mamabench and mamaretrieval: Benchmarks for Evaluating Medical Retrieval-Augmented Generation in Maternal, Neonatal, and Reproductive Health

REN YI 任一, École Polytechnique Fédérale de Lausanne, Switzerland bonjour@renyi.ch

Medical question-answering benchmarks rarely cover the maternal, neonatal, child, and reproductive-health questions a nurse-midwife asks, and, to our knowledge, no public chunk-level relevance benchmark exists for maternal-health guideline retrieval. We release two benchmarks that fill these gaps. **mamabench** is a scope-filtered QA set of 25,949 items assembled from seven existing expert-authored sources across multiple-choice, short-answer, and rubric-graded tracks; to help users calibrate the LLM judge that scores the rubric track, we re-scope HealthBench's physician-labelled meta-evaluation to the domain. **mamaretrieval** pairs 3,185 clinical queries with graded (0–6) relevance labels over a 63,650-chunk maternal-health guideline corpus, using a decomposed rubric that distinguishes a chunk that answers a query from one merely on its topic. Three decisions shape both: assemble and filter expert sources rather than author questions, grade relevance rather than binarise it, and measure and disclose the limits of the labels—scope-classifier agreement, a frontier-judge check, and a pooling-completeness audit—rather than treat them as an oracle. A companion paper uses the benchmarks to evaluate a deployed on-device assistant; both are released openly for research.

Additional Key Words and Phrases: benchmark, retrieval, medical question answering, maternal health, OBGYN, LLM-as-judge, TREC pooling

1 Introduction

Medical question-answering benchmarks have proliferated, but maternal, neonatal, child, and reproductive health—and the sub-Saharan, midwife-led primary-care setting in particular—are a thin slice of them. The widely used sets draw on US and Indian licensing exams (MedQA [Jin et al. 2021], MedMCQA [Pal et al. 2022]) or broad clinical chat (HealthBench [Arora et al. 2025]); AfriMed-QA [Nimo et al. 2025] is pan-African but multi-specialty. None is scoped to the maternal, neonatal, child, and reproductive-health questions a nurse-midwife actually asks at the point of care.

Retrieval is worse served. Medical retrieval-augmented-generation benchmarks such as MIRAGE [Xiong et al. 2024] score the *end-to-end* answer drawn from a corpus, not whether an individual retrieved passage is relevant; the classical medical-IR relevance collections (the TREC Clinical Decision Support and Precision Medicine tracks) judge full PubMed articles against case descriptions, not the short guideline chunks a deployed RAG system retrieves and injects into a prompt. To our knowledge there is no public chunk-level relevance benchmark for maternal-health guideline retrieval.

We release two benchmarks[1] that close these gaps for the domain. Both were built to supply the test data for a companion system paper [Ren Yi 2026]—a deployed, fully on-device RAG assistant for nurse-midwives in Zanzibar—but each is usable as a standalone instrument.

- **mamabench** (§2) is a scope-filtered question-answering benchmark: 25,949 items assembled from seven expert-authored sources across multiple-choice, short-answer, and rubric-graded tracks, plus a 6,853-triple judge-calibration side-file—not labels of our own, but HealthBench's physician annotations re-scoped to this domain—for vetting an LLM judge before it

[1]Datasets: https://huggingface.co/datasets/nmrenyi/mamabench, https://huggingface.co/datasets/nmrenyi/mamaretrieval. Construction code: https://github.com/nmrenyi/mamabench, https://github.com/nmrenyi/mamaretrieval. Released for research use; some sources are non-commercial (e.g. AfriMed-QA, CC BY-NC-SA 4.0), with each item's license recorded in its per-source manifest.

is trusted to score the open-ended items. Its scope classifier is cross-checked by agreement with a larger model and parity against prior labels.

- **mamaretrieval** (§3) pairs 3,185 clinical queries with graded (0–6) relevance labels over a 63,650-chunk maternal-health guideline corpus—a chunk-level relevance benchmark at the granularity a RAG system actually retrieves, built around a graded clinical-relevance rubric grounded in IR relevance theory, with its judge checked against a frontier reference model.

Three decisions shape both benchmarks, and they are the through-line of this paper:

(1) **Assemble and filter from expert-authored sources rather than author new questions.** The clinical curation already exists in licensing exams, physician panels, and published guidelines, so the benchmark's job is reliable scoping and uniform packaging, not de-novo clinical authoring.
(2) **Decompose clinical relevance into a graded, multi-dimensional rubric rather than a binary relevant/not-relevant label.** To a midwife, a chunk that names a drug and a chunk that gives its dose, route, and threshold are not equally useful, and a benchmark that cannot tell them apart cannot separate strong retrieval from adequate retrieval.
(3) **Measure and disclose the limits of the labels**—scope-classifier agreement, a frontier-judge check, and a pooling-completeness audit—rather than present the benchmark as an oracle.

The companion system paper uses these benchmarks to evaluate the deployed system and reports those results; this paper documents how the benchmarks were *built and validated*, and is what the system paper cites for their provenance. We do not reproduce the deployed-system evaluation here. mamabench is assembled from external QA sources and is corpus-independent, whereas mamaretrieval is coupled to a specific corpus chunk set (§3.1). Section 4 is candid about what each benchmark does and does not establish; the datasets and construction code are linked in the title footnote.

2 mamabench: A Scope-Filtered QA Benchmark

mamabench evaluates how well a model answers the questions a nurse-midwife asks—those spanning maternal, neonatal, child, and reproductive health—in formats from multiple-choice to free-text clinical scenarios. We build it by assembling such questions from seven existing, expert-authored sources and filtering each to that scope. This poses three construction problems: assembling and normalising the sources into one scorable schema, defining and applying the scope reliably, and giving consumers a way to trust the LLM judge that scores the open-ended items. We take each in turn.

2.1 Source assembly

The first decision is to assemble, not author: writing and validating maternal-health questions from scratch needs clinicians, whereas the seven sources already carry expert curation—medical-licensing boards, a pan-African physician panel, Kenyan clinicians, and OpenAI's physician annotators—so the benchmark's contribution is scoping and uniform packaging, not de-novo clinical authoring. Each source is normalised by a dedicated adapter into one canonical schema and sorted into one of three tracks (Table 1):

- **Multiple-choice** (23,241 items), for breadth of clinical knowledge: the OBGYN and paediatrics subset of MedMCQA, the in-scope portion of MedQA-USMLE, and the multiple-choice questions of AfriMed-QA. This track serves breadth; the open-ended tracks below carry the weight.

Table 1. The seven sources of mamabench, by track. *Yield* is the share of a source's upstream pool that our scope classifier (§2.2) keeps as in-scope; a dash marks sources that arrive already scoped to the domain—MedMCQA and AfriMed-QA by their subject or specialty labels, the Women's Health Benchmark by expert authoring—and so bypass the classifier entirely.

Track	Source	Items	Yield	Reference
Multiple-choice	MedMCQA (OBGYN & paediatrics)	18,508	–	[Pal et al. 2022]
	MedQA-USMLE	4,199	29.2%	[Jin et al. 2021]
	AfriMed-QA	534	–	[Nimo et al. 2025]
Short-answer	Kenya Clinical Vignettes	312	61.5%	[Mwaniki et al. 2025]
	AfriMed-QA (short-answer)	37	–	[Nimo et al. 2025]
	Women's Health Benchmark	20	–	[The Lumos AI Labs 2025]
Rubric-graded	HealthBench (oss_eval)	1,209	24.2%	
	HealthBench (consensus)	872	23.8%	[Arora et al. 2025]
	HealthBench (hard)	258	25.9%	
Total		**25,949**		

- **Short-answer** (369 items), closest to the deployment setting: 312 nurse-written primary-care cases from the Kenya Clinical Vignettes, the short-answer questions of AfriMed-QA, and a small set of expert-crafted items designed to expose model errors (the Women's Health Benchmark).
- **Rubric-graded** (2,339 items): the in-scope portion of HealthBench, whose items are scored against physician-written rubric criteria rather than a single gold answer.

Normalisation is mostly mechanical, but one source needs more than that: the AfriMed-QA multiple-choice items carry quality issues the others do not, and there we draw the line at *scorability*. Multiple-choice items are scored by the answer's letter, and we drop only what that scoring cannot represent: items with more than one correct option (a single answer index cannot encode them), and items whose correct-answer text is duplicated under another letter—if the same option text sits at two positions, a model could pick the right text yet be marked wrong for reporting the duplicate letter. Cosmetic issues that do not affect scoring (embedded option-letter prefixes, non-ASCII typography, terse stems) are left untouched and documented in the manifest. The principle is that the benchmark may refuse an item it cannot score, but must never change what an item means.

2.2 Defining scope with an LLM classifier

The central construction decision is how to decide whether an item is in scope. We define scope as four positive categories—MATERNAL, NEONATAL, CHILD HEALTH, and SEXUAL AND REPRODUCTIVE HEALTH (SRH)—plus NONE, and we classify each item by its *primary clinical concept, not the patient's demographics*. A question about managing a wrist fracture in a patient who happens to be pregnant is NONE—the fracture is the concept, the pregnancy incidental—whereas postpartum depression is MATERNAL, because here the postpartum context *is* the concept. The classifier is `Qwen3.6-27B-FP8`,[2] run with reasoning enabled at temperature 0; it applies this rule to every candidate item and emits a structured verdict with its reasoning retained for audit.

Trusting an LLM to draw the scope boundary needs evidence, and we offer two checks. First, **cross-model agreement**: re-classifying the HealthBench `oss_eval` items with the much larger

[2]Model card: https://huggingface.co/Qwen/Qwen3.6-27B-FP8.

`Qwen3.5-397B-A17B-FP8`[3]—a stronger model from the same family, so this checks self-consistency more than independence—agrees with the 27B on 98.1% of the 4,988 jointly classified items, and per-category counts differ by under one percent; the handful of items on which the 27B fails to converge are all NONE under the 397B, so they would be filtered out either way and the net effect on in-scope counts is zero. Second, **parity against prior labels**: the Kenya vignettes arrived with their own classifier labels, and ours agree with them on 86.8%, with the disagreements dominated by our rule being deliberately stricter—rejecting items where only the patient's being female, rather than the clinical concept, suggested OBGYN scope.

The per-source yield in Table 1 is itself a useful signal. The Kenyan primary-care vignettes are 61.5% in scope, against 24–29% for the general benchmarks: the maternal-health sources concentrate exactly where the deployment lives, while the general sets contribute a smaller, scope-filtered slice.

2.3 A judge-calibration set, re-scoped from HealthBench

The rubric-graded track is scored by an LLM judge, which raises the obvious question of whether the judge can be trusted. Rather than answer that once for our own judge, we ship the means to answer it for *any* judge: a calibration side-file of 6,853 (prompt, response, criterion) triples carrying physicians' met / not-met labels. We want to be precise about provenance, since it is easy to overclaim here. *We did not produce these labels.* They are the physician annotations HealthBench released for vetting rubric graders—its meta-evaluation set [Arora et al. 2025]; our only role is to re-scope that set to this domain—keeping the triples whose prompts the scope classifier (§2.2) places in the four positive categories (CHILD HEALTH 3,239, MATERNAL 2,284, SRH 942, NEONATAL 388, over 872 prompts)—and to package them as a ready-to-use calibration file alongside the benchmark. A consumer runs a candidate judge over these triples, compares its labels to the physicians', and decides whether to trust it before scoring the benchmark; the companion system paper selects its rubric judge in exactly this way [Ren Yi 2026]. One caveat: these 872 prompts *are* the HealthBench-consensus rubric track's prompts—HealthBench collected the side-file's physician labels on that same set—so calibrating a judge here and then scoring the consensus track reuses the same prompts; calibrate on the side-file, but read consensus-track scores with that overlap in mind.

3 mamaretrieval: A Chunk-Level Relevance Benchmark

mamaretrieval scores how well a retriever finds the guideline passages that answer a clinical query, at the chunk granularity a RAG system actually retrieves. Building it means choosing what to retrieve over (the corpus), what queries to ask, how to label a (query, chunk) pair's relevance, and how to keep those labels trustworthy at scale. We start with the corpus, then take query construction, the relevance rubric, pooling, and judge reliability in turn.

3.1 The guideline corpus

mamaretrieval is built over the same maternal-health guideline corpus the deployed system retrieves from, which the companion system paper [Ren Yi 2026] describes in full—its sources, construction, and packaging. We give here only what the benchmark depends on: the corpus's scale, its source-quality tiers, and the version coupling that makes the benchmark and its corpus one artifact.

Scale and sources. The corpus bundle (v0.2.0) holds 63,650 section-aware chunks drawn from 87 documents spanning international clinical guidance (WHO, NICE, MSF, the Hesperian midwifery references, and FIGO/EmONC materials), national and local protocols (the Tanzania and Zanzibar Ministry of Health guidelines), and standard obstetric and midwifery reference texts. Each chunk carries a canonical header recording its source, page, and a stable chunk identifier.

[3]Model card: https://huggingface.co/Qwen/Qwen3.5-397B-A17B-FP8.

Source-quality tiers. Sources are rated into quality tiers (from *very high* down to *low–moderate*) by their clinical depth and relevance to the Zanzibar OBGYN and midwifery scope. Query construction (below) uses these tiers as sampling weights, drawing proportionally more from the clinically richest sources.

Version coupling. Chunk identifiers are content-derived—a hash of the chunk text—so they survive source renames and page drift but change whenever a chunk's text changes. mamaretrieval references chunks by these identifiers, which couples it to a specific *chunk set* rather than to a specific embedding. The distinction is not hypothetical: the deployed system moved from corpus bundle v0.2.0 to v0.3.0, but that update only re-embedded the corpus—swapping the on-device embedder—and left all 63,650 chunks and their identifiers untouched, so every mamaretrieval label applies to the deployed v0.3.0 corpus unchanged. A genuine re-chunking would change the identifiers and require re-running the labelling pipeline; we therefore treat the (chunk set, queries, labels) triple as a single versioned artifact rather than versioning queries and labels on their own.

3.2 Query construction

A retrieval benchmark needs queries, and for this corpus there are none to mine—no search logs, no question bank tied to these guidelines. We generate them. Each query in mamaretrieval is written by an LLM from a single corpus chunk, asking a clinical question that the chunk answers; the chunk that seeded a query is recorded as a known-relevant anchor for it.

The queries come out of a three-step funnel over the corpus:

(1) **Tier-weighted sampling** draws 4,540 chunks, with each source's quota set by its quality tier (§3.1)—the clinically richest sources contribute proportionally more, so the query set inherits the corpus's quality signal rather than its raw size distribution.
(2) An **answerability gate** passes each sampled chunk to `Qwen3.6-27B-FP8` (reasoning on, temperature 0), prompted to keep the chunk only if a clinician could answer a specific question from it alone and, when so, to write exactly that question.[4]
(3) The kept chunks become the benchmark's **queries**: each generated question gets a stable identifier (`q_00001`, …), a pointer back to its seed chunk, and the source and tier metadata that chunk carried.

The gate keeps 3,185 of the 4,540 chunks (about 70%), and the 30% it discards are themselves informative: assessment rubrics, learning objectives, and table-of-contents fragments are rejected because no answerable clinical question follows from them.

One limitation of this construction is worth flagging up front, because it shapes which metrics we trust: a query written from a chunk tends to be phrased like that chunk, which gives dense embedding retrievers an artificial edge over lexical ones on the very chunk that seeded the query. We do not correct for this anchor-text effect; we disclose it (§3.5). Section 4 returns to the design's other limits, including that the queries are clean and each grounded in a single chunk.

3.3 The graded relevance rubric

With queries and a corpus, the benchmark needs a relevance label for each (query, chunk) pair: how useful is this chunk for answering this query? Our first rubric made that judgement coarse—effectively relevant or not. It saturated. Take a query asking for alternative iron and folic-acid dosages (Table 2): a footnote giving the exact substitute formulation, a dosing section padded with storage advice, and an adherence-counselling note that never states a dose are, by a binary test,

[4]The full filter-and-generate prompt ships with the mamaretrieval construction code (title footnote).

equally relevant—each is on topic, meaningful, and broadly actionable. A benchmark that cannot separate them cannot tell a retriever that surfaces the dose from one that surfaces the filler.

So we replaced the binary label with a graded one. Each (query, chunk) pair is scored

$$\text{score} = D_1 \times (D_2 + D_3 + D_4) \in \{0, \ldots, 6\},$$

one gate and three graded dimensions:

- D_1 — **topic (0/1).** Does the chunk address the query's clinical question in the same context—the same timing window (antenatal, intrapartum, postpartum)? A zero here forces a score of zero and halts scoring, which also keeps the judge from spending tokens on off-topic chunks.
- D_2 — **meaningful content (0/1/2).** How much clinical depth the chunk carries, from a bare mention to a multi-concept explanation.
- D_3 — **actionable guidance (0/1/2).** How specific the guidance is: none, a general instruction ("give a uterotonic"), or an exact one ("10 IU oxytocin IM").
- D_4 — **density (0/1/2).** What fraction of the chunk is useful for *this* query, penalising a relevant passage buried in unrelated text.

A structural rule ties them together: actionable guidance presupposes some meaningful content ($D_3 > 0 \Rightarrow D_2 > 0$).

This is not an ad-hoc scheme. A topical gate followed by complementary dimensions that a fixed rule aggregates is the design of the TREC Precision Medicine track [Roberts et al. 2017], and the topical/cognitive/situational split it rests on is Saracevic's [Saracevic 2007]—here D_1 is topical and D_3, D_4 situational; criteria-based decomposition is also the direction of recent LLM relevance-judgment work [Farzi and Dietz 2025]. Two further choices are specific to using an LLM as the judge: an explicit topical gate, and reasoning-before-verdict—the structured-output schema makes the judge emit its reasoning before the four scores, as in G-Eval [Liu et al. 2023]—both following the now-standard use of a strong LLM as the relevance assessor [Upadhyay et al. 2024]. The judge is `Qwen3.5-397B-A17B-FP8` at temperature 0, running the prompt reproduced in full in Appendix A; §3.5 reports how well its labels hold up.

Table 2 shows the graded rubric on that query. All four chunks pass the topic gate; the first three also clear a binary rubric's bar identically—each is meaningful and actionable—yet the graded rubric ranks them apart. The footnote giving the exact alternative formulation scores 6, specific and entirely on-query; a dosing section half-taken up by storage and supply logistics scores 4 ($D_4 = 1$); and an adherence note that never states a dose scores 2. The fourth chunk is on topic but pure cost commentary, and drops to 0—the rubric counts resource text as no meaningful clinical content ($D_2 = 0$). The judge's reasoning ships with every label and makes such calls explicit: of that last chunk, it notes the passage sits "literally under a … Resources heading and discusses cost," and scores $D_2 = 0$.

3.4 Pooling and label completeness

Scoring every chunk against every query is infeasible—63,650 chunks times 3,185 queries—so mamaretrieval labels a *pool*: for each query, the union of the top-ranked chunks returned by several retrievers, with every pooled pair judged and every unpooled chunk treated as not-relevant. This is the standard information-retrieval pooling protocol, and its one risk is well known: if the pool misses relevant chunks, the labels under-count them. We measured that risk rather than assume it away.

[5]https://huggingface.co/Octen/Octen-Embedding-8B
[6]https://www.voyageai.com

Table 2. The graded rubric on one query (q_00041 in release v0.2.0)—"What are the alternative dosages for iron and folic acid supplementation?"—as scored by the production judge over four *on-topic* chunks (D_1=1 throughout). Text is quoted from the released labels with markdown list formatting flattened to prose and elisions marked "…"; sources in brackets (MSF, the MSF obstetric manual; WHO PCPNC, the WHO pregnancy/childbirth guide; WHO ANC, the 2016 WHO antenatal-care guideline). A binary label cannot separate the first three; the graded score ranks them 6, 4, and 2. The four chunks (top to bottom) have identifiers 318a18d9, ad7678f8, 466a314b, and 27b615ab.

Retrieved chunk (verbatim)	D_1	D_2	D_3	D_4	Score
[MSF] "200 mg ferrous sulfate (65 mg elemental iron) + 400 micrograms folic acid tablets may be replaced by 185 mg ferrous fumarate (60 mg elemental iron) + 400 micrograms folic acid tablets."	1	2	2	2	**6**
[WHO PCPNC] "Give iron and folic acid. To all pregnant, postpartum and post-abortion women: routinely once daily in pregnancy and until 3 months after delivery or abortion; twice daily as treatment for anaemia (double dose). Check woman's supply … at each visit and dispense 3 months supply. Advise to store iron safely …"	1	2	1	1	**4**
[WHO PCPNC] "Motivate on adherence with treatments. Explore local perceptions about iron treatment (examples of incorrect perceptions: making more blood will make bleeding worse, iron will cause too large a baby)."	1	1	1	0	**2**
[WHO ANC] "Resources. Intermittent iron and folic acid supplementation might cost a little less than daily iron and folic acid supplementation due to the lower total weekly dose of iron."	1	0	0	0	**0**

Table 3. The six retrievers whose top-k unions form mamaretrieval's label pool, spanning retrieval paradigms. They *construct* the labels; the retriever scoreboard *over* them—which one best serves the deployed system—is reported in the companion paper [Ren Yi 2026], not here. Lateon is the HuggingFace model lightonai/GTE-ModernColBERT-v1.

Retriever	Paradigm	Reference
BM25	lexical (sparse)	[Robertson and Zaragoza 2009]
MedCPT	biomedical bi-encoder	[Jin et al. 2023]
Octen-Embedding-8B	general-purpose dense	model card[5]
voyage-4-large	frontier API dense	Voyage AI[6]
Lateon	late-interaction	[Khattab and Zaharia 2020]
Gecko	on-device dense	[Lee et al. 2024]

The first pool took the top-10 chunks from three retrievers (BM25, MedCPT, and Octen-Embedding-8B): about 24.7 candidates per query and 78,571 judged pairs. A completeness audit on a 100-query sample then judged a wider pool (five retrievers at top-20, adding voyage-4-large and

Lateon) and found the original pool had captured only about half of all the relevant chunks known across both judging rounds (lenient recall near 0.49): half the relevant chunks were being labelled not-relevant by omission.

That finding drove the released label set. mamaretrieval v0.2.0 pools the top-20 chunks from six retrievers—the five above plus Gecko—chosen to span retrieval paradigms so the pool is not biased toward any one (Table 3). Judged across all 3,185 queries, this yields 230,964 graded labels: the pool the companion system paper draws its label provenance from. A residual gap remains—chunks that no retriever ranks in its top-20 are still unjudged—but the same audit bounds it: against that same reference, widening from the initial pool to the five-retriever top-20 union raised lenient recall from about 0.49 to about 0.99, and the released pool adds a sixth retriever, so it is near-saturated at the top of the ranking. Only absolute coverage-sensitive numbers need the residual caveat (§3.5).

3.5 Judge reliability and metric guidance

A single LLM produces every relevance label, so the benchmark is only as trustworthy as that judge. We check it against a frontier reference model (Claude Opus 4.7) on a 62-pair pilot: the production judge agrees on 59 of 62 pairs at the lenient cut (score ≥ 3, any useful content) and 53 of 62 at the strict cut (score ≥ 5, complete and specific), and a stratified manual spot-check of its reasoning traces found its calls defensible, including on hard cases such as drug-name collisions across clinical contexts. Those two cuts—lenient ≥ 3 and strict ≥ 5—are the relevance thresholds the benchmark reports against.

We are deliberate about what this establishes. It is an agreement check against another model, not against clinicians: it shows the judge is a reasonable relevance classifier and that retriever *rankings* are stable under it, but it does not make any single label ground truth. Each label is one judge's opinion, and absolute, per-item numbers carry more uncertainty than relative comparisons.

That uncertainty, together with the pool incompleteness of §3.4, is why we guide consumers on metric choice. Metrics that turn on the presence or rank of a relevant chunk—Hit Rate@k (is any of the top k relevant?) and MRR (the mean over queries of $1/r$, where r is the rank of the *first* relevant chunk)—are robust to a pool that misses some relevant chunks: one hit suffices, and a label missed elsewhere cannot change that a retrieved relevant chunk is relevant. nDCG@k is moderately sensitive: it credits every relevant chunk by grade and rank and normalises against an ideal ranking built from the judged-relevant set, so missing labels distort both. Precision@k understates any retriever that surfaces a relevant-but-unjudged chunk (counted as a miss), and Recall@k is the most exposed, since its denominator—the total count of relevant chunks—is exactly what pooling under-counts. We recommend leading with the robust metrics and reading the rest as relative comparisons with the completeness caveat attached—the order the companion system paper follows when it evaluates retrievers on mamaretrieval.

4 Limitations

We have flagged limits where they arise; here we collect what a user should weigh before relying on either benchmark.

No human gold standard. The scope of mamabench and every relevance label in mamaretrieval are decided by LLMs, not clinicians. We anchor them externally—cross-model agreement and prior labels for the scope classifier (§2.2), a frontier reference model for the relevance judge (§3.5)—and we ship HealthBench's physician labels so a consumer can vet a rubric judge (§2.3). But no midwife reviewed either benchmark at scale, and those physician labels come from HealthBench's annotator pool, not midwifery specialists. Each label is best read as one judge's opinion: model and retriever *rankings* are robust, absolute per-item numbers less so.

mamaretrieval queries are clean and single-chunk. Every query is LLM-written from one chunk and answerable from it (§3.2), with two consequences. First, the benchmark exercises single-passage retrieval: it does not test questions that can only be answered by combining several chunks, nor robustness to the noisy, code-switched, or mistyped queries a hurried nurse might enter. Second, because a query reads like its source chunk, dense retrievers gain an anchor-text edge that inflates their absolute scores—another reason to lead with the pooling-robust metrics (§3.5).

The relevance labels are pooled. They cover the six-retriever top-20 union; a chunk no retriever surfaced is unjudged and counted not-relevant (§3.4). Coverage-sensitive metrics—Recall and Precision—carry that bias, which is why we report Hit Rate and MRR first.

Scope, language, and transfer. Both benchmarks are English-only and scoped to maternal, neonatal, child, and reproductive health for a Zanzibar nurse-midwife setting; the numbers need not transfer to other languages, specialties, or care settings. A few mamabench items the scope classifier could not converge on are disclosed and have zero net effect on in-scope counts (§2.2).

5 Conclusion

We release two benchmarks for evaluating medical retrieval-augmented generation in maternal, neonatal, child, and reproductive health: **mamabench**, a scope-filtered QA set assembled from seven expert sources, and **mamaretrieval**, a chunk-level relevance benchmark with a graded rubric that distinguishes a chunk that answers the query from one that is merely on-topic. Three choices run through both: assemble and filter rather than author, decompose relevance into a graded rubric rather than a binary label, and measure and disclose the limits of the labels rather than treat them as an oracle. The companion system paper uses them to locate where a deployed on-device assistant succeeds and fails, but they also stand alone—resources for maternal-health RAG that did not previously exist. The clearest next steps are the ones we left open: queries that demand multi-passage synthesis and survive noisy phrasing, and validation against midwives rather than against other models.

References

Rahul K. Arora, Jason Wei, Rebecca Soskin Hicks, et al. 2025. HealthBench: Evaluating Large Language Models Towards Improved Human Health. arXiv:2505.08775 [cs.CL] https://arxiv.org/abs/2505.08775 Dataset: https://huggingface.co/datasets/openai/healthbench.

Naghmeh Farzi and Laura Dietz. 2025. Criteria-Based LLM Relevance Judgments. arXiv:2507.09488 [cs.IR] https://arxiv.org/abs/2507.09488

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What Disease Does This Patient Have? A Large-Scale Open Domain Question Answering Dataset from Medical Exams. *Applied Sciences* 11, 14 (2021). https://github.com/jind11/MedQA

Qiao Jin, Won Kim, Qingyu Chen, Donald C. Comeau, Lana Yeganova, W. John Wilbur, and Zhiyong Lu. 2023. MedCPT: Contrastive Pre-trained Transformers with large-scale PubMed search logs for zero-shot biomedical information retrieval. *Bioinformatics* 39, 11 (2023), btad651.

Omar Khattab and Matei Zaharia. 2020. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*.

Jinhyuk Lee, Zhuyun Dai, Xiaoqi Ren, et al. 2024. Gecko: Versatile Text Embeddings Distilled from Large Language Models. arXiv:2403.20327 [cs.CL] https://arxiv.org/abs/2403.20327

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Paul Mwaniki, Wycliffe Musau, Lynda Isaaka, Conrad Wanyama, Vinod Menon, Alastair K. Denniston, Xiaoxuan Liu, Mphatso Emmanual-Fabula, Gerald Williams, Bilal A. Mateen, and Ambrose Agweyu. 2025. Benchmarking Large Language Models

and Clinicians Using Locally Generated Primary Healthcare Vignettes in Kenya. https://www.medrxiv.org/content/10.1101/2025.10.25.25338798v1 medRxiv preprint 2025.10.25.25338798. Data and code: https://github.com/pmwaniki/vignette.

Charles Nimo, Tobi Olatunji, et al. 2025. AfriMed-QA: A Pan-African, Multi-Specialty, Medical Question-Answering Benchmark Dataset. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*. 1948–1973. https://aclanthology.org/2025.acl-long.96/ Dataset: https://huggingface.co/datasets/intronhealth/afrimedqa_v2.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. MedMCQA: A Large-scale Multi-Subject Multi-Choice Dataset for Medical Domain Question Answering. In *Proceedings of the Conference on Health, Inference, and Learning (CHIL)*. https://huggingface.co/datasets/openlifescienceai/medmcqa

Ren Yi. 2026. MAM-AI: An On-Device Medical Retrieval-Augmented Generation System for Nurses and Midwives in Zanzibar. arXiv:2606.29580 [cs.CL] Companion paper.

Kirk Roberts, Dina Demner-Fushman, Ellen M. Voorhees, William R. Hersh, Steven Bedrick, Alexander J. Lazar, and Shubham Pant. 2017. Overview of the TREC 2017 Precision Medicine Track. In *Proceedings of the Twenty-Sixth Text REtrieval Conference (TREC)*.

Stephen Robertson and Hugo Zaragoza. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends in Information Retrieval* 3, 4 (2009), 333–389.

Tefko Saracevic. 2007. Relevance: A Review of the Literature and a Framework for Thinking on the Notion in Information Science. Part II. *Journal of the American Society for Information Science and Technology* 58, 13 (2007), 1915–1933.

The Lumos AI Labs. 2025. Women's Health Benchmark (WHB). https://huggingface.co/datasets/TheLumos/WHB_subset HuggingFace dataset (unrefereed); expert-authored maternal/reproductive-health question set.

Shivani Upadhyay, Ronak Pradeep, Nandan Thakur, Nick Craswell, and Jimmy Lin. 2024. UMBRELA: The Open-Source Reproduction of the Bing Relevance Assessor. arXiv:2406.06519 [cs.IR] https://arxiv.org/abs/2406.06519

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. Benchmarking Retrieval-Augmented Generation for Medicine. In *Findings of the Association for Computational Linguistics: ACL 2024*.

A Relevance-Judge Prompt

The prompt the relevance judge (`Qwen3.5-397B-A17B-FP8`) runs for every (query, chunk) pair is reproduced below: the system message carrying the four-dimension rubric, the structural rule, the output schema, and four worked examples, followed by the per-pair user template. It is the canonical prompt from `scripts/judge_relevance.py` in the construction repository,[7] and ships verbatim with the dataset as `audit/judge_relevance_prompt.txt`; its hash is pinned in the release manifest. The box rules and a few Unicode symbols (e.g. ≥, ×, μ) are rendered in ASCII here, and long lines are wrapped to the column; the content is otherwise verbatim.

```
# Judge prompt (v2 graded rubric)
# ----------------------------------------------------------------
# Sourced verbatim from scripts/judge_relevance.py at release time.
# prompt_hash and schema_version in manifest.json pin this content.
# ----------------------------------------------------------------

## SYSTEM message

You are a clinical relevance judge for MAMAI, a RAG system serving midwives and doctors in Zanzibar. Given a
clinical query and a retrieved guideline chunk, score the chunk on four dimensions. Think carefully, then emit
JSON.

--------------------------------------------------------------------------
D1 -- Topic (boolean: true / false)
--------------------------------------------------------------------------
Does this chunk address the same clinical problem as the query?

  TRUE when the chunk concerns the SAME:
    - condition, diagnosis, or syndrome the query asks about
    - clinical event, complication, or procedure the query asks about
    - drug or intervention the query asks about
    - clinical timing/context (antenatal / intrapartum / postpartum /
      neonatal) -- a query about postpartum monitoring is NOT satisfied
      by a chunk about antenatal monitoring of the same parameter.

  FALSE when the chunk:
```

[7] https://github.com/nmrenyi/mamaretrieval/blob/main/scripts/judge_relevance.py

```
    - covers a different condition, procedure, or clinical event
    - covers the same parameter in a different timing/context window
    - shares only a body system or patient population while addressing a
      different clinical problem
    - is pure metadata (TOC, references, learning objectives,
      assessment / test rubrics, course-content outlines)

  If D1 is FALSE, set D2 / D3 / D4 all to 0 and stop evaluating them.

------------------------------------------------------------------------
D2 -- Meaningful clinical content (0 / 1 / 2)
------------------------------------------------------------------------
How rich is the chunk's clinical content (independent of whether it
specifically answers the query)?

  0 -- no meaningful clinical content beyond a topic label:
       section headers, page footers, TOC entries, references,
       citations, learning objectives ("students should be able to..."),
       administrative notes, supersession statements, cost / resource
       text, assessment marking criteria, or one-line topic mentions
       with no body.

  1 -- some clinical content; one or two clinical facts; brief background;
       definition or peripheral mention; epidemiology / statistics
       without management content; a single clinical principle stated.

  2 -- rich clinical content; multiple concepts developed in depth:
       pathophysiology + risk factors + symptoms; formal recommendation
       with indications; detailed mechanism explanation; substantive
       clinical paragraph(s) with named entities and relationships.

------------------------------------------------------------------------
D3 -- Actionable guidance (0 / 1 / 2)
------------------------------------------------------------------------
How specific is the actionable guidance a clinician could use directly?

  0 -- no actionable guidance:
       pure background, definitions, pathophysiology, epidemiology,
       references; or evidence-summary text that names doses studied
       without making a recommendation.

  1 -- general or partial guidance; action specified but missing
       specifics:
       "give a uterotonic"; "monitor vital signs"; "consult an
       obstetrician"; "massage the uterus"; "estimate blood loss";
       "look for these signs" without thresholds; assess-and-record
       statements; bullet-list of clinical signs without diagnostic
       cutoffs.

  2 -- specific complete guidance: exact doses with route and frequency,
       numeric thresholds with action triggers, full step-by-step
       procedure with all steps specified, scheduled monitoring with
       intervals:
       "10 IU oxytocin IM immediately, then 5 IU/hr infusion for 4 hr";
       "BP > 160/110 mmHg + platelets < 100 x10^9/L -> severe pre-eclampsia
       criteria";
       "BP and pulse every 30 min, temperature every 4 hr";
       "200 mg ferrous sulfate (65 mg elemental iron) + 400 ug folic
       acid daily; may be replaced by 185 mg ferrous fumarate".

  STRUCTURAL RULE: if you assign D3 >= 1, then D2 must be >= 1 -- any
  actionable clinical guidance carries some meaningful content. If you
  find yourself wanting D3 >= 1 with D2 = 0, reconsider D2 (the guidance
  itself IS clinical content).

------------------------------------------------------------------------
D4 -- Density relative to THIS query (0 / 1 / 2)
------------------------------------------------------------------------
What fraction of the chunk text is directly useful for answering the
specific query? (Not the broader topic -- the specific query.)
```

```
  0 -- useful content is < 25% of the chunk:
       long chunk with one buried relevant sentence; mostly off-topic
       surrounding text; the answer exists somewhere but is dwarfed by
       irrelevant procedural detail or related-but-not-query-specific
       content.

  1 -- useful content is 25-75% of the chunk:
       mixed -- relevant content interleaved with adjacent-but-not-query-
       specific material (e.g., the right protocol followed by a long
       complications list when the query is about diagnosis only; useful
       dosing followed by task-shifting / cost discussion).

  2 -- useful content is > 75% of the chunk:
       focused -- the chunk is essentially the answer plus minor
       framing. Short focused chunks count as D4=2: signal-to-noise
       from the LLM's point of view, not absolute length, is what
       matters. A 50-character footnote that's 100% on the query is
       D4=2; a 1500-character chunk that's 70% off-the-query is D4=1.

  Note: D4 is judged AGAINST THE QUERY, not against the broader topic.
  A chunk fully about postpartum monitoring is D4=2 only if the query
  is also about postpartum monitoring. If the query is specifically
  "how often to check BP", and the chunk also covers fundal height
  and lochia at length, that's D4=1.

------------------------------------------------------------------------
Reasoning instruction
------------------------------------------------------------------------
Think carefully before producing the JSON output. Walk through each
dimension in order (D1 -> D2 -> D3 -> D4), referring to specific
sentences or passages in the chunk where appropriate. Your reasoning
trace is captured separately; do NOT embed it in the JSON.

------------------------------------------------------------------------
Output format
------------------------------------------------------------------------
After reasoning, respond with valid JSON only -- no "score" field, no
"reasoning" field; the score is computed downstream from D1 / D2 / D3 /
D4 and the reasoning is captured in the raw response.

{
  "d1_topic": true,
  "d2_meaningful": 2,
  "d3_actionable": 2,
  "d4_density": 2
}

------------------------------------------------------------------------
Worked examples
------------------------------------------------------------------------

### Example 1
Query: What are the alternative dosages for iron and folic acid supplementation?
Chunk (WHO ANC 2016): "RECOMMENDATION A.2.1: Daily oral iron and folic acid supplementation with 30 mg to 60 mg
 of elemental iron and 400 ug (0.4 mg) folic acid is recommended for pregnant women to prevent maternal anaemia
, puerperal sepsis, low birth weight, and preterm birth."
Example reasoning:
D1: Iron and folic acid dosing in antenatal care -- same topic, same context.
D1 = true.
D2: A formal recommendation with named indications (maternal anaemia,
puerperal sepsis, LBW, preterm birth) -- multiple clinical concepts in one
paragraph. D2 = 2.
D3: Specific dose ranges (30-60 mg elemental iron, 400 ug folic acid daily)
with route (oral) -- matches the "exact doses with route and frequency"
anchor. D3 = 2.
D4: The entire chunk is the recommendation statement. D4 = 2.
Expected JSON:
{"d1_topic": true, "d2_meaningful": 2, "d3_actionable": 2, "d4_density": 2}

### Example 2
Query: What clinical signs and laboratory findings indicate severe pre-eclampsia and HELLP syndrome?
```

```
Chunk (midwifery-preparation-for-practice): "HELLP SYNDROME -- HELLP syndrome is a rare but life-threatening
liver disorder thought to be a type of severe preeclampsia. Characterised by Haemolysis (destruction of red
blood cells), Elevated Liver enzymes (indicating liver damage), and Low Platelet count. Many hypotheses attempt
 to define the pathogenesis... Both HELLP and preeclampsia occur during the later stages of pregnancy, and
sometimes after childbirth..."
Example reasoning:
D1: HELLP syndrome and severe preeclampsia -- same topic. D1 = true.
D2: Defines HELLP, breaks down the acronym with clinical meaning (RBC
destruction, liver damage, platelet drop), and discusses timing and
relationship to preeclampsia. Substantive clinical definitions. D2 = 2.
D3: The chunk defines the syndrome but provides no diagnostic thresholds
(e.g., AST > X, platelets < Y) and no management steps. Pure
background/pathophysiology. D3 = 0.
D4: The chunk is short and focused on HELLP. The pathogenesis text is
adjacent context but the core (definition + acronym) directly addresses
the query. D4 = 2.
Expected JSON:
{"d1_topic": true, "d2_meaningful": 2, "d3_actionable": 0, "d4_density": 2}

### Example 3
Query: How should a midwife assess and manage uterine tone during the third stage of labour?
Chunk (who-midwifery-education-modules-2): "CHECKLIST OF SUB-TASKS FOR THE MANAGEMENT OF THE THIRD STAGE OF
LABOUR | Sub-task: Ensure uterus well contracted | Knowledge: Consistency of contracted uterus | Skill:
Palpation of uterus and massage to promote contraction | Attitudes: Accuracy, Gentleness"
Example reasoning:
D1: Uterine tone in the third stage of labour -- same topic, same timing.
D1 = true.
D2: Checklist row identifies the sub-task ("Ensure uterus well contracted"),
the knowledge ("Consistency"), and the skill ("Palpation and massage"). Brief
listing, no elaboration. D2 = 1.
D3: Names actions (palpate, massage) but no procedural specifics -- how to
palpate, frequency, when soft vs firm, technique nuance. Partial guidance.
D3 = 1.
D4: The entire table row is on-topic for the query. D4 = 2.
Expected JSON:
{"d1_topic": true, "d2_meaningful": 1, "d3_actionable": 1, "d4_density": 2}

### Example 4
Query: How often should maternal blood pressure, pulse, and temperature be checked after birth?
Chunk (hesperian-a-book-for-midwives Chapter 8 -- Prenatal checkups): "How to check blood pressure: The needle
will begin to go back down. As the air leaks out, you will start to hear the mother's pulse... Check the mother
's blood pressure at each visit. If her blood pressure is going up, ask her to come back every week..."
Example reasoning:
D1: Query asks about postpartum monitoring; chunk is from "Chapter 8
Prenatal checkups" and discusses BP "at each visit" and "come back every
week" -- antenatal context. Same parameter, different clinical timing.
D1 = false.
D2-D4: zeroed per the structural rule (D1 = false).
Expected JSON:
{"d1_topic": false, "d2_meaningful": 0, "d3_actionable": 0, "d4_density": 0}

## USER message template (rendered per query x chunk pair)

## Query
<query_text>

## Guideline chunk
Source: <source> | Page: <page>
---
<chunk text>
---
```