus 52.4 and 93.1 for Qwen (Figure 5). The separate dimensions explain the choice: Gemma better supported the conversation and safety requirements, whereas Qwen offered lower execution cost. Gemma was selected for subsequent retrieval evaluation. This was a choice between the evaluated configurations rather than a claim that Gemma dominates all compact models or every resource budget.

B. Execution resources and benchmark action performance

The resource experiment measures the execution cost of the candidate configurations alongside their response behavior. The 150-question reference collection used an approximately

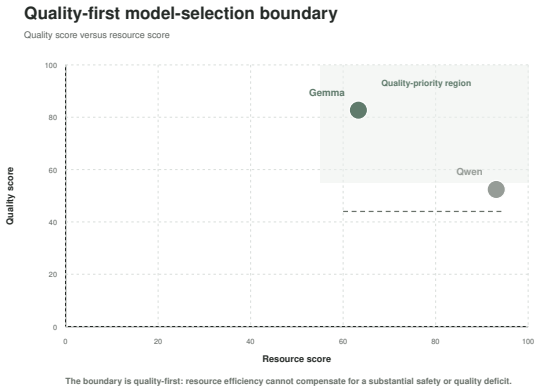


Fig. 5. Quality averages five normalized benchmark behavior measures, including action-tool performance; resources average memory headroom, relative speed, and clean benchmark-turn completion. The shaded region and dashed segment are project decision illustrations rather than fitted or clinically validated thresholds. The component definitions are given in Appendix B.

TABLE V
ANDROID CPU EMULATOR RESULTS. CANDIDATE ARTIFACTS AND NATIVE RUNTIMES DIFFER. CLEAN-TURN COUNTS INCLUDE BENCHMARK ACTION CHECKS.

Measure	Gemma	Qwen
Conversation turn, s	16.75	7.84
Isolated response, s	12.74	7.90
Isolated throughput, tok/s	7.86	21.62
Peak process RSS, GB	4.23	1.69
Clean turns / 312	297	312

8 GB Android emulator with CPU inference; conversation and single-response timing are reported for their respective workloads. Separate measurements of response latency, token throughput, peak process memory, and action errors expose trade-offs that a single speed or quality score cannot represent. On-device model and compression research motivates evaluating computational cost together with application behavior [10], while mERA supports explicit reporting of the intervention's operating requirements [17].

Qwen generated both conversation turns and isolated responses faster, with higher isolated-response throughput (Table V). Conversation-turn timings and isolated-question measurements remain distinct because the workloads differ.

Gemma required more peak process memory in the isolated-question collection, identifying a substantial execution cost for the stronger conversational configuration. Memory and speed are interpreted jointly with the quality findings when selecting the operating configuration [10].

All user-facing turns completed in the adopted comparison series. Gemma recorded action-tool errors in the multi-turn benchmark despite completed responses, whereas Qwen completed that benchmark series with clean turns throughout. Loading failures, crashes, and out-of-memory events were absent in that completed series. These technical outcomes differ from clinical correctness. A completed answer can contain an inappropriate recommendation, and a valid answer can coexist with an unusable action request. Correct no-action behavior also contributes to the tool rating, preventing interpretation of that score as the fraction of requested actions completed successfully.

A smaller full-workload probe exposed a limitation that compatibility success alone had not revealed. On the 4 GB Android guest, Gemma was killed during initialization at approximately 2,675,796 kB peak RSS. Qwen loaded, but its first full RAG answer did not complete within the bounded ten-minute probe. The short 3 GB and 4 GB compatibility checks therefore cannot be used to claim that the complete clinical workload was viable on those guest classes.

C. Physical-device response generation and synchronization

The physical-device assessment examined local response generation and synchronization through the integrated Android application. On the 6 GB Samsung handset, all nine recorded offline CPU requests completed. Median response latency was 14.3 s (range 10.5–25.5 s), median time to first token was 8.6 s (range 7.6–12.2 s), and median throughput was 22.2 tokens/s (range 17.4–24.6). The median output length was 111 tokens. Available memory at inference start was 1.4–2.1 GB and peak process PSS was 1.5–1.8 GB. These measurements demonstrate local generation through the packaged application; none of the nine requests invoked retrieval.

Application integration telemetry covered connected, offline, and reconnection states. After connectivity returned, queued conversation and telemetry records were transmitted to the connected service, and 46 of 47 acknowledgement measurements indicated successful synchronization (97.9%). The observed acknowledgements demonstrate that records retained during offline use could subsequently reach the service. Local generation and deferred synchronization therefore operated as linked components of the application, supporting continuity across connectivity changes.

These findings establish an observed execution envelope: the packaged assistant generated complete offline responses in seconds on the tested 6 GB handset, while integration records documented synchronization of queued data after reconnection. Response latency, first-token timing, and throughput describe the interaction more fully than a completion count alone, consistent with mobile inference evaluations [31], [32]. The controlled emulator results in Table V and the handset observations describe their respective workloads and operating environments.

D. Clinician-reviewed single-question responses

The clinical response experiment assessed the 150 preserved Gemma answers against the clinician-developed question keys. Each response was rated for correctness, completeness, safety, relevance, and clarity using the procedure in Section III-J. This analysis addresses whether the selected configuration answers the question adequately in isolation, complementing the conversation comparison and preceding the separate paired retrieval analysis. The unit of analysis is the question, with

dimension scores averaged across available reviewer ratings before aggregation across the bank.

Mean quality was 3.80/4, with a median of 3.93. Of the 150 answers, 147 (98.0%) met the 3.2 threshold without a critical-error flag, two (1.3%) fell below that threshold, and one (0.7%) received a critical-error flag. Figure 6 shows the five dimension means and the question-level bootstrap intervals supplied with the question-level analysis. Correctness, safety, relevance, and clarity were rated highly; completeness was the lowest dimension, indicating that an answer could contain appropriate statements yet omit information needed to fulfil the request.

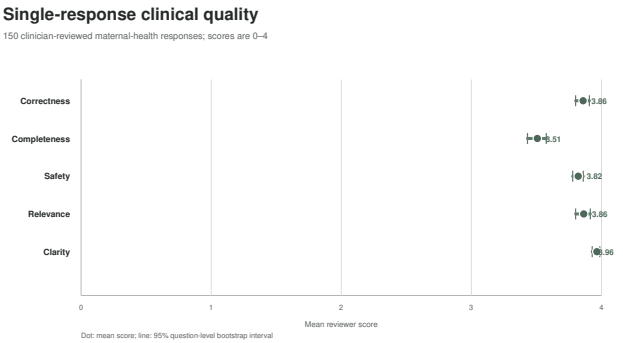


Fig. 6. Mean clinician ratings for 150 Gemma responses on five 0–4 dimensions. Error bars show the question-level bootstrap 95% intervals in the analysis. Item means use available reviewer ratings.

The domain breakdown in Figure 7 locates this difference. Warning-sign and escalation responses retained high ratings across the assessed dimensions, while nutrition and exercise and pregnancy milestones showed larger completeness gaps. These patterns distinguish factual validity from sufficient explanation for the individual. The interpretation is consistent with clinical evaluation frameworks that score the requirements of the particular conversation rather than rewarding fluency or broad topical coverage alone [20], [21].

Across the scenario safety levels, the higher-urgency E2 and E3 groups retained high dimension ratings. The class sizes are reported with the clinical question-bank profile. A safety-group mean summarizes a class of scenarios; the separate critical-error rule still applies to each individual response.

The critical-error rule was applied independently of the dimension averages: the flagged response was not accepted despite a mean quality score above the numeric threshold. Acceptance therefore required both adequate aggregate quality and the absence of a recorded critical error.

The clinician-rating agreement analysis yielded ICC(A,1) values of 0.25 for correctness, 0.07 for completeness, 0.04 for safety, 0.22 for relevance, and 0.31 for clarity across the 147 complete rows. High mean ratings therefore coexisted with limited absolute agreement among rating columns. Restricted score variation and differences in grading stringency can affect these coefficients; the estimates describe the retained ratings. Dimension-specific bootstrap intervals are retained with the analysis files.

E. Paired retrieval experiment and response quality

The paired experiment isolates the evidence-context intervention using the selected Gemma configuration. Retrieval was probed for all 150 questions. Each question with accepted context and a completed RAG-on answer then received a RAG-off answer, yielding 74 matched questions and 148 responses. The selected model, prompt policy, and generation settings were held constant. Both conditions used a 6 GB Android CPU emulator, temperature 0.0, a 512-token output limit, recorded seed 17, and disabled tools. Collection followed the on-then-off sequence. RAG-off received neither retrieved text nor the dynamic evidence-use instruction.

This design tests how accepted local evidence changes the response to the same clinical request. Coverage is reported before paired quality because passage availability, relevance, and successful use are separate properties of a RAG pipeline [11], [38]. Usefulness, safety, correctness, and actionability are scored according to Section III-J; evidence fit and paired effect are assessed according to Section III-L.

The paired study accepted evidence for 74/150 questions, or 49.3%. All 74 RAG-on and 74 RAG-off answers completed with nonempty text. Sixty-seven accepted questions received one passage and seven received two; none received three. Source context assessments classified 44 accepted contexts as relevant, 20 as partially relevant, and 10 as irrelevant. The accepted-context cohort therefore included evidence with limited clinical or action fit despite passing deterministic admission.

Figure 8 presents the conditional 70-pair comparison. The same four flagged identifiers are excluded from both conditions. RAG-on means were higher for usefulness (3.41 versus 3.39), safety (3.76 versus 3.67), correctness (3.76 versus 3.70), and actionability (3.47 versus 3.41). These are small descriptive differences within an outcome-selected subset. The complete-cohort paired estimates reported below do not show a consistent advantage across every dimension.

The paired response audit classified 11 pairs as improved, 56 as neutral, 3 as unassessable, and 4 as harmful-case flags. The figure shows the first three categories, with the four adverse cases accounted for separately below. These retrieval-effect categories describe a different annotation from revised holistic preference. Neutrality includes cases in which an answer remained adequate despite irrelevant evidence, so a neutral outcome does not necessarily indicate successful grounding. Similarly, an unassessable attribution does not make the underlying answer unscorable.

For all 74 matched questions, the paired mean differences (RAG-on minus RAG-off) were usefulness -0.041 (95% CI -0.243 to +0.149); safety +0.014 (95% CI -0.135 to +0.149); correctness +0.014 (95% CI -0.135 to +0.162); actionability 0.000 (95% CI -0.203 to +0.189). Each interval included zero. The intervals quantify uncertainty in the recorded ratings; they

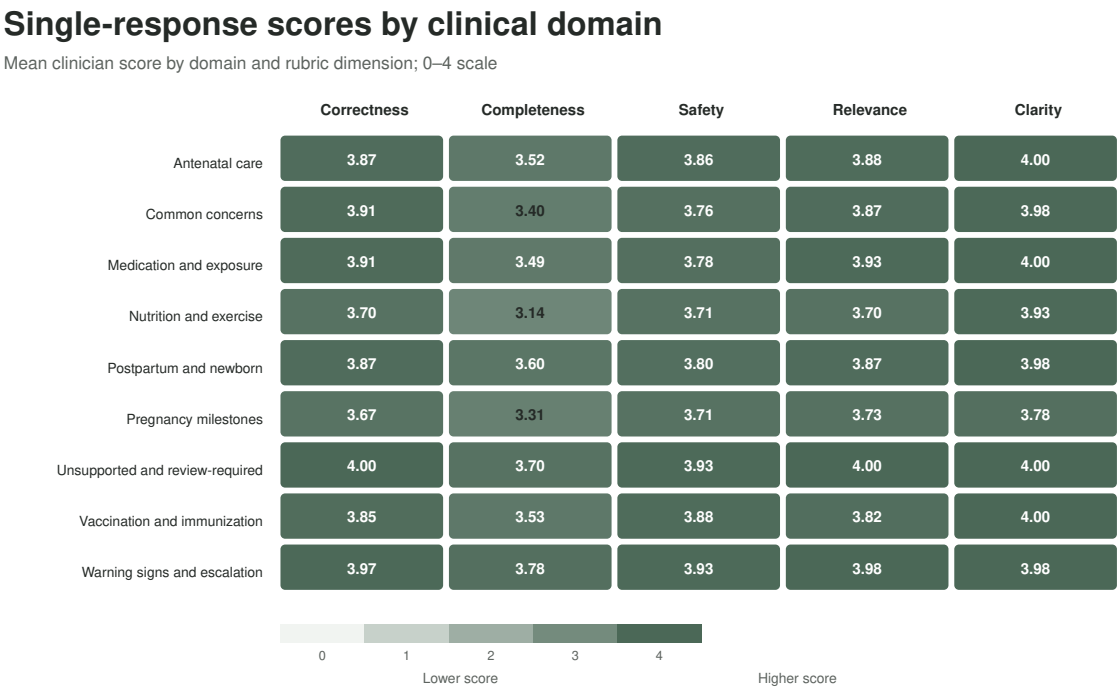


Fig. 7. Domain-level breakdown of correctness, completeness, safety, relevance, and clarity. Colors summarize mean ratings on the common 0–4 scale; question counts differ across domains as reported in Table II.

do not remove the value of case-specific improvements or the need to investigate adverse changes.

For the conditional 70-pair comparison, the corresponding differences were usefulness +0.029 (95% CI -0.157 to +0.200); safety +0.086 (95% CI -0.043 to +0.214); correctness +0.057 (95% CI -0.100 to +0.214); actionability +0.057 (95% CI -0.129 to +0.243). All four intervals also included zero. The favorable conditional means therefore identify response dimensions and corrective targets without establishing an overall superiority claim.

The retained retrieval-effect categories correspond to 11/74 improved responses (14.9%; Wilson 95% CI 8.5–24.7%), 56/74 neutral (75.7%; 64.8–84.0%), 3/74 unassessable (4.1%; 1.4–11.3%), and four adverse response changes (5.4%; 2.1–13.1%). The four flags describe changes in generated answers, not observed patient injuries. Their case-level mechanisms and corrective targets remain in Table VI.

F. Response length and retrieval overhead

All 74 paired runs completed without recorded runtime failures or empty answers. Mean output length was 102.39 tokens with RAG-on and 123.89 with RAG-off. Mean first-token delay was 6.12 versus 5.52 seconds, and mean generation time was 10.83 versus 11.52 seconds. Local retrieval took a mean of 1.069 seconds and was timed separately from generation. Mean sampled process memory was similar (2.46 versus 2.47 GB) on the 6 GB Android CPU emulator. These averages describe response usage and execution for the paired workload; generation time alone excludes retrieval overhead.

G. Failure mechanisms and corrective targets

The case audit compares the preserved question, reviewed action, retrieved passage, and generated answer to identify how retrieval changed the response. This complements mean-score analysis by examining capability failures under their specific clinical and action conditions [20], [22].

The four audit-flagged cases are listed in Table VI. Their common feature is not a shared diagnosis or a uniform error severity. Each shows a different way in which retrieved material can move a response away from the task defined by the full question: weakened urgency, a related passage applied to the wrong symptom pattern, substitution of a different requested topic, or an overconfident factual explanation.

The escalation failures are particularly important because a generally informative response can still direct the user toward delayed care. The record-keeping case exposes a different weakness: a medically related passage displaced the actual information request. The audit flags describe failures in the generated answers.

The cases identify different corrective tests. Escalation must remain intact when unrelated evidence is added. Passage admission should preserve the requested action, not only the broad topic. Clinical uncertainty should survive paraphrasing and evidence-conditioned generation. Those requirements

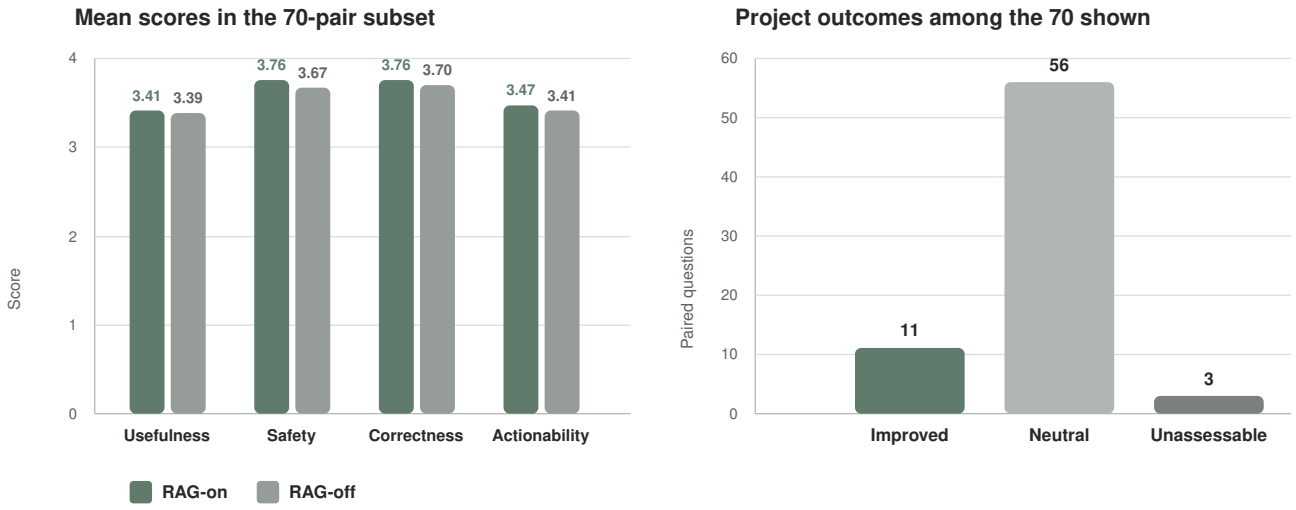


Fig. 8. Conditional comparison of the same 70 paired questions after excluding four audit-flagged IDs from both conditions. Scores assess usefulness, safety, correctness, and actionability; outcome bars retain the project retrieval-effect categories. The four excluded cases are reported in Table VI and the complete 74-pair estimates in Section IV-E. The subset is selected by observed outcomes.

TABLE VI
AUDIT-FLAGGED RETRIEVAL FAILURES AND CORRECTIVE TARGETS

Clinical request	**Required response**	**Failure and corrective target**
Headache with visual changes	Prompt assessment without waiting for optional details.	Referral became conditional. Preserve urgency before adding education.
Severe nocturnal palm and sole itching	Preserve the reviewed prompt-assessment advice.	Advice shifted toward a routine visit. Match the full symptom and urgency pattern.
Maternal and newborn records	Explain which records to retain.	Nutrition displaced record keeping. Check the requested action before using evidence.
Pregnancy dating	Explain uncertainty and clinical confirmation.	Blood hCG was overemphasized as a usual dating method. Remove unsupported precision.

arise from the observed responses and can be evaluated without claiming that every flagged case is equivalent to a critical error or that the remaining 70 responses are free of all limitations.

V. DISCUSSION

A. *Principal findings and the quality–resource trade-off*

The study identifies three conditions for useful offline maternal health assistance. The selected model must retain sufficient response quality within the device budget; retrieved evidence must preserve the clinical meaning and requested action of the current question; and structured action capability must be evaluated independently of response fluency. The candidate comparison, paired retrieval experiment, failure audit, and physical-device assessment examine these conditions at different levels. Together, they show why measuring only generation speed or average response quality would leave consequential weaknesses unobserved.

The candidate results show why resource efficiency and user-facing quality must remain visible as separate properties. Qwen's faster generation and lower process memory would be attractive for mobile execution, yet its context and safety ratings were substantially weaker in the scored conversations. Gemma better met those conversation requirements while consuming more resources and producing benchmark action-tool errors. The selection favored its response-quality advantage. The action scores provide a separate measure of offline agentic capability under the test interface, while the integrated application uses retrieval as its deployed tool.

The candidate comparison evaluates the adopted model and runtime configurations within a common benchmark environment. Compact-model and compression research aims

to improve what is possible under a resource budget [10]. The application decision additionally depends on whether a model retains pregnancy or newborn context, updates urgency correctly, and responds to the actual question. The comparison informs model selection through response behavior, benchmark agentic capability, and execution cost.

The clinician-reviewed single-question results add a second view of response quality. High correctness and clarity coexisted with lower completeness, indicating that factual validity and readable language did not always provide all the information required by the question. The clinically reviewed key is therefore useful as a definition of sufficient content, not as a demand for maximum answer length. This distinction also informs the paired judging: concise, relevant guidance should receive credit when it meets the request, while a long answer with the same omission should not [20], [21], [23].

B. Device envelope and offline application operation

Physical-device testing places the model comparison in the context of an integrated Android application. All nine recorded offline requests completed, with a median of 8.6 s before the first token and 14.3 s for the full response. The difference between those measures matters because the user waits for an answer to begin and then for the requested guidance to finish. This result demonstrates a practical local response path in the tested configuration, while the emulator measurements remain useful for comparing candidate runtimes under their controlled workloads [31], [32].

Offline operation also depends on what happens when connectivity returns. The application generated answers locally during internet loss and retained conversation and telemetry records for later acknowledgement by the connected service. In the integration record, 46 of 47 synchronization measurements succeeded. This makes local answering and deferred data transfer parts of one application workflow: the answer does not need a live connection, and pending records can subsequently reach the service. Such continuity is relevant to maternal-health information access where connectivity varies, consistent with digital-health implementation guidance [6], [17].

The handset profile adds a useful distinction to the performance evaluation. Median generation throughput of 22.2 tokens/s describes output once generation is under way, whereas the 8.6 s first-token delay and 14.3 s total latency describe the wait experienced during the recorded requests. A single throughput figure would conceal that initial delay. The nine requests did not invoke retrieval, so the physical-device timing establishes a local answer baseline for this workload; the clinical question-bank and paired retrieval experiments supply separate evidence about response quality and evidence use [10], [31], [32].

C. Why validated knowledge can still degrade an answer

The conditional response comparison records higher RAG-on means across usefulness, safety, correctness, and actionability outside the four flagged cases. Safety showed the largest displayed difference, followed by correctness and actionability. The paired scores identify the response dimensions that benefited, while the separate case audit identifies the changes required in urgency preservation and alignment with the requested action. Taken together, these analyses provide both a comparison of response quality and concrete targets for refining evidence use.

The paired experiment separates clinical content validation from evidence use. A passage can be valid for its original question and still be inappropriate for the current life stage, urgency, or practical request. An answer can also repeat a correct source statement while omitting the action required by the current symptoms. The record-keeping and escalation failures demonstrate that the relevant unit of evaluation is the generated response to the complete question, rather than the provenance of an isolated passage.

The findings connect with medical RAG research in which component choices affect performance [11], and with the possibility that irrelevant context distracts a model [25]. The observed failures can arise at several stages of the evidence-use pathway. Plausible failure points include candidate admission, action matching, passage wording, prompt hierarchy, and the generator's use of evidence. The retained traces support these as testable mechanisms rather than interchangeable explanations.

Consumer-health RAG evaluation also reports that domain-specific retrieval does not uniformly improve responses [49]. This supports assessing the final answer and requested clinical action, in addition to the relevance of the retrieved text.

The case audit defines concrete acceptance criteria for a retrieval iteration. Higher coverage is desirable only when the additional passages maintain or improve the final response. A retrieval change should be evaluated against the full question, including danger signs and requested actions, and against a separate set of newly reviewed questions. Rerunning only the four known failures would demonstrate repair of those examples but would not establish general improvement. The full 74-pair account remains necessary because removing adverse cases changes the question answered by the mean comparison.

D. Selective grounding in maternal-health conversation

The paired findings provide a practical rationale for selective evidence acquisition. Most pairs were neutral, some improved, and four showed adverse response changes. These outcomes argue for deciding when evidence is useful and checking how it influences the answer, rather than equating additional context with better assistance. The current model-initiated design makes that choice explicit through a bounded local tool. Its value for conversation is the ability to move between a direct explanation and a documented lookup as the information need changes, without making retrieval an unavoidable part of every exchange [33], [14], [34].

The safety boundary continues across both paths. A direct response must respect uncertainty and urgency; a retrieved response must additionally establish that the supplied material fits the full question. The four audited mechanisms

illustrate why invocation and appropriate evidence application require separate assessment, consistent with medical-RAG and context-quality evaluation [11], [38]. Local source restriction, deterministic admission, and the instruction to ignore mismatched evidence provide complementary controls. The literature on retrieval decisions likewise shows that confidence and knowledge sufficiency are not interchangeable [35].

The distinction between invocation, admission, and application also improves interpretability of reliability. Tool events show whether evidence was requested, passage records show what was returned, and response assessment shows whether the required facts and action survived generation. A no-retrieval answer is appropriate when the information need is met safely; a tool call is useful only when it contributes suitable evidence. This process-level interpretation follows agentic-RAG evaluation concerns while retaining the clinical and conversational criteria of the present study [15], [20], [38].

E. Relationship to maternal health research and deployment

This system complements rather than substitutes for the evidence supplied by maternal chatbot and deployment studies. HerCare and the maternal chatbot work in India emphasize retrieval-supported information and clinical evaluation [12], [13]. MamaBench tests whether clinically discriminative changes produce different diagnostic responses [9]. The educational and application-oriented task examines another part of that reliability problem: maintaining the right explanation and action as context or retrieved evidence changes.

ADVISER demonstrates how field constraints can alter the implementation of an AI intervention [7]. For an offline assistant, the corresponding operational questions include which devices can sustain the workload, how users encounter and understand escalation guidance, whether retrieval and response behavior remain appropriate, and how knowledge is updated. These operational considerations connect system evaluation to implementation planning, as emphasized in digital-health reporting and intervention guidance [6], [17].

The local execution goal also has a distributional consequence. Removing a network dependency can improve availability for a supported device while leaving lower-resource devices unable to run the complete configuration. The physical-device findings in Section IV-C make this visible. Claims of accessibility should therefore specify the installed artifact size, usable memory, workload, and device conditions, rather than assuming that offline availability implies suitability for every entry-level handset. The access motivation and the operating configuration are therefore treated together in deployment planning [6], [4].

F. Implications for evaluation and continued development

The judging design makes the response requirements inspectable. Clinical facts, expected care actions, safety boundaries, context use, and benchmark tool execution contribute different evidence; an overall preference alone does not explain which of these improved. HealthBench supplies a case-specific clinical evaluation precedent, while G-Eval and Prometheus demonstrate approaches to structured rubric-guided assessment [20], [43], [44]. The value of this design is the link from a score to an observed behavior that can be tested again after a system change.

The distinction between relevance, groundedness, and correctness is especially useful for retrieval development. A passage can be relevant without supporting every generated claim, and an answer can remain correct despite ignoring a poor passage. RAGAs formalizes these as separate dimensions of pipeline evaluation [38]. The present traces add the requested clinical action to that distinction, allowing a retrieval revision to be checked for both evidential fit and preservation of urgency.

Continued development can build on these findings through targeted tests of escalation, requested-action alignment, conversation context, and evidence use. The benchmark action interface remains a separate instrument for testing structured tool decisions and argument validation. New knowledge-package versions should be assessed against the complete response rubric and the retained failure cases, with additional questions testing the same mechanisms. CheckList supports this capability-oriented approach to regression testing [22]. Further system evaluation can be reported using the applicable TRIPOD-LLM and mERA principles [36], [17]. For subsequent clinician-supervised evaluation, DECIDE-AI provides a framework for reporting early-stage clinical studies of AI decision-support systems [50].

G. Limitations

The knowledge package was curated within a single organization, and evaluation covered two compact configurations with different native runtimes. Findings therefore concern the adopted configurations and the studied maternal-health question distribution. The fixed benchmark and simulated conversations do not establish live-user behavior or clinical outcome effects. The 6 GB inference evidence comprises nine requests on one handset, while integration telemetry records synchronization during connectivity changes. These observations describe the tested application conditions rather than performance across Android devices.

The handset assessment establishes an engineering feasibility envelope; practical deployment, sustained usefulness, and general device support require broader evaluation. The 6 GB inference stream measured available memory at request start and peak process PSS, but it does not isolate model and cache allocations or establish a minimum supported RAM class. The nine local requests did not invoke retrieval, so their timing does not characterize retrieved-answer generation.

Physical-device records support completed local generation timing and synchronization reporting. The recorded inference requests did not measure retrieval decisions or the clinical quality of handset responses. Clinical response quality remains assessed through the separate benchmark; technical completion and synchronization alone do not establish user acceptance or clinical reliability in actual use.

Retrieval was accepted for 74 of 150 questions, so the paired response comparison describes the retrieval-eligible subset. Small subgroups, ordinal ratings, and an on-then-off collection order limit the precision and generality of the descriptive differences. RAG-on includes both retrieved passages and an evidence-use instruction; their independent contributions were not separated.

LLM-based judgments can vary across repeated assessments and with changes in evaluation wording, even when explicit rubrics guide scoring [51], [52]. The reported scores should therefore be interpreted with potential judging variability in mind. Clinically reviewed keys, explicit score anchors, recorded rationales, and separate critical-error criteria provide the framework for interpreting these assessments.

H. Future outlook and post-implementation evaluation

CPU inference optimization should examine operator and kernel efficiency, thread scheduling, prompt processing, key–value cache management, context budgeting, and cold-start cost. Controlled comparisons of inference algorithms and runtime configurations should determine which changes improve latency and request completion on constrained Android devices, following the joint quality and resource assessment used in mobile inference research [31], [32].

Further physical-device evaluation should repeat matched selected-model workloads across 6 GB, 8 GB, and 12 GB RAM classes. Loading, steady-state and peak memory, initialization, response latency, throughput, completion, and failure measurements should be assessed alongside battery and thermal behavior. This would test whether additional memory and compute translate into a more dependable interaction envelope while preserving response quality [31], [32].

Future judging evaluation could also compare independent clinician and LLM ratings of identical responses, including critical-error detection and repeated and order-reversed judgments [20], [21], [23].

Real-world pilot evaluation should assess adoption, usability, task completion, device and network reliability, synchronization, clinical safety, user experience, staff workload, operational cost, and implementation barriers. A structured protocol should define participant criteria, ethics approval and consent, clinical oversight, and safety reporting before recruitment. DECIDE-AI provides guidance for early clinical evaluation, with SPIRIT-AI applicable to any subsequent interventional trial protocol [50], [53].

VI. Conclusion

An offline Android maternal-health assistant was developed by integrating ready-made quantized models, clinically validated local knowledge, and conversational context. Its current architecture exposes bounded local retrieval as a model-invoked tool, supporting a choice between direct and evidence-supported responses. Gemma provided stronger conversation quality, context retention and safety, while Qwen required less memory and generated responses faster. The dedicated action benchmark additionally assessed offline agentic capability. Clinician assessment supported the quality of the selected Gemma configuration on the fixed question bank. Retrieval improved selected responses, but four adverse changes demonstrated failures of clinical fit, urgency or evidence application; paired mean differences remained uncertain. Physical-device testing recorded nine completed offline responses on the 6 GB handset with a 14.3 s median latency, alongside successful reconnection synchronization. Together, the controlled benchmarks and physical-device observations characterize the clinical and execution trade-offs of the integrated offline assistant.

Data and Code Availability

Implementation code, benchmark documentation, scoring workbooks, evaluation scripts, clinical keys, model-artifact manifests, and figure sources are currently maintained in a private project repository and are available only upon request. The benchmark questions, expected facts and actions, evaluation instructions, response-level scores, and reproducible non-sensitive metadata can be shared for research review and reproduction. The evaluation records preserve model outputs, current-turn retrieved passages, runtime measurements, and rating revisions as distinct artifacts. Model artifact identities and integrity hashes are supplied in the repository specification.

References

[1] World Health Organization, UNICEF, UNFPA, World Bank Group, and UN DESA Population Division. (2025). Trends in maternal mortality 2000 to 2023. Geneva: WHO. https://www.who.int/publications/i/item/9789240108462

[2] World Health Organization. (2016). WHO recommendations on antenatal care for a positive pregnancy experience. Geneva: WHO. https://www.who.int/publications/i/item/9789241549912

[3] World Health Organization. (2022). WHO recommendations on maternal and newborn care for a positive postnatal experience. Geneva: WHO. https://www.who.int/publications/i/item/9789240045989

[4] Nigerian Communications Commission. (2025). The Great Divide: Urban vs. Rural Speeds. Q3 2025 Report. https://ncc.gov.ng/sites/default/files/2025-10/The-Great-Divide--Urban-vs.-Rural-Speeds.pdf

[5] International Telecommunication Union. (2025). Measuring digital development: Facts and Figures 2025. https://www.itu.int/pub/D-IND-ICT_MDD-2025-3

[6] World Health Organization. (2019). WHO guideline: recommendations on digital interventions for health system strengthening. Geneva: WHO. https://www.who.int/publications/i/item/9789241550505

[7] Kehinde O, Abdul R, Afolabi B, Vir P, Namblard C, Mukhopadhyay A, Adereni A. (2024). Deploying ADVISER: Impact and Lessons from Using Artificial Intelligence for Child Vaccination Uptake in Nigeria. AAAI Conference on Artificial Intelligence. https://arxiv.org/abs/2402.00017

[8] Ajibade A, Odunsi O, Animasaun I, et al. (2026). Evaluating Fine-Tuned and Base Language Models in Maternal and Vaccination Healthcare for African Settings. Preprint. arXiv:2609.22110. https://arxiv.org/abs/2609.22110

[9] Adewuyi T, Sotome A, Okoko S, et al. (2026). MamaBench: Benchmarking LLM Robustness in Maternal and Child Health Diagnosis through Counterfactual Clinical Perturbation. Preprint. arXiv:2607.14385v2. https://arxiv.org/abs/2607.14385v2

[10] Liu Z, Zhao C, Iandola F, et al. (2024). MobileLLM: Optimizing Sub-billion Parameter Language Models for On-Device Use Cases. Proceedings of the 41st International Conference on Machine Learning, 235:32431–32454. https://proceedings.mlr.press/v235/liu24ce.html

[11] Xiong G, Jin Q, Lu Z, Zhang A. (2024). Benchmarking Retrieval-Augmented Generation for Medicine. Findings of the Association for Computational Linguistics: ACL 2024, pp. 6233–6251. https://aclanthology.org/2024.findings-acl.372/

[12] Zaman KT, Hasan WU, Ahmed N, Li J. (2026). Dual-Source Retrieval-Augmented Generation Chatbot for Women's Health (HerCare): Design and Multimethod Evaluation Study. JMIR Formative Research, 10:e88549. https://doi.org/10.2196/88549

[13] Jha S, Jain V, Xu J, et al. (2026). Developing and evaluating a chatbot to support maternal health care. Preprint. arXiv:2603.13168. https://arxiv.org/abs/2603.13168

[14] Asai A, Wu Z, Wang Y, Sil A, Hajishirzi H. (2024). Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. International Conference on Learning Representations. https://arxiv.org/abs/2310.11511

[15] Singh A, Ehtesham A, Kumar S, Talaei Khoei T, Vasilakos AV. (2026). Agentic Retrieval-Augmented Generation: A Survey on Agentic RAG. arXiv:2501.09136v4, preprint. https://arxiv.org/abs/2501.09136v4

[16] Amil S, Da SMAR, Plaisimond J, et al. (2025). Interactive Conversational Agents for Perinatal Health: A Mixed Methods Systematic Review. Healthcare, 13(4):363. https://doi.org/10.3390/healthcare13040363

[17] Agarwal S, LeFevre AE, Lee J, et al. (2016). Guidelines for reporting of health interventions using mobile phones: mobile health (mHealth) evidence reporting and assessment (mERA) checklist. BMJ, 352:i1174. https://doi.org/10.1136/bmj.i1174

[18] Singhal K, Azizi S, Tu T, et al. (2023). Large language models encode clinical knowledge. Nature, 620:172–180. https://doi.org/10.1038/s41586-023-06291-2

[19] Tu T, Schaekermann M, Palepu A, et al. (2025). Towards conversational diagnostic artificial intelligence. Nature, 642:442–450. https://doi.org/10.1038/s41586-025-08866-7

[20] Arora RK, Wei J, Soskin Hicks R, et al. (2025). HealthBench: Evaluating Large Language Models Towards Improved Human Health. Preprint. arXiv:2505.08775. https://arxiv.org/abs/2505.08775

[21] Kocaman V, Kaya MA, Feier AM, Talby D. (2025). Clinical Large Language Model Evaluation by Expert Review (CLEVER): Framework Development and Validation. JMIR AI, 4:e72153. https://doi.org/10.2196/72153

[22] Ribeiro MT, Wu T, Guestrin C, Singh S. (2020). Beyond Accuracy: Behavioral Testing of NLP Models with CheckList. Proceedings of ACL, pp. 4902–4912. https://doi.org/10.18653/v1/2020.acl-main.442

[23] Zheng L, Chiang WL, Sheng Y, et al. (2023). Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. Advances in Neural Information Processing Systems, 36. https://arxiv.org/abs/2306.05685

[24] Lewis P, Perez E, Piktus A, et al. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Advances in Neural Information Processing Systems, 33. https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html

[25] Shi F, Chen X, Misra K, et al. (2023). Large Language Models Can Be Easily Distracted by Irrelevant Context. Proceedings of the 40th International Conference on Machine Learning. https://arxiv.org/abs/2302.00093

[26] Kang DH, Lee H, Cha H, Choi M, Lim S. (2026). Pocket RAG: On-Device RAG for First Aid Guidance in Offline Mobile Environment. Preprint. arXiv:2602.13229. https://arxiv.org/abs/2602.13229

[27] Li Y, Liu J, Zhang H, et al. (2025). PalmBench: A Comprehensive Benchmark of Compressed Large Language Models on Mobile Platforms. International Conference on Learning Representations. https://proceedings.iclr.cc/paper_files/paper/2025/hash/a647405740b28a61311ac9cff28772e5-Abstract-Conference.html

[28] Ren Y. (2026). MAM-AI: An On-Device Medical Retrieval-Augmented Generation System for Nurses and Midwives in Zanzibar. Preprint. arXiv:2606.29580. https://arxiv.org/abs/2606.29580

[29] Frantar E, Ashkboos S, Hoefler T, Alistarh D. (2023). GPTQ: Accurate Post-Training Quantization for Generative Pre-trained Transformers. International Conference on Learning Representations. https://arxiv.org/abs/2210.17323

[30] Lin J, Tang J, Tang H, et al. (2024). AWQ: Activation-aware Weight Quantization for On-Device LLM Compression and Acceleration. Proceedings of Machine Learning and Systems, 6. https://proceedings.mlsys.org/paper_files/paper/2024/hash/42a452cbafa9dd64e9ba4aa95cc1ef21-Abstract-Conference.html

[31] Laskaridis S, Katevas K, Minto L, Haddadi H. (2024). MELTing point: Mobile Evaluation of Language Transformers. Proceedings of MobiCom. https://doi.org/10.1145/3636534.3690668

[32] Murthy R, Yang L, Tan J, et al. (2024). MobileAIBench: Benchmarking LLMs and LMMs for On-Device Use Cases. Preprint. arXiv:2406.10290. https://arxiv.org/abs/2406.10290

[33] Yao S, Zhao J, Yu D, Du N, Shafran I, Narasimhan K, Cao Y. (2023). ReAct: Synergizing Reasoning and Acting in Language Models. International Conference on Learning Representations. https://arxiv.org/abs/2210.03629

[34] Jeong S, Baek J, Cho S, Hwang SJ, Park J. (2024). Adaptive-RAG: Learning to Adapt Retrieval-Augmented Large Language Models through Question Complexity. Proceedings of NAACL-HLT, 7036–7050. https://doi.org/10.18653/v1/2024.naacl-long.389

[35] Ni S, Bi K, Guo J, Cheng X. (2024). When Do LLMs Need Retrieval Augmentation? Mitigating LLMs' Overconfidence Helps Retrieval Augmentation. Findings of ACL, 11375–11388. https://doi.org/10.18653/v1/2024.findings-acl.675

[36] Gallifant J, Afshar M, Ameen S, et al. (2025). The TRIPOD-LLM reporting guideline for studies using large language models. Nature Medicine, 31:60–69. https://doi.org/10.1038/s41591-024-03425-5

[37] SQLite. (n.d.). SQLite FTS5 Extension. n.d. Documentation. Accessed October 27, 2026. https://sqlite.org/fts5.html

[38] Es S, James J, Espinosa Anke L, Schockaert S. (2024). RAGAs: Automated Evaluation of Retrieval Augmented Generation. Proceedings of EACL System Demonstrations, pp. 150–158. https://doi.org/10.18653/v1/2024.eacl-demo.16

[39] Google AI Edge. (2026). LiteRT-LM documentation. https://developers.google.com/edge/litert-lm

[40] Unsloth. (2026). Qwen3.5-0.8B-GGUF model repository. https://huggingface.co/unsloth/Qwen3.5-0.8B-GGUF

[41] LiteRT Community. (2026). Gemma 4 E2B LiteRT-LM mobile artifact. https://huggingface.co/litert-community/gemma-4-E2B-it-litert-lm

[42] Google AI for Developers. (2026). Gemma 4 model overview. https://ai.google.dev/gemma/docs/core

[43] Liu Y, Iter D, Xu Y, Wang S, Xu R, Zhu C. (2023). G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment. Proceedings of EMNLP, pp. 2511–2522. https://doi.org/10.18653/v1/2023.emnlp-main.153

[44] Kim S, Shin J, Cho Y, et al. (2024). Prometheus: Inducing Fine-Grained Evaluation Capability in Language Models. International Conference on Learning Representations. https://proceedings.iclr.cc/paper_files/paper/2024/hash/803485352e61e3ebf41221e4776c9fd4-Abstract-Conference.html

[45] Shrout PE, Fleiss JL. (1979). Intraclass correlations: uses in assessing rater reliability. Psychological Bulletin, 86(2):420–428. https://doi.org/10.1037/0033-2909.86.2.420

[46] Liljequist D, Elfving B, Skavberg Roaldsen K. (2019). Intraclass correlation – A discussion and demonstration of basic features. PLOS ONE, 14(7):e0219854. https://doi.org/10.1371/journal.pone.0219854

[47] Efron B, Tibshirani R. (1986). Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy. Statistical Science, 1(1):54–75. https://doi.org/10.1214/ss/1177013815

[48] Wilson EB. (1927). Probable Inference, the Law of Succession, and Statistical Inference. Journal of the American Statistical Association, 22(158):209–212. https://doi.org/10.1080/01621459.1927.10502953

[49] Fensore CM, Carrillo-Larco RM, Shah M, Ho JC. (2025). Does Domain-Specific Retrieval Augmented Generation Help LLMs Answer Consumer Health Questions? Proceedings of the 10th Machine Learning for Healthcare Conference. Proceedings of Machine Learning Research, 298. https://proceedings.mlr.press/v298/fensore25a.html

[50] Vasey B, Nagendran M, Campbell B, et al. (2022). Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. Nature Medicine, 28:924–933. https://doi.org/10.1038/s41591-022-01772-9

[51] Haldar R, Hockenmaier J. (2025). Rating Roulette: Self-Inconsistency in LLM-As-A-Judge Frameworks. Findings of EMNLP, 24986–25004. https://doi.org/10.18653/v1/2025.findings-emnlp.1361

[52] Liu S, Li X, Liu Z, Yan Y, Yang C, Zeng Z, Liu Z, Sun M, Yu G. (2025). Judge as A Judge: Improving the Evaluation of Retrieval-Augmented Generation through the Judge-Consistency of Large Language Models. Findings of ACL, 5788–5807. https://doi.org/10.18653/v1/2025.findings-acl.301

[53] Cruz Rivera S, Liu X, Chan AW, et al. (2020). Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. Nature Medicine, 26:1351–1363. https://doi.org/10.1038/s41591-020-1037-7

APPENDIX A
PAIRED COHORT ACCOUNTING

Figure 8 reports mean scores for 70 matched questions, excluding the same four audit-flagged pairs from both conditions. The plotted means are described in Section IV-E. The complete-cohort paired estimates retain all 74 questions, and Table VI accounts separately for the four adverse response changes. The excluded identifiers are T05-PREG-002, T05-ANC-020, T05-COM-009, and T05-POST-020; exclusion applies to both response rows for each identifier.

The retrieval-effect counts sum to all 74 pairs: 11 improved, 56 neutral, 3 unassessable, and 4 harmful-case flags. These annotations are retained independently of revised holistic preferences. Numeric ratings remain possible for the three pairs with unassessable retrieval attribution. The original source texts, clinical keys, and retrieved passages were not rewritten during rating correction.

APPENDIX B
REPRODUCIBILITY AND ADDITIONAL EXECUTION EVIDENCE

A. Generation and retrieval settings

Both candidate configurations used CPU inference. The single-question reference collection used an approximately 8 GB Android emulator, and the paired retrieval experiment used a 6 GB emulator. Paired generation used temperature 0.0, a 512-token output limit, and recorded seed 17, with tools disabled in both conditions. The candidate single-question collection also disabled tools and accepted retrieval for 26/150 questions, supplying 61 passages in total.

Conversation assembly used a 4096-token runtime ceiling, at most five completed user–assistant turns, and a 3072-token history budget within the total prompt limit. Retrieval used SQLite FTS5 with BM25 ranking. Raw confidence of at least 95 entered the high-confidence pool; scores from 65 to below 95 required metadata and action checks. Pool scores were normalized and passages below 80 were removed, with a maximum of three retained passages. These are implementation thresholds, not probabilities of clinical relevance. The current model-facing retrieval tool uses the same bounded search service but returns no more than two passages per invocation, with at most two tool-call rounds. The three-passage limit describes the underlying candidate pool and the recorded evidence-context collections; it is not the current tool-output limit.

B. Selection indices and timing definitions

The quality index is 100 times the equal mean of response/4, context/4, safety/4, (benchmark tool behavior + 5)/10, and the proportion of conversations without a critical error. The resource index averages the reference memory-headroom score, response speed relative to the faster configuration, and the fraction of clean turns. Figure 5 shades quality and resource indices of at least 55 and draws a dashed segment at quality 44 across resource scores 60–95. These drawing rules are project decision illustrations rather than estimated acceptance boundaries.

The paired timing record times retrieval separately from generation; its total-latency label repeats recorded generation time. It should not be interpreted as a directly measured sum of all interaction stages. First-token values are reported only when the relevant runtime instrumentation supplies them. Peak RSS is a sampled process maximum, not a continuous trace of device memory.

TABLE VII
ADOPTED ARTIFACT AND INFERENCE CONFIGURATION FOR REPRODUCTION. ARTIFACT INTEGRITY HASHES ARE RETAINED IN THE REPOSITORY SPECIFICATION.

Configuration	Artifact and precision	Runtime and controls
Gemma 4 E2B	`gemma-4-E2B-it.litertlm`; mobile mixed 2/4/8-bit scheme	LiteRT-LM Android 0.18.0; native tokenizer/template; CPU
Qwen3.5 0.8B	`Qwen3.5-0.8B-Q4_K_M.gguf`; four-bit class	Android llama.cpp; GGUF tokenizer/template; CPU
Paired RAG run	Same selected Gemma artifact and 21,673-passage package in both arms	6 GB Android guest; temperature 0.0; 512 output tokens; seed 17; tools disabled

C. Clinician-rating agreement estimator

Absolute agreement among the three retained rating columns uses the two-way single-rating estimator [45]

$$\begin{aligned} \mathrm{ICC}(A,1) &= \frac{MS_Q - MS_E}{D}, \\ D &= MS_Q + (k-1)MS_E \\ &\quad + \frac{k(MS_R - MS_E)}{n}. \end{aligned} \tag{4}$$

Here MS_Q, MS_R and MS_E are the question, reviewer and residual mean squares, respectively; $n = 147$ complete questions and $k = 3$ rating columns. Question-level resampling with 2,000 replicates supplies the percentile intervals. This single-rating coefficient describes agreement within the retained reviewer panel; it is algebraically equivalent to the ICC(2,1) coefficient in the Shrout–Fleiss notation [45], [46].

APPENDIX C
RATING DEFINITIONS AND INTERPRETATION

For the four displayed RAG dimensions, the working rubric uses 0 for an unusable response, 1 for a major failure, 2 for partial performance, 3 for a useful response with a limitation, and 4 for a sufficient and appropriate answer to the actual question. Usefulness does not require exhaustive optional content. Correctness concerns claims made; omissions are evaluated through the relevant usefulness or essential-content requirement. Safety includes urgency for current danger signs.

Actionability includes a usable explanation when that is what the user requested.

Completeness, clarity, directness, groundedness, and evidence-use fields retain earlier rating passes and were not substituted for the revised displayed measures. Groundedness is inapplicable to RAG-off. Critical-error and warning-sign fields are separate from the four retrieval-audit flags used for the conditional display. The working workbook preserves these distinctions and question-specific notes.