MedicaLLM-Eval : Benchmarking AI Systems for Multilingual Reproductive and Maternal Health

Anonymous ACL submission

Abstract

Maternal health chatbots are already deployed to hundreds of millions of non-English speakers, but the benchmarks that evaluate them are almost entirely English. We built the Sakhi benchmark to close that gap. Sakhi is a parallel English, Hindi, and Marathi dataset of 380 maternal and reproductive health questions in two tracks, 149 doctor-edited pairs and 231 community-sourced pairs, validated by Accredited Social Health Activist (ASHA) workers and doctors in rural India. We evaluate 13 frontier LLMs spanning proprietary and open-weight systems with a two-layer framework: automated linguistic and semantic metrics, and a five-axis, fifteen-criterion clinical rubric (Accuracy, Completeness, Context Awareness, Communication, Terminology Accessibility) scored by an LLM judge with a three-judge calibration on a stratified 500-response subset. Four findings. No model exceeds Medical Quality Score 0.32 on the expert arm in English under the GPT-4o-mini judge, and the stricter GPT-5.1 judge drops every model a further 0.17 to 0.20 MQS on identical outputs. Terminology Accessibility is the weakest axis for every model (pass rates 0.123 to 0.206), below Context Awareness, so the dominant failure is lexical rather than broadly cultural. Apparent cross-lingual parity under a mid-tier judge is largely a judge-leniency artifact and does not survive a stricter judge. Open-weight 27B models (MedGemma 27B, Gemma 3 27B) match proprietary flagship MQS at roughly 20 times lower per-response cost, and medical fine-tuning alone does not lift rubric compliance in this cross-cultural setting. We release the dataset, reference answers, rubric, and judge-calibration records for the NeurIPS Datasets and Benchmarks track.

1 Introduction

Roughly 609 million people in India speak Hindi as a first or second language, and another 99 million speak Marathi (Express, 2025). Many of them now get maternal-health information from conversational AI agents on messaging and social media platforms, including Meta AI and a growing set of government and NGO-run health bots (Chen et al., 2025; Longchar et al., 2025). The benchmarks that evaluate those agents are almost entirely in English (Singhal et al., 2025; Arora et al., 2025; Sellergren et al., 2025). Safety and accuracy assessments in Hindi or Marathi sit at the edge of what the field currently measures (Ghosh et al., 2025; Khullar et al., 2025), and the gap is not neutral. A response that correctly explains a pregnancy medication in English may drop the same warning in Marathi, or give the dose in a unit the reader cannot decode (Jin et al., 2024; Alonso et al., 2024; Rohera et al., 2025). Recent work documents that medical LLMs can produce unsafe recommendations and are vulnerable to generating harmful misinformation, particularly when tested outside their primary evaluation language (Moëll and Sand Aronsson, 2025; Han et al., 2024).

This paper introduces Sakhi, a community-validated benchmark for maternal and reproductive health LLMs in English, Hindi, and Marathi. We built it with Accredited Social Health Activist (ASHA) workers and doctors in rural India, anchored in a bot already deployed to that population, and we release it with two reference-answer tracks that a benchmark of this kind has not previously combined: a doctor-edited expert arm of 149 question-answer pairs and a community-sourced non-expert arm of 231 pairs, so that the MQS gap between the two is directly measurable within the same model. Nearby work touches pieces of this problem and stops short of the design we need. Khullar et al. (2025) evaluate triage classification accuracy on Indian-language maternal-health conversations, and Bhanusree et al. (2026) compare LLM responses to pregnancy queries in Telugu, but neither provides a domain-specific clinical rubric evaluation of open-ended generation across three

languages.

Our evaluation covers 13 instruction-tuned LLMs drawn from a pre-declared eligibility frame and stratified by access mode, training specialization, and capability tier. Each model produces three independent responses per question per language under identical prompts, for just under 45,000 total generations, scored by GPT-4o-mini against a five-axis, fifteen-criterion clinical rubric (Section 4) and calibrated on a stratified 500-response subset with two further judges (Claude Opus 4.6 and GPT-5.1). The headline numbers in Section 5 support three claims that bear directly on deployment. No model exceeds MQS 0.32 on the doctor-edited arm in English, and the stricter GPT-5.1 judge drops every model a further 0.17 to 0.20. Terminology Accessibility is the weakest axis for every model, lower than Context Awareness in every case, so the dominant failure mode is lexical rather than broadly cultural. Open-weight 27B models (MedGemma 27B, Gemma 3 27B) reach the same MQS as proprietary flagships at roughly one-twentieth the per-response cost, and medical fine-tuning alone does not lift rubric compliance in this cross-cultural setting.

Contributions.

1. Sakhi, a parallel English, Hindi, and Marathi maternal and reproductive health benchmark with two reference-answer tracks (doctor-edited and community-sourced), validated through a purpose-built Q&A review platform across three stakeholder channels: practicing medical doctors for clinical accuracy, Accredited Social Health Activist (ASHA) frontline community health workers for sociolinguistic grounding in rural India, and healthcare nonprofit staff for local cultural and infrastructural expertise.
2. A stratified 13-model evaluation with within-family flagship/efficient pairs and a base-versus-medical-finetune matched comparator, scored under a five-axis clinical rubric and a three-judge calibration that quantifies judge-model bias.
3. Axis-level and cost-level findings that change the recommended deployment choice for NGO-scale maternal-health bots: prioritize lexical and terminology alignment over medical fine-tuning, and open-weight 27B models over proprietary flagships.

2 Related Works

Automated Evaluation Metrics Early work on automatic evaluation relied on surface overlap methods such as BLEU (Papineni et al., 2002), which measures n-gram precision against reference translations with a brevity penalty to discourage overly short outputs, and ROUGE (Lin, 2004), which provides recall-oriented metrics including n-gram overlap (ROUGE-N) and longest common subsequence (ROUGE-L). These metrics are reproducible and easy to compute, yet they correlate poorly with human judgments (Callison-Burch et al., 2006) as their reliance on exact lexical overlap means they may fail to fully capture semantic adequacy or fluency in cases of valid paraphrasing or synonym substitution, a limitation acknowledged even in BLEU's foundational paper (Papineni et al., 2002). Such weaknesses are particularly visible in low-resource and morphologically rich languages (Sai et al., 2022; Datta et al., 2022). To improve reproducibility, BLEU score reporting guidelines and tools like SacreBLEU were introduced (Post, 2018). The community also adopted character-based metrics such as chrF (Popović, 2015) and chrF++ (Popović, 2017), which use character n-gram F-scores and are more robust to orthographic variation, better suited for morphologically rich languages, along with METEOR (Banerjee and Lavie, 2005), which improves alignment with human preferences by combining precision and recall, using flexible matching via stemming and synonym databases (like WordNet), and adding a fragmentation penalty to account for word order. Even when combined in metric stacks (BLEU/ROUGE/chrF), these methods have been shown in multiple studies (Callison-Burch et al., 2006; Sai et al., 2022; Kocmi et al., 2023) to miss semantic and factual distortions, a limitation that becomes critical in clinical evaluation, where traditional metrics are known to diverge from clinical judgment and fail to assess semantic adequacy or factuality (Tam et al., 2024; Croxford et al., 2025; Abbasian et al., 2024).

To address the limitations of the aforementioned metrics, recent work adopts semantic embedding–based approaches that capture contextual meaning beyond lexical overlap. Sentence-BERT (SBERT) (Reimers and Gurevych, 2019) introduced a Siamese network structure to generate fixed-size sentence embeddings optimized for semantic similarity, while SimCSE (Gao et al.,

2021) refined this idea through contrastive learning, producing robust representations. Similarly, contextual methods such as BERTScore (Hanna and Bojar, 2021) and MoverScore (Zhao et al., 2019) capture word-level and semantic alignment. These methods consistently outperform n-gram metrics for semantic comparisons (Sai et al., 2022; Datta et al., 2022). These approaches have been adapted for specific linguistic contexts; for instance, L3Cube-IndicSBERT (Deode et al., 2023) improves cross-lingual alignment for Hindi and Marathi, and benchmarks like IndicGenBench (Singh et al., 2024) support systematic evaluation of Indic text. Other studies assess translation quality for Indic languages using prompting and fine-tuning strategies (Nair et al., 2024), while XNLI benchmark tasks remain a standard for multilingual sentence evaluation (Conneau et al., 2018). Despite their advantages, embedding-based measures alone can overlook clinically important semantic drift (Tam et al., 2024; Chen et al., 2025) and are insufficient to ensure clinical safety, positioning them as one component of a multi-pronged evaluation anchored in explicit clinical validation.

LLM-as-an-Evaluator Approaches A third approach, which aims to scale up evaluation while mitigating the cost of human judges, involves using LLM-as-a-Judge frameworks to augment and validate other automated metrics, particularly for assessing safety and factual correctness. Benchmarks like OpenAI's HealthBench (Arora et al., 2025) establish this foundation with physician-designed, multi-dimensional rubrics. This has spurred the development of LLM-based evaluators such as G-Eval (Liu et al., 2023) and LLM-Rubric (Hashemi et al., 2024), which calibrate judgments against human ratings to improve reliability. LLM judges, however, should not be treated as standalone assessors. They work best when cross-validated against human or rubric-based evaluations. Empirical studies provide evidence that LLM judges can exhibit biases, including self-preference where models favor their own outputs (Wataoka et al., 2025; Stureborg et al., 2024; Wang et al., 2024), and a general "agreeableness" bias, and even sycophancy (Malmqvist, 2024), which may lead to increased false-positive rates in certain evaluation contexts (Zhou et al., 2024). LLM judges also miss nuanced factual errors, which makes them less reliable for specialist correctness without oversight (Szymanski et al., 2025). In high-stakes healthcare settings, these limitations translate into safety and equity risks, and some studies find demographic biases in evaluation outcomes themselves (Ye et al., 2024; Ji et al., 2025). Recent work now recommends anchoring LLM-judge outputs in deterministic, clinician-weighted rubrics and cross-validating them with other metrics and human review, a hybrid approach that matters most in sensitive contexts like maternal health (Chen et al., 2025; Szymanski et al., 2025).

Human Evaluation and Methodological Rigor Beyond automated metrics, evaluation research is moving past generic benchmarks toward domain-specific frameworks and tighter methodological standards. Recent examples include new proposals for conversational medical AI (Abbasian et al., 2024), structured clinical documentation assessment (Seo et al., 2024), and taxonomies for medical LLM performance (Lacerda et al., 2025). Alongside these proposals, researchers are adopting more standardized human evaluation with detailed rubrics that define clear criteria such as accuracy, coherence, and safety, rather than relying on subjective ratings (Bojic et al., 2023), and foundational guides on designing reliable user studies in NLP (Schuff et al., 2023) now inform the practice. This tighter rigor does not make human evaluation a clean gold standard. Clark et al. (Clark et al., 2021) demonstrate that human evaluation is often not fully reliable and is prone to sampling bias and low inter-annotator agreement. LLMs used as automated annotators are also vulnerable to adversarial prompt injection that can manipulate them into producing poisoned datasets and compromised findings (Baumann et al., 2025). The field now recommends using human studies to validate and contextualize automated approaches, not to replace them, with both requiring careful design to mitigate their respective limitations.

Domain-Specific Models and Evaluation Parallel to methodological advances in evaluation (Sai et al., 2022), model development has diverged into two main trajectories: general-purpose reasoning engines and domain-specialized biomedical systems. Generalist models such as the GPT-4 family, Gemini, and the multilingual Aya Expanse (Dang et al., 2024) demonstrate strong linguistic fluency and broad reasoning capabilities, yet their reliability in clinical contexts remains inconsistent, a gap that domain-specific models are explicitly designed to address (Sellergren et al., 2025; Chen

et al., 2023). By contrast, domain-adapted systems such as Meditron-70B (Chen et al., 2023) and MedGemma (Sellergren et al., 2025) are fine-tuned on biomedical corpora and clinical guidelines, often achieving superior performance on medical QA benchmarks when correctness and safety are prioritized. At the same time, comparative analyses caution that domain adaptation does not guarantee uniform improvements; robust gains depend on task-specific fine-tuning strategies (Anisuzzaman et al., 2025; Savage et al., 2025) and on careful, context-sensitive evaluation (Sai et al., 2022). This underscores the need for evaluation frameworks that are sensitive to linguistic and clinical variation, particularly in multilingual healthcare applications, a challenge highlighted by the non-English limitations of models like MedGemma and the multilingual focus of emerging generalist models (Dang et al., 2024; Sellergren et al., 2025). Multilingual medical benchmarks such as Multi-OphthaLingua (Restrepo et al., 2025) have revealed substantial cross-lingual bias in ophthalmological QA, while L2M3 (Gangavarapu, 2024) demonstrates the potential of translation-augmented LLMs for health equity in low-resource settings.

We construct a hybrid method built on methodical cross-validation with human evaluations. Our evaluation leverages automated metrics (e.g., chrF, BERTScore) (Popović, 2015; Zhao et al., 2019; Banerjee and Lavie, 2005) for reproducible, large-scale benchmarking, while acknowledging their inability to assess clinical safety by themselves (Tam et al., 2024; Croxford et al., 2025; Chen et al., 2025). Expert-grounded rubrics inspired by frameworks like HealthBench (Arora et al., 2025) are employed to encode medical correctness. LLM-judge outputs are employed but are anchored by human-expert validation (Chen et al., 2025; Szymanski et al., 2025), in line with best practices for safely deploying models in sensitive, multilingual contexts (Dang et al., 2024; Sellergren et al., 2025). While recent benchmarks such as CLINIC (Ghosh et al., 2025) evaluate multilingual trustworthiness across 15 languages and Script Gap (Khullar et al., 2025) assess triage accuracy on Indian-language maternal health conversations, neither provides domain-specific clinical rubrics for evaluating the quality of generated maternal health advice, which is the gap our framework closes.

3 Dataset

Multilingual datasets of maternal health queries from rural and semi-urban populations remain scarce (Singhal et al., 2025; Antoniak et al., 2024), and the few that exist typically mix synthetic generation with automated validation alone. Sakhi separates the two. We release the benchmark as 380 question-answer pairs across two tracks in parallel English, Hindi, and Marathi (1,140 query-language pairs total), drawn from a larger generation pool of 822 pairs through deduplication, medical-expert review, and, for the stricter track, full doctor rewriting.

Generation Pipeline. We compiled a knowledge corpus from trusted maternal health guidelines: the WHO Recommendations on Maternal Health, India's National Antenatal Care Guideline (noa), the Auxiliary Nurse and Midwife Training Manual, and National Health Mission protocols. These documents were segmented into over 3,000 semantic chunks following retrieval-augmented generation (RAG) practices for specialized healthcare domains (Wu et al., 2025; Amugongo et al., 2025). A two-stage generator-validator produced the initial question-answer pool: Aya Expanse (Dang et al., 2024) drafted question-answer pairs from the chunks, and MedGemma (Sellergren et al., 2025) validated each pair for medical accuracy, safety, and contextual relevance.

A Purpose-Built Review System for Three Stakeholder Groups. AI-generated medical content needs human verification before it can serve as a benchmark reference, especially in high-stakes clinical domains (Amugongo et al., 2025; Fraile Navarro et al., 2025). Off-the-shelf annotation tools did not fit the workflow we needed, so we built a dedicated Q&A review platform that routes each generated pair through three distinct stakeholder groups, each contributing the kind of judgment they are actually qualified to make.

The first channel is **practicing medical doctors** who provide clinical validation. Doctors receive a pair, a review rubric (accuracy, safety, completeness of contraindications and warnings), and an editor to rewrite the reference answer if they find it insufficient. The pairs they rewrite in this channel form the **expert arm** of 149 doctor-edited question-answer pairs. The second channel is **Accredited Social Health Activist (ASHA) workers**, frontline community health workers in rural India who are

the daily point-of-care contact for most of our target users. ASHA workers provide the benchmark's sociolinguistic grounding: they flag questions that do not sound like a patient would ask them, flag reference answers that use clinic-only vocabulary a rural user cannot parse, and flag omissions around locally-available health infrastructure (government schemes, anganwadi centres, district health posts). The third channel is **healthcare nonprofit staff** with deep local expertise in cultural norms, family dynamics, and the specific trust architecture through which health information actually reaches women in rural India. This group caught the subtle failures that clinical review alone misses, such as advice that is medically correct but culturally implausible for a daughter-in-law to act on without her mother-in-law's approval, or for a first-time mother to raise without her husband present.

Following the gold standard for clinical NLP validation (Tam et al., 2024; Bojic et al., 2023), the three review channels feed a two-tier pipeline that produces the two released tracks. In the first tier, the medical-expert channel and the nonprofit-staff channel review the machine-generated English pairs for clinical accuracy and sociocultural appropriateness; pairs that pass both form the **non-expert arm** of 231 community-sourced pairs with doctor-reviewed reference answers. In the second tier, validated pairs are routed to practicing doctors and medical students who write their own answer versions; the 149 rewritten pairs form the **expert arm**. Both arms are then professionally translated into Hindi and Marathi by native speakers, and the ASHA worker channel reviews the translated pairs for patient-voice fidelity before release. Each review action is logged in the platform with reviewer role, timestamp, and edit diff, so that the provenance of every released reference answer is recoverable and auditable.

What the two tracks enable. The expert and non-expert arms are not redundant. They let us measure, within the same model, how much MQS drops when the reference answer gets stricter: the 0.10 to 0.13 MQS expert-minus-non-expert gap reported in Section 5.2. Benchmarks that report on a single reference-quality level cannot surface this gap. We release both arms as `sakhi_benchmark_expert.csv` and `sakhi_benchmark_non_expert.csv` along with the rubric, judge prompts, and judge-calibration records.

4 Methodology

Evaluation Framework and Problem Setting Our work introduces a two-stage evaluation framework designed to address limitations of LLM benchmarks for medical applications. The framework emphasizes clinical validity, linguistic fidelity, and contextual appropriateness, focusing on maternal health scenarios across English, Hindi, and Marathi. Maternal health is a high-stakes domain where misinformation has direct consequences for patient safety. In multilingual settings, linguistic nuance and cultural appropriateness also shape whether a patient understands and follows the advice. Our framework evaluates model-generated responses along three complementary dimensions:

1. Semantic fidelity, which measures the preservation of clinical meaning;

2. Linguistic quality, assessing fluency and coherence;

3. Expert-based clinical validity, which evaluates the factual and ethical soundness.

The evaluation dataset (Section 5 and Section 3) comprises 380 patient-style English queries split across two reference-quality tracks, 149 doctor-edited (expert arm) and 231 community-sourced (non-expert arm), spanning ten maternal health themes. Each English query was professionally translated into Hindi and Marathi, yielding parallel evaluation sets of 380 questions per language (1,140 query-language pairs in total). The evaluation proceeds sequentially: automated benchmarking establishes reproducible linguistic and semantic baselines (Layer 1), followed by structured rubric review (Layer 2). The two layers are then synthesized into multi-dimensional indicators of model reliability.

Models Evaluated The evaluation covers 13 instruction-tuned LLMs drawn from a pre-declared eligibility frame and stratified along three axes: access mode (proprietary vs. open-weight), training specialization (general-purpose, multilingual-specialized, medical-specialized), and capability tier (flagship vs. efficient). The stratification, the full panel composition, and the matched base-versus-medical-finetune comparator (Gemma 3 27B alongside MedGemma 27B) are described in Section 5.1. Model selection prioritized public API or open-weight availability, documented

multilingual capabilities, and relevance to healthcare dialogue. Models were accessed through OpenRouter, the local Claude Code CLI (Claude Opus 4.7), and HuggingFace Inference Endpoints (MedGemma variants), with identical prompts and per-family generation parameters documented in Appendix B.4. Each model generated three independent responses per (question, language) under no multi-turn context, yielding up to $13 \times 3 \times 3 \times 380 = 44{,}460$ total model outputs before coverage gaps are applied (see Section 5.1 and the Limitations).

Layer 1: Automated Benchmarking This layer enables scalable, language-agnostic evaluation of two core properties: semantic fidelity and linguistic quality. Automated scoring provides standardized quantitative baselines across models and languages prior to clinical review.

Semantic Fidelity Assessment Preserving clinical meaning is central to the safety of automated responses. To evaluate semantic fidelity, two complementary similarity metrics are employed. Embedding-based alignment measures overall conceptual correspondence using cosine similarity between dense sentence embeddings generated by OpenAI's text-embedding-3-small model within a shared multilingual space. While additional embedding models (Cohere, Voyage, and alternative architectures) were evaluated during development, text-embedding-3-small was selected as the primary metric for its consistent performance across all three languages; comparative results from alternative embedding models are provided in Appendix A. BERTScore, computed with xlm-roberta-base for multilingual text and roberta-base for English, captures token-level semantic overlap and contextual entailment across Indic and Latin scripts. All configurations, model versions, and scripts are documented for reproducibility. To mitigate single-reference bias, sampled outputs were manually inspected for semantic entailment and corrected when necessary.

Linguistic Quality Assessment Linguistic quality is evaluated through complementary metrics capturing fluency, coherence, and readability. SacreBLEU assesses n-gram precision; chrF++ provides robustness to morphological variation; METEOR accounts for synonymy and paraphrasing; and ROUGE-L quantifies structural coherence through sequence overlap. Each metric provides a normalized score, with scores reported individually to preserve diagnostic granularity across evaluation dimensions.

For English outputs, we also compute perplexity as an intrinsic measure of fluency (Jelinek et al., 2005; Meister and Cotterell, 2021). Perplexity was not computed for Hindi and Marathi due to the limited availability of validated language models for reliable perplexity estimation in these languages at the time of evaluation. Scores are normalized per response and averaged across languages. To validate the reliability of the metric, a small-scale human evaluation was conducted. Correlations between human ratings and automated scores demonstrated consistent directional alignment, supporting their interpretive adequacy.

Layer 2: Expert-Based Clinical Validation Automated metrics cannot fully capture clinical accuracy or the safety implications of a response. The second layer introduces structured rubric evaluation, grounded in a standardized criterion set validated by domain experts.

Thematic Rubric Development and Validation A panel of clinical experts (see Appendix A) designed and validated the evaluation rubrics. Through iterative content analysis, we identified ten core maternal health themes:

- Antenatal & Maternal Health Care
- Nutrition, Diet & Supplementation
- Mental, Emotional & Social Well-being
- Clinical Procedures & Guidelines
- Medication & Vaccination Safety
- Reproductive & Sexual Health (Beyond Pregnancy)
- Health Systems, Access & Provider Support
- Infection Prevention & Hygiene Practices
- Symptom Interpretation & Danger Sign Recognition
- Risk & Complication Management

Each theme was assigned a uniform 5×3 rubric structure comprising 15 binary criteria organized under five axes: Accuracy, Completeness, Context Awareness, Communication, and Terminology Accessibility.

Axis weights were determined through expert consultation: Accuracy = 0.30, Completeness = 0.25, Context Awareness = 0.20, Communication = 0.15, and Terminology Accessibility = 0.10.

Example (Antenatal & Maternal Health Care theme): under the Accuracy axis, criteria include:

- Correctly emphasizing the importance and recommended schedule of regular antenatal check-ups.
- Accurate identification of key assessments to be conducted during pregnancy.
- Recognition of the roles of healthcare providers in maternal care.

The complete rubric framework and thematic codebook appear in Appendix A.

The Medical Quality Score (MQS) represents a single unified metric applied consistently across all themes. Each theme utilizes the same five-axis rubric structure with identical weighting, ensuring comparability across maternal health topics. Theme-specific rubrics differ only in the concrete criteria under each axis, tailored to the clinical requirements of each domain.

Thematic Classification Process Each question was automatically assigned to a maternal health theme using a dspy-based classifier implementing few-shot structured prompting. DSPy (Differentiable Structured Prompting Yield optimizer) refines prompt templates to enhance thematic consistency across languages (Khattab et al., 2023). Each theme was initialized with 10–15 expert-labeled seed examples to calibrate decision boundaries.

Model classifications were manually reviewed for a representative subset to ensure alignment with expert expectations, particularly where thematic overlap occurred (e.g., Symptom Interpretation vs. Risk Management). Low-confidence predictions were reviewed by clinicians before final assignment. This hybrid process combined efficiency with expert oversight, ensuring the integrity of theme-level labeling.

Automated Rubric Scoring To enable evaluation at scale, rubric scoring is performed by an LLM judge (GPT-4o-mini) that receives each question, its reference answer, the model's response, and the 15 rubric criteria for that theme, and returns binary pass/fail judgments as structured JSON. The judge runs with deterministic parameters (temperature = 0) for reproducibility. The exact prompt is given in Appendix B.3; we note in the Limitations that the prompt lists the rubric as predicates over a labeled Question/Reference/Response triplet rather than explicitly instructing the judge to compare response against reference, and we test how much that structural choice matters through cross-judge calibration below.

Cross-Judge Calibration Because a single judge can bias the absolute MQS (Arora et al., 2025), we re-score a stratified 500-response calibration subset (balanced across 13 models, 2 dataset arms, 3 languages, and 10 themes) with two further judges, Claude Opus 4.6 and GPT-5.1, under the same prompt (Section 5.7). Reporting the pairwise Pearson correlations and mean signed bias between judges lets readers see how much of the reported MQS is judge-calibrated rather than intrinsic to the model. A comparable human-expert scoring pass on this subset is bounded by reviewer capacity and is reported as a planned extension in the Limitations; the human validation that was feasible at this stage is concentrated in the upstream creation of reference answers (Section 3), not in per-response scoring of the 44,460 model outputs.

Score Synthesis and Analytical Approach The two-stage framework produces a multi-dimensional evaluation matrix integrating automated and expert-derived metrics. For each response, five axis-level pass ratios (A_i) are computed. The overall Medical Quality Score (MQS) is defined as:

$$\text{MQS} = \sum(w_i \times A_i)$$

where weights w_i correspond to expert-defined importance factors: Accuracy = 0.30; Completeness = 0.25; Context Awareness = 0.20; Communication = 0.15; Terminology Accessibility = 0.10, with $\sum w_i = 1$.

The framework produces separate metrics for semantic fidelity (embedding similarity and BERTScore), linguistic quality (SacreBLEU, chrF++, METEOR, ROUGE-L, and perplexity for English), and clinical validity (MQS). These dimensions are reported independently rather than combined into a single composite score, as each captures distinct aspects of model performance critical to different deployment considerations. Aggregate interpretations consider performance across all three dimensions jointly.

Per-theme and per-language analyses expose asymmetries such as high fluency combined with low medical correctness. Together, the two layers give a transparent, interpretable picture of LLM reliability in maternal healthcare and support the development of safer, context-aware clinical dialogue systems.

Reproducibility and Transparency We release the full pipeline alongside the paper. The supplementary code repository contains the prompt templates, API configurations (temperature, top_p, max_tokens), random seeds, preprocessing scripts, rubric materials, thematic classification examples, and expert evaluation protocols. Generation calls used identical API parameters within each model family (reasoning vs. standard; Appendix B.4), and every generation and judge call is logged to a timestamped JSON Lines file so that individual scores can be re-audited without re-running the API. Computational infrastructure details are in Appendix B.4.

5 Experiments and Results

This section presents the concrete experimental instantiation of the two-layer framework in Section 4 and the findings it produced. The main body carries one headline table, four figures, and the judge-calibration table; the full per-language, per-theme, per-axis, per-cost, and per-validity tables are relocated to Appendix A so that the narrative stays readable and the backing numbers remain available for review.

5.1 Experimental Setup

Stratified model panel. Rather than an ad-hoc collection of currently-popular LLMs, we fix a reproducible sampling frame and draw a stratified panel from it. The eligible frame, as of the snapshot date **2026-04-01**, is any instruction-tuned LLM that (i) offers a public inference API or open weights, (ii) documents generation in English and at least one Indic script, (iii) has been released or substantively updated within the prior 24 months, and (iv) is not a base language model. We stratify the frame along three axes shown in Section 4 to explain variance in the outcomes of interest: *access mode* (proprietary API vs. open-weight), *training specialization* (general-purpose, multilingual-specialized, medical-specialized), and *capability tier* (flagship vs. efficient). From each populated cell we include the provider's own publicly-positioned flagship, plus, where a matched-generation efficient sibling exists in the same family, the efficient variant as well. This rule yields a 13-model panel spanning five providers, with within-family flagship/efficient pairs for OpenAI (GPT-5 Mini / GPT-4o Mini), Google (Gemini 3 Pro / Gemini 3 Flash / Gemini 3.1 Flash-Lite), Anthropic (Claude Opus 4.7 / Claude Haiku 4.5), Meta (Llama 4 Maverick / Llama 3.3 70B), and Google DeepMind (MedGemma 27B / MedGemma 4B), together with Cohere's Command A and Aya Expanse 32B. We additionally include Gemma 3 27B as a controlled base-model comparator for MedGemma to allow attribution of medical fine-tuning gains. Empty cells (no open-weight proprietary frontier; no proprietary medical-specialized) are reported rather than hidden.

Two-track dataset. All 13 models are evaluated on two disjoint question sets.

- **Expert (149):** doctor-validated pairs with professionally-translated gold answers in all three languages, released as `sakhi_benchmark_expert.csv`. This is the primary track for clinical-quality claims.
- **Non-expert (231):** community-sourced pairs with doctor-reviewed reference answers, released as `sakhi_benchmark_non_expert.csv`. This track provides broader thematic coverage and supports robustness checks.

The two tracks are reported separately throughout; no score is pooled across them. A response is generated independently for each of three runs per (model, language, question), following Section 4. Total evaluated model outputs are $13 \times 3 \times 3 \times (149 + 231) = 44{,}460$, minus the open-weight MedGemma variants that the OpenRouter endpoint used to dispatch the other models does not host. MedGemma is evaluated on the non-expert track only and flagged accordingly in every table.

Judge panel and calibration. Our primary rubric judge is **GPT-4o-mini**, which scores every response in both dataset arms under the zero-shot rubric prompt in Appendix B.3. To test how much the findings depend on the judge model, a methodological concern that prior work flagged with a 0.118-point MQS gap between judges on the same responses, we additionally score a **stratified 500-response calibration subset** (balanced across 13 models, 2 datasets, 3 languages, and 10 themes) with two further judges: **Claude Opus 4.6** (via local Claude Code, zero API cost) and **GPT-5.1**

(via OpenRouter). The same zero-shot prompt is used by all three judges. Judge-agreement statistics on this subset quantify how much reported MQS depends on the choice of evaluator.

Generation configuration. Proprietary models are accessed via OpenRouter with `temperature`=0.7 and `max_tokens`=400, except reasoning-capable models (GPT-5 family, Gemini 3 family, o-series) which use `temperature`=1.0, `max_tokens`=2000, and OpenRouter's `reasoning={effort=low, exclude=true}` to prevent reasoning-token exhaustion while keeping reasoning content out of the scored response. Claude Opus 4.7 is invoked via the local `claude -p` CLI. All generation and judge calls are appended to timestamped per-(model, dataset, lang) JSON Lines files; any interruption resumes by skipping already-completed `(run, q_idx)` keys without re-spending API tokens. Full configuration, API parameters, and reproducibility scripts live in the supplementary repository.

5.2 Overall Cross-Model Performance

Table 1 reports MQS for each of the 13 models across the two dataset arms and three languages, computed as the mean over three runs times all questions for which the model produced a valid response. *MQS is systematically lower on the expert arm than on the non-expert arm* for every model and every language where both arms are covered. The gap is 0.110 to 0.137 on English cells, 0.079 to 0.105 on Hindi, and 0.063 to 0.109 on Marathi, with an across-cell median of 0.103. The doctor-edited expert references contain more specific clinical content, which the 15 binary rubric criteria correspondingly enforce more strictly. Collapsing the two arms into a single score would obscure this gap; we keep them separate throughout.

Under the expert arm (English), Claude Haiku 4.5 and Claude Opus 4.7 lead with 0.305 and 0.297 respectively, closely followed by GPT-5 Mini at 0.298. *No model exceeds 0.32 on expert-EN; fewer than one in three rubric criteria are reliably met.* Under the non-expert arm (English), the ordering is similar: Claude Haiku 4.5 (0.422), GPT-5 Mini (0.418), Claude Opus 4.7 (partial coverage; see below), Gemini 3 Flash (0.396), and the H2-baseline Gemma 3 27B (0.395) cluster tightly. Llama 3.3 70B and GPT-4o Mini are the weakest non-specialist proprietary / open-weight pair. Medical-specialized MedGemma 27B (0.394) is indistinguishable from its base Gemma 3 27B (0.395) on non-expert EN, so medical fine-tuning alone does not lift rubric compliance in this cross-cultural setting. We call this the **H2-baseline null** and the Discussion revisits it.

Figure 1 shows the same non-expert EN ranking with per-model run-to-run error bars computed by grouping each cell's judge outputs by run index, averaging within each of the three runs, then taking the standard deviation across the three per-run means. Error bars are small (at most ~ 0.015 MQS), so rankings are stable across reruns even where the judge-vs-judge disagreement is large (Section 5.7).

Coverage gaps. Claude Opus 4.7 is the model with the largest coverage footprint in this release. The expert arm has full coverage in English and Hindi but only one complete run for Marathi (n=108 usable responses of the 447 expected), and the non-expert arm has partial coverage only on Marathi (n=113), with non-expert English and non-expert Hindi missing entirely due to sustained Claude Code CLI session limits during the dispatch window. We exclude partial and missing Claude Opus 4.7 cells from the top-3 bolding in Table 1 and show the exact coverage in Appendix Table 7. Aya Expanse 32B was not dispatched to the expert arm because its native Cohere-API route requires credentials we did not have; it is reported on the non-expert arm only. MedGemma 4B and 27B are reported on the non-expert arm only because OpenRouter does not host them and the HuggingFace Inference Endpoint route we set up was scoped to the non-expert track in this release. Gemini 3.1 Flash-Lite is retained in Appendix Table 7 to show that it failed to produce valid output in every cell; it is excluded from scored comparisons elsewhere.

5.3 Theme-Level and Cross-Lingual Structure

Figure 2 summarises the MQS surface across the 10 maternal-health themes and all 13 models on the deployment-facing non-expert arm, averaged over the three languages. Two patterns are visible without having to look at the numbers. First, themes sort left-to-right from strongest (Nutrition, Symptoms) to weakest (Clinical Procedures, Risk Management), and the ordering is consistent across models; every model in the panel is strongest on Nutrition and weakest on Risk Management. Second, models sort top-to-bottom into a tight cluster: the flagship proprietary pair (Claude Haiku 4.5,

Table 1: Medical Quality Score (MQS) by model, dataset, and language under the GPT-4o-mini judge. Cells are the mean across three runs and all questions for which the model produced a valid response. **Bold** marks the best score per column; underline marks the second best. A midrule separates the 13-model evaluated panel from the baseline Gemma 3 27B, which is reported only to anchor the medical-finetune comparison against MedGemma 27B.

Model	Expert (149)			Non-expert (231)		
	EN	HI	MR	EN	HI	MR
GPT-5 Mini	0.298	0.325	0.310	0.418	0.419	**0.419**
GPT-4o Mini	0.241	0.270	0.250	0.376	0.375	0.359
Command A	0.256	0.302	0.282	0.393	0.381	0.351
Gemini 3 Pro	0.255	0.296	0.289	0.387	0.386	0.382
Gemini 3 Flash	0.266	0.307	0.321	0.396	0.408	0.404
Claude Opus 4.7	0.297	**0.354**	**0.353**	–	–	0.363
Claude Haiku 4.5	**0.305**	0.338	0.343	**0.422**	**0.434**	0.406
Llama 3.3 70B	0.233	0.268	0.261	0.369	0.363	0.349
Llama 4 Maverick	0.254	0.280	0.279	0.364	0.376	0.369
Aya Expanse 32B	–	–	–	0.382	0.369	0.337
MedGemma 27B	–	–	–	0.394	0.392	0.383
MedGemma 4B	–	–	–	0.326	0.328	0.331
Gemma 3 27B (baseline)	0.266	0.292	0.281	0.395	0.396	0.388

GPT-5 Mini) and Gemini 3 Flash are visibly lighter than Llama 3.3 70B and MedGemma 4B, but the separation is small relative to the theme-level separation. The within-model row is flatter than the within-theme column, meaning theme-level difficulty explains more variance than model choice.

The cross-lingual story is more subtle and hinges on judge calibration. For 11 of 13 models, Hindi MQS equals or exceeds English MQS under the GPT-4o-mini judge, with ΔHI ranging from -0.013 (Aya Expanse) to $+0.058$ (Claude Opus 4.7) and a median of $+0.014$. Marathi shows a similar pattern, with ΔMR in $[-0.045, +0.061]$ and a median of $+0.005$. We read this result carefully: it does *not* mean LLMs are better at maternal-health reasoning in Hindi or Marathi. The judge-agreement analysis in Section 5.7 shows that GPT-4o-mini is systematically more lenient on non-English outputs, and GPT-5.1 as a stricter judge collapses the EN/HI/MR gap substantially. The conservative interpretation is that apparent cross-lingual parity under a mid-tier judge masks judge leniency toward Devanagari-script responses. Full per-model ΔHI and ΔMR values are in Appendix A.1 (Table 3).

5.4 Axis-Level Diagnostics

Figure 3 reports the per-axis pass rate on the non-expert EN arm. Three patterns dominate.

Communication is uniformly high. Every model scores above 0.58 on Communication (GPT-5 Mini 0.700, Gemini 3 Flash and Gemma 3 27B 0.688, Claude Haiku 4.5 0.689), confirming that frontier LLMs reliably adopt warm, supportive phrasing and recommend professional consultation.

Terminology Accessibility is the lowest axis for every model on the non-expert EN arm (0.123–0.206), consistently below Context Awareness in this slice. The same pattern does not hold uniformly on the expert arm, where Completeness is the lowest axis for Claude Opus 4.7 expert-EN and a handful of other (model, language) cells (Appendix A.3). The criteria under Terminology Accessibility cover defining medical abbreviations (ANC, BP, FHR), translating micronutrient and medication names into the local language, and avoiding jargon at first use. Most rubric items in this axis are failed in the non-expert EN slice that the deployed bot most directly reflects. A response can be correct, empathetic, and culturally appropriate while still using the untranslated abbreviation "FHR" that a rural user cannot decode.

Context Awareness is the second-lowest axis (0.226–0.264), failing criteria about local cultural beliefs, resource constraints, and community-health infrastructure (ASHA workers, anganwadi centres, government health schemes). Combined with the Terminology failures, this axis pair characterises the benchmark's central finding on the non-expert EN arm: models produce responses that are clinically correct and communicatively warm but *linguistically inaccessible* and *infrastructurally generic*, fitting Western-clinic assumptions rather than rural Indian deployment contexts.

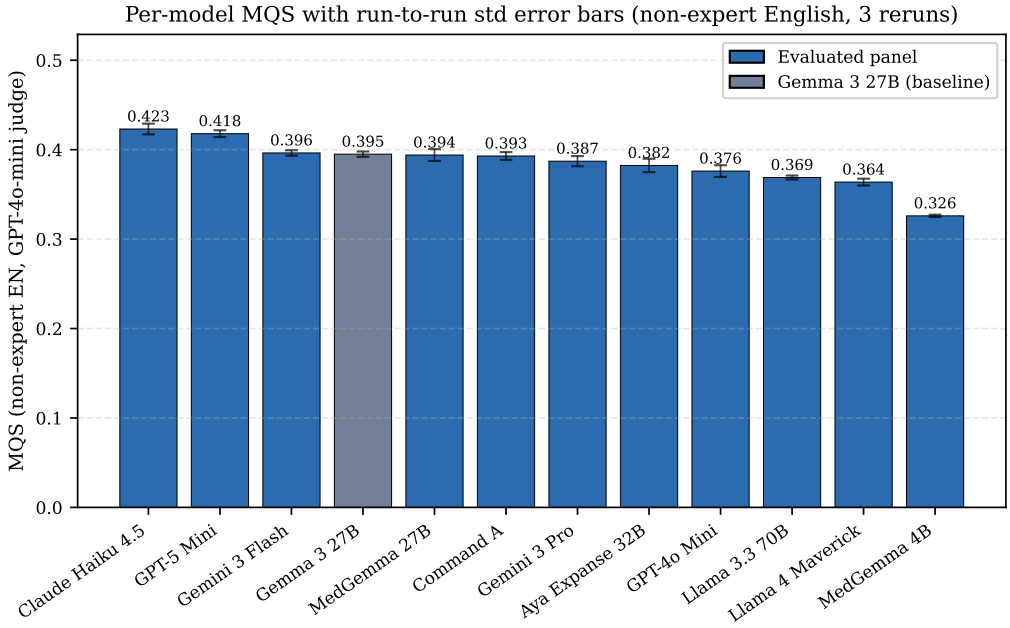

Figure 1: Per-model MQS on the non-expert EN arm with run-to-run std error bars. Means are computed across the 3 reruns $\times$ 231 questions; error bars are the std across the three per-run means. Gemma 3 27B is shown in grey as the matched base-model baseline.

5.5 Cost-Performance Trade-Offs

Figure 4 plots per-response generation cost against non-expert EN MQS, on a log-scaled cost axis. The picture is cleaner than a table: MedGemma 27B and Gemma 3 27B sit at roughly \$0.00007 to \$0.00008 per response with MQS near 0.395, while Claude Haiku 4.5 and Gemini 3 Pro sit at \$0.0015 to \$0.0019 per response with MQS in the 0.387 to 0.422 range. The frontier flattens above \$0.0002 per response: paying 20$\times$ more moves MQS by at most 0.03. Claude Opus 4.7 is accessed through the local Claude Code subscription at zero marginal per-response cost; it is a research-evaluation convenience, not a fair commercial point, so we exclude it from the cost axis.

The deployment consequence is concrete: for an NGO running a maternal-health bot at million-queries-per-month scale, switching from Claude Haiku 4.5 to Gemma 3 27B saves roughly \$17,000 per month for a -0.027 MQS change. Our axis-level evidence (Figure 3) suggests that amount would produce more safety gain if spent on Hindi and Marathi medical-terminology alignment than on a frontier-tier model.

5.6 Generation Reliability

Not every response logged in a run constitutes a valid answer: some are empty, truncated, or return refusals. Failure rates are highest in Marathi, reflecting the familiar brittleness of LLM generation in lower-resource scripts; Llama-family models in particular show degraded Marathi coverage. Gemini 3.1 Flash-Lite fails in every cell of the panel and is excluded from scored comparisons. We report the full per-(dataset, language) valid-response rate in Appendix Table 7; deployment decisions should pair MQS with this "generation dropout" rate, since a model that scores well on valid responses is only useful to the extent that valid responses come back in the first place.

Table 2: Three-judge agreement on the stratified 500-response calibration subset ($n = 383$ rows with all three judges present). Pearson ρ is computed on MQS and **bold** marks the judge pair with highest rank agreement. Mean ΔMQS is the signed bias; underline marks the smallest absolute bias.

Pair	Pearson ρ	Mean ΔMQS
GPT-4o-mini $\leftrightarrow$ Claude Opus 4.6	**0.429**	+0.037
GPT-4o-mini $\leftrightarrow$ GPT-5.1	0.238	+0.201
Claude Opus 4.6 $\leftrightarrow$ GPT-5.1	0.338	+0.163

5.7 Judge-Agreement Calibration

Our primary GPT-4o-mini scores are validated on a stratified 500-response calibration subset ($n = 383$ rows with all three judges present) by two further

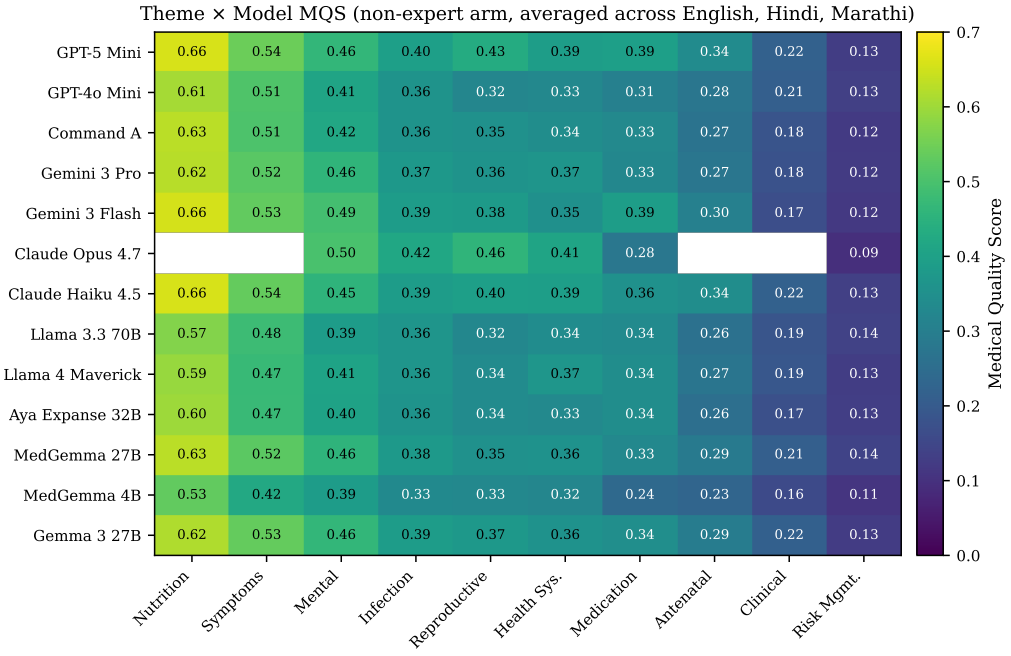

Figure 2: Theme × Model MQS (non-expert arm, averaged across English, Hindi, and Marathi under the GPT-4o-mini judge). Themes are sorted left-to-right by cross-model mean. Blank cells for Claude Opus 4.7 on non-expert EN / HI reflect the coverage gap described in Section 5.2. Full numeric table in Appendix A.2 (Table 4).

judges, Claude Opus 4.6 and GPT-5.1. Table 2 reports pairwise Pearson correlations on MQS and the mean signed difference (bias) between each pair.

Two effects emerge. First, **GPT-5.1 is systematically stricter** than both other judges by roughly 0.17 to 0.20 MQS on identical responses. Every model's headline score in Table 1 would shrink by that offset under the stricter evaluator. This bias is larger than the MQS gap between most pairs of models in our panel, so absolute rankings should be interpreted in light of judge choice. Second, **rank correlations are modest** ($\rho = 0.24$ GPT-4o-mini $\leftrightarrow$ GPT-5.1; $\rho = 0.43$ GPT-4o-mini $\leftrightarrow$ Claude Opus 4.6), meaning that judges disagree not only on absolute level but also on which responses are stronger. GPT-4o-mini and Claude Opus 4.6 agree most closely on bias ($\Delta = +0.037$), matching the pattern noted in prior work that mid- and frontier-tier proprietary models converge on lenient rubric interpretations, while GPT-5.1 applies a more discerning standard.

We read the absolute MQS values in Table 1 as *GPT-4o-mini-calibrated* scores. Under the GPT-5.1 judge the model ordering is broadly preserved (Claude Haiku 4.5 and GPT-5 Mini remain top) but the magnitudes compress by a known, reported offset. The judge panel is itself the methodological contribution: reporting a single judge's scores without this triangulation understates the uncertainty that deployers should account for.

5.8 Summary of Findings

The stratified 13-model evaluation produces four findings that hold across both dataset arms:

1. **Rubric-compliance is low overall.** No model exceeds MQS 0.32 on expert-EN under GPT-4o-mini, and under the stricter GPT-5.1 judge the panel-wide mean compresses by a further $\sim$ 0.17–0.20 MQS. Frontier LLMs meet fewer than one in three doctor-edited rubric criteria (§5.2, §5.7).
2. **On the deployment-facing non-expert EN arm, the dominant failure is linguistic.** Terminology Accessibility is the lowest axis for every model in this slice (pass rates 0.123–0.206), below Context Awareness (0.226–0.264). Responses are empathetic and clinically accurate but use untranslated medical shorthand and Western-clinic infrastructure assumptions. On the stricter expert arm, Completeness competes with Terminology Accessibility for the lowest axis in some cells (§5.4).
3. **Apparent cross-lingual parity is a judge artifact.** Under GPT-4o-mini, 11 of 13 models score equal or higher on Hindi than English.

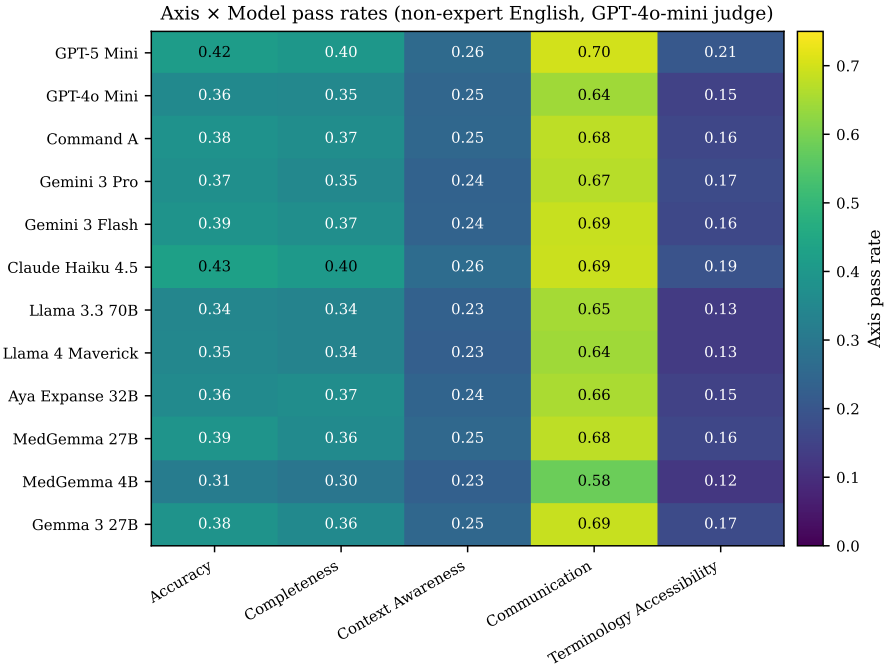

Figure 3: Axis × Model pass rates on the non-expert English arm (GPT-4o-mini judge). Each cell is the mean fraction of binary rubric criteria satisfied within that axis. Communication is uniformly high (0.58–0.70); Terminology Accessibility is uniformly low (0.12–0.21) across every model. Full numeric table in Appendix A.3 (Table 5).

Under GPT-5.1 (stricter; §5.7) the compression is not matched in the Indic direction, suggesting the parity is partly an artifact of mid-tier-judge leniency toward non-English text rather than genuine cross-lingual competence (§5.3).

4. **Open-weight medium-scale dominates the cost-performance frontier.** MedGemma 27B ($0.00007 per response, MQS 0.394) and Gemma 3 27B ($0.00008 per response, MQS 0.395) reach non-expert EN MQS matching proprietary flagships at over an order of magnitude lower per-response cost. The MQS/$ curve flattens above $0.0002 per response (§5.5).

Discussion (Section 6) revisits the implications of each finding for deployment of AI systems in rural multilingual health contexts and the methodological requirements for multi-judge clinical benchmarking.

6 Discussion

Clinically correct, linguistically inaccessible, infrastructurally generic. Across the full 13-model panel, one pattern holds in every cell: the responses are clinically accurate and warmly phrased, yet they assume a Western clinic and an English-speaking reader. No model exceeds MQS 0.32 on the expert arm in English under the GPT-4o-mini judge, and under the stricter GPT-5.1 judge (§5.7) every model loses a further 0.17 to 0.20 MQS on identical outputs. Fewer than one in three doctor-validated rubric criteria are reliably met. The axis decomposition tells us where the shortfall lives, and where it does not. Communication is uniformly high, above 0.58 for every model (§5.4), so warm tone and the habit of recommending a doctor are not the bottleneck. Accuracy and Completeness pass rates cluster in the middle of the panel, neither the failure mode nor the strong suit. Context Awareness and, more sharply, Terminology Accessibility are the weakest axes. What this looks like on the ground: a response that correctly flags pre-eclampsia danger signs but uses the untranslated shorthand BP, never gives the Hindi or Marathi word for blood pressure, and tells the reader to visit a "healthcare provider" without mentioning the ASHA worker who is, for most of our target users, the nearest healthcare contact at all.

The dominant failure is linguistic, not just cultural. Our earlier two-model analysis, and most prior work in this space, described the shortfall as a "cultural awareness" gap. The 13-model axis decomposition locates it more precisely, and the location is actionable. Terminology Accessibility pass rates sit between 0.123 and 0.206, lower than

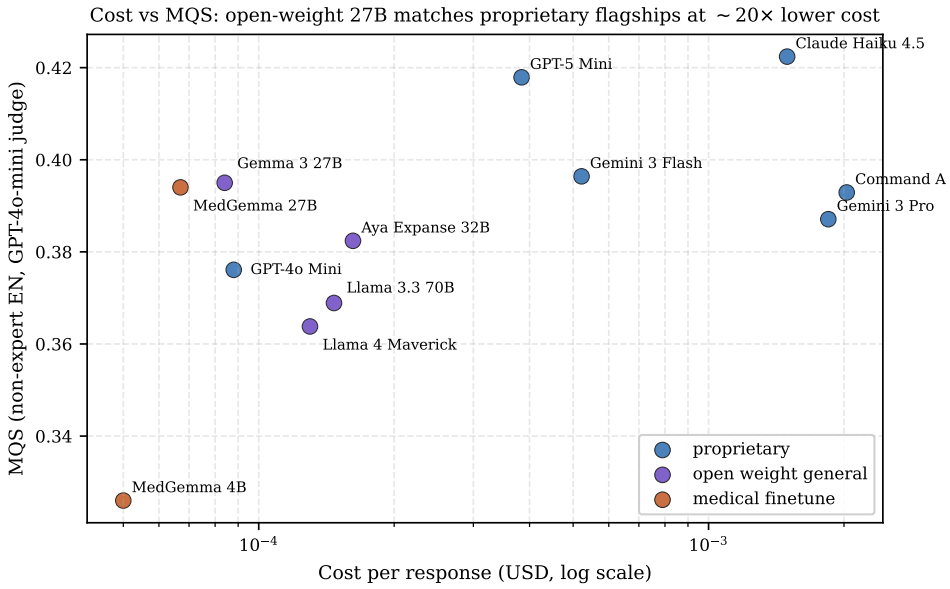

Figure 4: Cost per response (USD, log scale) versus non-expert EN MQS under the GPT-4o-mini judge. Colour encodes training family: proprietary API, open-weight general-purpose, and medical-fine-tuned. The efficient frontier is dominated by open-weight 27B models; the proprietary cluster (top right) does not separate from them on MQS despite costing $\sim 20\times$ more. Full numeric table in Appendix A.4 (Table 6).

Context Awareness (0.226–0.264) for every model in the panel. The failures are lexical: English abbreviations left untranslated at first use, Indian generic drug names rendered in Latin script instead of Devanagari, and metric shorthand that stays in English even when the question is posed in Marathi. Pointing at "culture" is too diffuse to act on. Pointing at terminology is actionable: curated parallel medical glossaries, abbreviation expansion at first use, and alignment fine-tuning on Indian Pharmacopoeia drug names are each narrow, measurable interventions. This sharpens the CLINIC finding that multilingual trustworthiness varies across scripts (Ghosh et al., 2025) by locating the weakness at the lexical level rather than at the reasoning level.

Apparent cross-lingual parity is a judge artifact. Under GPT-4o-mini, 11 of 13 models score equal or higher on Hindi than on English, and Marathi shows the same pattern for most of the panel (§5.3). Taken alone this would imply that frontier LLMs handle Indic-script maternal-health queries at least as well as English queries. Our judge-agreement analysis shows the result is not load-bearing. GPT-5.1 applied to the stratified 500-response subset scores every model 0.17 to 0.20 MQS lower on identical outputs, and the offset is not uniform across languages. Rank correlations across judges are modest: Pearson 0.24 between GPT-4o-mini and GPT-5.1, and 0.43 between GPT-4o-mini and Claude Opus 4.6. The judges disagree on absolute level and, separately, on which responses are better. Any paper that reports apparent cross-lingual parity on the basis of a single mid-tier judge is reporting a potential artifact. At minimum, multilingual medical benchmarks should triangulate across one stricter judge and publish the per-language bias offset alongside the headline scores.

Open-weight medium-scale models are the rational deployment choice. The cost-performance frontier (§5.5) is unambiguous for the operator profile that motivates this benchmark, namely NGOs running maternal-health bots at rural-India scale. MedGemma 27B and Gemma 3 27B both reach MQS $\approx$ 0.395 on the non-expert English arm at roughly \$0.00008 per response. Claude Haiku 4.5, the panel's strongest proprietary model, reaches 0.422 at about \$0.0015 per response, so the extra \$0.00142 per call buys +0.027 MQS. For a single bot serving one million queries per month, that is \$17,040 per month for three percentage points of rubric compliance, money that, on the evidence in §5.4, would produce a larger safety gain if spent on Hindi and Marathi medical-terminology alignment instead. MedGemma 27B and its base Gemma 3 27B are indistinguishable on non-expert EN (MQS 0.394 vs. 0.395), a controlled matched-pair result that confirms the hypothesis we called H2-baseline: medical fine-tuning alone does not lift rubric com-

pliance in this cross-cultural setting. The lever is translation and terminology, not medical specialization.

Automated metrics are not a substitute for the rubric. Embedding similarity and BERTScore can rate an answer highly on topical alignment while the rubric flags the same response as missing specific clinical criteria that matter for patient safety: drug-interaction warnings, the correct antenatal visit schedule, the right danger-sign threshold for seeking immediate care. That is the pattern prior work on metric-versus-clinical-quality divergence already documents (Tam et al., 2024; Croxford et al., 2025), and it is why the two-layer design is the floor rather than an optional add-on. The automated layer scales and reproduces; the rubric is what the hospital or the NGO programme officer will actually read before deciding to deploy.

Situating Sakhi among related benchmarks. HealthBench (Arora et al., 2025) is the closest methodological comparator, with rubric-style evaluation across 26 medical specialties, but it is predominantly English and uses synthetic conversations rather than community-validated questions. CLINIC (Ghosh et al., 2025) covers 15 languages but assesses generic trustworthiness, not domain-specific clinical quality. Script Gap (Khullar et al., 2025) evaluates triage classification accuracy on Indian-language maternal-health conversations but stops short of open-ended generation quality. Sakhi contributes three pieces that the prior benchmarks do not jointly offer. First, reference answers in all three languages on a single question set, both doctor-edited (expert arm) and community-sourced (non-expert arm), so the MQS gap between them is measurable within the same model. Second, a stratified 13-model panel drawn from a pre-declared eligibility frame, with within-family flagship/efficient pairs and a base-versus-medical-finetune matched comparator that lets us attribute, or refuse to attribute, gains to medical fine-tuning. Third, a three-judge calibration on a stratified 500-response subset that quantifies how much the headline MQS depends on the judge model. We read these as depth where existing work offers breadth. The paper is written for the NeurIPS Datasets and Benchmarks track, and the intended users are NGO deployers and research teams who need an evaluation that fails for the same reasons patients actually get hurt.

Limitations

Several limitations shape what our findings can and cannot support. First, the benchmark covers a focused, doctor-validated subset of maternal and reproductive health queries (149 expert-edited plus 231 community-sourced questions) and does not exhaust the full space of questions a bot will receive in deployment. Second, Hindi and Marathi are represented through professional translations of English questions and reference answers. This controls for topic drift across languages at the cost of not reflecting the full variation of spontaneous Indic-script patient queries, which often code-switch between Hindi and English, use non-standard transliteration, and lean on regional idiom. A fully native-authored Indic-first question set is the natural next dataset release.

Third, several coverage gaps in the 13-model panel are reported rather than hidden. Claude Opus 4.7 has full coverage on the expert English and expert Hindi cells, one complete run on expert Marathi (n=108), partial coverage on non-expert Marathi (n=113), and no valid responses on non-expert English or non-expert Hindi, all driven by sustained Claude Code CLI session limits during the dispatch window. Aya Expanse 32B was dispatched on the non-expert arm only because its native Cohere-API route required credentials we did not have available for the expert track. MedGemma 4B and 27B are reported on the non-expert arm only because the OpenRouter endpoint used to dispatch the other models does not host them and the HuggingFace Inference Endpoints route we set up for MedGemma was scoped to the non-expert track in this release. Gemini 3.1 Flash-Lite produced no valid output in any cell and is excluded from scored comparisons but kept in Table 7 so the failure is visible. None of these gaps changes the panel-wide ranking, but each qualifies the specific cell it touches.

Fourth, the rubric-judge prompt lists the rubric criteria as predicates over a labeled Question / Reference / Response triplet and asks for binary JSON scores, but does not explicitly instruct the judge to compare the response against the reference answer (Appendix B.3). The design is deliberate: each criterion is written as a predicate that a correct response should satisfy (for example, "Correctly emphasizing the recommended schedule of regular antenatal check-ups"), and the reference is provided as clinically-grounded context rather than as a gold-

standard target to be matched. This differs from HealthBench (Arora et al., 2025), whose physician-designed rubrics are tied to conversation-specific criteria scored against a single reference. The trade-off is that a judge is left to infer the evaluation task from the structural layout of the prompt, and the three-judge calibration on the 500-response subset (Section 5.7) is what tells us how much that under-specification matters in practice. Pearson $\rho = 0.24$ between GPT-4o-mini and GPT-5.1 means the two judges disagree not just on strict-vs-lenient calibration but on the relative ordering of responses. A revision of the prompt that explicitly asks the judge to compare response against reference is one of the first follow-ups we would run, and we kept the prompt unchanged across this release to preserve comparability with earlier scored runs.

Fifth, the primary rubric judge is GPT-4o-mini. The cross-judge calibration with Claude Opus 4.6 and GPT-5.1 on the stratified 500-response subset (Section 5.7) quantifies the bias this introduces, and we report judge identity and offset throughout, but the full $13\times2\times3$ grid is scored by a single primary judge. Extending the three-judge calibration to the entire grid is bounded by API cost, not by methodology.

Sixth, human reviewer capacity is the hardest constraint on this release and the reason this paper is about LLM-judge rubric scores rather than human rubric scores at scale. The review platform described in Section 3 was sized for doctor-edited reference answers on the 380 pairs in the two released tracks and for ASHA-worker and nonprofit-staff review of those pairs, not for per-response scoring of the 44,460 generated model outputs. A capacity-bounded human-judge pass that mirrors the LLM three-judge calibration, covering a stratified subset balanced across the 13 models, the 10 themes, and the 3 languages, is the critical next data release for tightening the MQS estimates and the judge-agreement analysis. We will report the sampling design, theme distribution, and per-cell N alongside that release rather than reconstruct a human-vs-LLM agreement claim that the current data does not support.

Seventh, the rubric is binary (pass/fail on 15 criteria per theme). This sharpens agreement but misses the difference between a criterion that is met weakly and one that is met well. A graded follow-up rubric is left to future work. Eighth, responses are scored single-turn, without clarifying exchanges. Real maternal-health conversations involve follow-up and disambiguation, and a model that would have clarified under a multi-turn protocol is penalized here for its single-turn output. Finally, a benchmark is necessary but not sufficient for deployment. Any real-world use of these models in maternal-health advice requires additional safety validation, regulatory and ethics review, and ongoing monitoring that sits outside the scope of this paper.

Acknowledgments

References

National Antenatal Care Guideline 2022.pdf.

Mahyar Abbasian, Elahe Khatibi, Iman Azimi, David Oniani, Zahra Shakeri Hossein Abad, Alexander Thieme, Ram Sriram, Zhongqi Yang, Yanshan Wang, Bryant Lin, Olivier Gevaert, Li-Jia Li, Ramesh Jain, and Amir M. Rahmani. 2024. Foundation metrics for evaluating effectiveness of healthcare conversations powered by generative AI. *NPJ Digital Medicine*, 7:82.

Iñigo Alonso, Maite Oronoz, and Rodrigo Agerri. 2024. MedExpQA: Multilingual benchmarking of Large Language Models for Medical Question Answering. *Artificial Intelligence in Medicine*, 155:102938.

Lameck Mbangula Amugongo, Pietro Mascheroni, Steven Brooks, Stefan Doering, and Jan Seidel. 2025. Retrieval augmented generation for large language models in healthcare: A systematic review. *PLOS digital health*, 4(6):e0000877.

D. M. Anisuzzaman, Jeffrey G. Malins, Paul A. Friedman, and Zachi I. Attia. 2025. Fine-Tuning Large Language Models for Specialized Use Cases. *Mayo Clinic Proceedings: Digital Health*, 3(1):100184.

Maria Antoniak, Aakanksha Naik, Carla S. Alvarado, Lucy Lu Wang, and Irene Y. Chen. 2024. NLP for maternal healthcare: Perspectives and guiding principles in the age of LLMs. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*, New York, NY, USA. Association for Computing Machinery.

Rahul K. Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, Johannes Heidecke, and Karan Singhal. 2025. HealthBench: Evaluating Large Language Models Towards Improved Human Health. *arXiv preprint*. ArXiv:2505.08775 [cs].

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor,

Michigan. Association for Computational Linguistics.

Joachim Baumann, Paul Röttger, Aleksandra Urman, Albert Wendsjö, Flor Miriam Plaza-del Arco, Johannes B. Gruber, and Dirk Hovy. 2025. Large Language Model Hacking: Quantifying the Hidden Risks of Using LLMs for Text Annotation. *arXiv preprint*. ArXiv:2509.08825 [cs].

Anagani Bhanusree, Sai Divya Vissamsetty, K VenkataKrishna Rao, and Rimjhim. 2026. Comparative analysis of large language models in generating Telugu responses for maternal health queries. *Preprint*, arXiv:2603.18898.

Iva Bojic, Jessica Chen, Si Yuan Chang, Qi Chwen Ong, Shafiq Joty, and Josip Car. 2023. Hierarchical Evaluation Framework: Best Practices for Human Evaluation. In *Proceedings of the 3rd Workshop on Human Evaluation of NLP Systems*, pages 11–22, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the Role of Bleu in Machine Translation Research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 249–256, Trento, Italy. Association for Computational Linguistics.

Xiaolan Chen, Jiayang Xiang, Shanfu Lu, Yexin Liu, Mingguang He, and Danli Shi. 2025. Evaluating large language models and agents in healthcare: key challenges in clinical applications. *Intelligent Medicine*, 5(2):151–163.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, Alexandre Sallinen, Alireza Sakhaeirad, Vinitra Swamy, Igor Krawczuk, Deniz Bayazit, Axel Marmet, Syrielle Montariol, Mary-Anne Hartley, Martin Jaggi, and Antoine Bosselut. 2023. MEDITRON-70B: Scaling Medical Pretraining for Large Language Models. Publisher: arXiv Version Number: 1.

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. All That's 'Human' Is Not Gold: Evaluating Human Evaluation of Generated Text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7282–7296, Online. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating Cross-lingual Sentence Representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Emma Croxford, Yanjun Gao, Nicholas Pellegrino, Karen Wong, Graham Wills, Elliot First, Frank Liao, Cherodeep Goswami, Brian Patterson, and Majid Afshar. 2025. Current and future state of evaluation of large language models for medical summarization tasks. *npj Health Systems*, 2(1):6. Publisher: Nature Publishing Group.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, and 26 others. 2024. Aya Expanse: Combining Research Breakthroughs for a New Multilingual Frontier. *arXiv preprint*. ArXiv:2412.04261 [cs].

Goutam Datta, Nisheeth Joshi, and Kusum Gupta. 2022. Analysis of Automatic Evaluation Metric on Low-Resourced Language: BERTScore vs BLEU Score. In *Speech and Computer*, pages 155–162, Cham. Springer International Publishing.

Samruddhi Deode, Janhavi Gadre, Aditi Kajale, Ananya Joshi, and Raviraj Joshi. 2023. L3Cube-IndicSBERT: A simple approach for learning cross-lingual sentence representations using multilingual BERT. *arXiv preprint*. ArXiv:2304.11434 [cs].

Indian Express. 2025. Top 10 most spoken languages globally in 2025: 5 indian languages ranked. Accessed: 2025-11-03.

David Fraile Navarro, Enrico Coiera, Thomas W. Hambly, Zoe Triplett, Nahyan Asif, Anindya Susanto, Anamika Chowdhury, Amaya Azcoaga Lorenzo, Mark Dras, and Shlomo Berkovsky. 2025. Expert evaluation of large language models for clinical dialogue summarization. *Scientific Reports*, 15(1):1195. Publisher: Nature Publishing Group.

Agasthya Gangavarapu. 2024. Introducing L2M3, a multilingual medical large language model to advance health equity in low-resource regions. *Preprint*, arXiv:2404.08705.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Akash Ghosh, Srivarshinee Sridhar, Raghav Kaushik Ravi, Muhsin Muhsin, Sriparna Saha, and Chirag Agarwal. 2025. CLINIC: Evaluating multilingual trustworthiness in language models for healthcare. *Preprint*, arXiv:2512.11437.

Tianyu Han, Sven Nebelung, Firas Khader, Tianci Wang, Gustav Müller-Franzes, Christiane Kuhl, Sebastian Försch, Jens Kleesiek, Christoph Haarburger, Keno K. Bressem, Jakob Nikolas Kather, and Daniel Truhn. 2024. Medical large language models are

susceptible to targeted misinformation attacks. *npj Digital Medicine*, 7(1):288. Publisher: Nature Publishing Group.

Michael Hanna and Ondřej Bojar. 2021. A Fine-Grained Analysis of BERTScore. In *Proceedings of the Sixth Conference on Machine Translation*, pages 507–517, Online. Association for Computational Linguistics.

Helia Hashemi, Jason Eisner, Corby Rosset, Benjamin Van Durme, and Chris Kedzie. 2024. LLM-Rubric: A Multidimensional, Calibrated Approach to Automated Evaluation of Natural Language Texts. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13806–13834, Bangkok, Thailand. Association for Computational Linguistics.

F. Jelinek, R. L. Mercer, L. R. Bahl, and J. K. Baker. 2005. Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, 62(S1):S63.

Yuelyu Ji, Hang Zhang, and Yanshan Wang. 2025. Bias Evaluation and Mitigation in Retrieval-Augmented Medical Question-Answering Systems. *arXiv preprint*. ArXiv:2503.15454 [cs].

Yiqiao Jin, Mohit Chandra, Gaurav Verma, Yibo Hu, Munmun De Choudhury, and Srijan Kumar. 2024. Better to ask in english: Cross-lingual evaluation of large language models for healthcare queries. In *Proceedings of the ACM Web Conference 2024*, WWW '24, page 2627–2638, New York, NY, USA. Association for Computing Machinery.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. 2023. Dspy: Compiling declarative language model calls into self-improving pipelines. *Preprint*, arXiv:2310.03714.

Manurag Khullar, Utkarsh Desai, Poorva Malviya, Aman Dalmia, and Zheyuan Ryan Shi. 2025. Script gap: Evaluating LLM triage on Indian languages in native vs romanized scripts in a real world setting. *Preprint*, arXiv:2512.10780.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Masaaki Nagata, Toshiaki Nakazawa, Martin Popel, and 3 others. 2023. Findings of the 2023 Conference on Machine Translation (WMT23): LLMs Are Here but Not Quite There Yet. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42, Singapore. Association for Computational Linguistics.

Anisio Lacerda, Gisele Pappa, Adriano Pereira, Wagner Meira Jr, and Alexandre Barros. 2025. Evaluation of medical large language models: Taxonomy, review, and directions. pages 10528–10536.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

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*, pages 2511–2522, Singapore. Association for Computational Linguistics.

W. Longchar and 1 others. 2025. Trends and determinants of maternal health services utilization in india. *Scientific Reports*.

Lars Malmqvist. 2024. Sycophancy in large language models: Causes and mitigations. *Preprint*, arXiv:2411.15287.

Clara Meister and Ryan Cotterell. 2021. Language Model Evaluation Beyond Perplexity. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5328–5339, Online. Association for Computational Linguistics.

Birger Moëll and Fredrik Sand Aronsson. 2025. Harm Reduction Strategies for Thoughtful Use of Large Language Models in the Medical Domain: Perspectives for Patients and Clinicians. *Journal of Medical Internet Research*, 27:e75849.

Aarathi Rajagopalan Nair, Deepa Gupta, and B. Premjith. 2024. Investigating translation for Indic languages with BLOOMZ-3b through prompting and LoRA fine-tuning. *Scientific Reports*, 14:24202.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Matt Post. 2018. A Call for Clarity in Reporting BLEU Scores. In *Proceedings of the Third Conference on*

Machine Translation: Research Papers, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

David Restrepo, Chenwei Wu, Zhengxu Tang, Zitao Shuai, Thao Nguyen Minh Phan, Jun-En Ding, Cong-Tinh Dao, Jack Gallifant, and 1 others. 2025. Multi-OphthaLingua: A multilingual benchmark for assessing and debiasing LLM ophthalmological QA in LMICs. In *Proceedings of the AAAI Conference on Artificial Intelligence (AISI Track)*.

Pritika Rohera, Chaitrali Ginimav, Gayatri Sawant, and Raviraj Joshi. 2025. Better to ask in english? evaluating factual accuracy of multilingual llms in english and low-resource languages. *Preprint*, arXiv:2504.20022.

Ananya B. Sai, Akash Kumar Mohankumar, and Mitesh M. Khapra. 2022. A Survey of Evaluation Metrics Used for NLG Systems. *ACM Comput. Surv.*, 55(2):26:1–26:39.

Thomas Savage, Stephen P. Ma, Abdessalem Boukil, Ekanath Rangan, Vishwesh Patel, Ivan Lopez, and Jonathan Chen. 2025. Fine-Tuning Methods for Large Language Models in Clinical Medicine by Supervised Fine-Tuning and Direct Preference Optimization: Comparative Evaluation. *Journal of Medical Internet Research*, 27(1):e76048. Company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research Publisher: JMIR Publications Inc., Toronto, Canada.

Hendrik Schuff, Lindsey Vanderlyn, Heike Adel, and Ngoc Thang Vu. 2023. How to do human evaluation: A brief introduction to user studies in NLP. *Natural Language Engineering*, 29(5):1199–1222.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, Justin Chen, Fereshteh Mahvar, Liron Yatziv, Tiffany Chen, Bram Sterling, Stefanie Anna Baby, Susanna Maria Baby, Jeremy Lai, Samuel Schmidgall, and 62 others. 2025. MedGemma Technical Report. *arXiv preprint*. ArXiv:2507.05201 [cs].

Junhyuk Seo, Dasol Choi, Taerim Kim, Won Chul Cha, Minha Kim, Haanju Yoo, Namkee Oh, YongJin Yi, Kye Hwa Lee, and Edward Choi. 2024. Evaluation Framework of Large Language Models in Medical Documentation: Development and Usability Study. *Journal of Medical Internet Research*, 26(1):e58329. Company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research Publisher: JMIR Publications Inc., Toronto, Canada.

Harman Singh, Nitish Gupta, Shikhar Bharadwaj, Dinesh Tewari, and Partha Talukdar. 2024. IndicGenBench: A Multilingual Benchmark to Evaluate Generation Capabilities of LLMs on Indic Languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11047–11073, Bangkok, Thailand. Association for Computational Linguistics.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R. Pfohl, Heather Cole-Lewis, Darlene Neal, Qazi Mamunur Rashid, Mike Schaekermann, Amy Wang, Dev Dash, Jonathan H. Chen, Nigam H. Shah, Sami Lachgar, Philip Andrew Mansfield, and 16 others. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31(3):943–950. Publisher: Nature Publishing Group.

Rickard Stureborg, Dimitris Alikaniotis, and Yoshi Suhara. 2024. Large Language Models are Inconsistent and Biased Evaluators. *arXiv preprint*. ArXiv:2405.01724 [cs].

Annalisa Szymanski, Noah Ziems, Heather A. Eicher-Miller, Toby Jia-Jun Li, Meng Jiang, and Ronald A. Metoyer. 2025. Limitations of the LLM-as-a-Judge Approach for Evaluating LLM Outputs in Expert Knowledge Tasks. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, IUI '25, pages 952–966, New York, NY, USA. Association for Computing Machinery.

Thomas Yu Chow Tam, Sonish Sivarajkumar, Sumit Kapoor, Alisa V. Stolyar, Katelyn Polanska, Karleigh R. McCarthy, Hunter Osterhoudt, Xizhi Wu, Shyam Visweswaran, Sunyang Fu, Piyush Mathur, Giovanni E. Cacciamani, Cong Sun, Yifan Peng, and Yanshan Wang. 2024. A framework for human evaluation of large language models in healthcare derived from literature review. *npj Digital Medicine*, 7(1):258. Publisher: Nature Publishing Group.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024. Large Language Models are not Fair Evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450, Bangkok, Thailand. Association for Computational Linguistics.

Koki Wataoka, Tsubasa Takahashi, and Ryokan Ri. 2025. Self-Preference Bias in LLM-as-a-Judge. *arXiv preprint*. ArXiv:2410.21819 [cs].

Junde Wu, Jiayuan Zhu, Yunli Qi, Jingkun Chen, Min Xu, Filippo Menolascina, Yueming Jin, and Vicente

Grau. 2025. Medical Graph RAG: Evidence-based Medical Large Language Model via Graph Retrieval-Augmented Generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28443–28467, Vienna, Austria. Association for Computational Linguistics.

Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, Nitesh V. Chawla, and Xiangliang Zhang. 2024. Justice or Prejudice? Quantifying Biases in LLM-as-a-Judge. *arXiv preprint*. ArXiv:2410.02736 [cs].

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text Generation Evaluating with Contextualized Embeddings and Earth Mover Distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578, Hong Kong, China. Association for Computational Linguistics.

Hongli Zhou, Hui Huang, Yunfei Long, Bing Xu, Conghui Zhu, Hailong Cao, Muyun Yang, and Tiejun Zhao. 2024. Mitigating the Bias of Large Language Model Evaluation. In *Proceedings of the 23rd Chinese National Conference on Computational Linguistics (Volume 1: Main Conference)*, pages 1310–1319, Taiyuan, China. Chinese Information Processing Society of China.

A Full Results Tables

This appendix holds the full numeric tables underpinning the figures in Section 5. Each table is produced by `scoring/pipeline/render_tables.py` from the aggregated JSONL judge logs; re-running that script regenerates every number.

A.1 Cross-Lingual MQS

A.2 MQS by Theme

A.3 Axis Pass Rates, Full Panel

A.4 Cost-Performance Details

A.5 Generation Validity

B Prompt Templates and API Configuration

All models were evaluated using the same prompt templates to ensure comparability. No few-shot examples were provided to the evaluated models or to the rubric judge, following a zero-shot design to avoid biasing evaluation toward specific response patterns.

Table 3: Cross-lingual MQS averaged over both dataset arms. ΔHI and ΔMR show the signed change from English; positive means the model scores higher in Hindi or Marathi than English under the GPT-4o-mini judge. **Bold** marks the best (highest) score per column; underline marks the second best.

Model	EN	HI	MR	ΔHI	ΔMR
GPT-5 Mini	0.358	0.372	0.365	+0.014	+0.007
GPT-4o Mini	0.309	0.323	0.305	+0.014	-0.004
Command A	0.324	0.341	0.317	+0.017	-0.008
Gemini 3 Pro	0.321	0.341	0.335	+0.019	+0.014
Gemini 3 Flash	0.331	0.357	0.362	+0.026	+0.031
Claude Opus 4.7	0.297	0.354	0.358	**+0.058**	**+0.061**
Claude Haiku 4.5	0.364	0.386	0.375	+0.022	+0.011
Llama 3.3 70B	0.301	0.316	0.305	+0.015	+0.004
Llama 4 Maverick	0.309	0.328	0.324	+0.019	+0.015
Aya Expanse 32B	0.382	0.369	0.337	-0.013	-0.045
MedGemma 27B	**0.394**	**0.392**	**0.383**	-0.002	-0.011
MedGemma 4B	0.326	0.328	0.331	+0.002	+0.005
Gemma 3 27B (baseline)	0.330	0.344	0.335	+0.014	+0.004

B.1 English Response Generation Prompt

The following prompt was used for all models generating English responses. The `{question}` placeholder was replaced with each question from the dataset.

```
You are a knowledgeable and caring
assistant trained to support pregnancy-
related health. Your task is to provide
accurate, empathetic, and reliable answers
to user questions specifically about
pregnancy, prenatal care, and postnatal
well-being.

LANGUAGE: You MUST answer strictly in
English. Use a warm, supportive, and
informative tone. Avoid medical jargon;
explain any technical term in plain,
easy-to-understand language.

SCOPE: Respond to questions related to
women's health, including pregnancy,
prenatal care, postnatal well-being,
reproductive health and contraception,
sexual health and wellness, menstrual
health and disorders, common gynecological
conditions, nutrition, mental health, and
general wellness unique to women.

For very short or unclear questions, assume
a pregnancy-related intent and restate the
implied question clearly before answering.
Examples:
- "Food?" -> "What kind of food should I
  eat during pregnancy?"
- "Exercise?" -> "What kind of exercise is
  safe during pregnancy?"
- "Swelling feet" -> "Is swelling in feet
  normal during pregnancy and what can I
  do?"

ANSWER FORMAT: Provide a single-paragraph
answer that is clear, concise, and around
```

Table 4: MQS by maternal-health theme (English, both dataset arms averaged, GPT-4o-mini judge). **Bold** marks the best score per theme column; underline marks the second best. Theme labels are rotated; short names correspond to the ten themes described in Section 4 (e.g., Antenatal = Antenatal & Maternal Health Care).

Model	Antenatal	Clinical	Health Sys.	Infection	Medication	Mental	Nutrition	Reproductive	Risk Mgmt.	Symptoms
GPT-5 Mini	0.308	0.207	0.391	0.409	**0.379**	0.506	0.635	0.320	0.290	0.491
GPT-4o Mini	0.249	0.203	0.365	0.394	0.291	0.461	0.612	0.276	0.262	0.419
Command A	0.260	0.181	0.359	0.358	0.328	0.499	0.648	0.310	0.245	0.431
Gemini 3 Pro	0.266	0.205	0.373	0.335	0.318	0.476	0.626	0.272	0.232	0.419
Gemini 3 Flash	0.271	0.154	0.344	0.405	0.338	**0.511**	0.642	0.289	0.274	0.460
Claude Opus 4.7	0.302	0.050	–	**0.525**	0.236	–	–	0.222	**0.478**	0.350
Claude Haiku 4.5	**0.313**	**0.221**	**0.393**	0.405	0.363	0.509	**0.676**	0.351	0.305	**0.512**
Llama 3.3 70B	0.232	0.176	0.350	0.388	0.345	0.445	0.607	0.293	0.204	0.439
Llama 4 Maverick	0.250	0.169	0.359	0.382	0.340	0.445	0.586	0.290	0.247	0.399
Aya Expanse 32B	0.261	0.173	0.352	0.352	0.372	0.452	0.645	0.342	0.202	0.457
MedGemma 27B	0.278	0.193	0.371	0.349	0.355	0.492	0.631	**0.370**	0.209	0.510
MedGemma 4B	0.242	0.134	0.323	0.281	0.256	0.434	0.551	0.356	0.103	0.381
Gemma 3 27B (baseline)	0.273	0.213	0.392	0.433	0.356	0.488	0.621	0.305	0.242	0.427

Table 5: Axis-level pass rates on the non-expert EN arm (GPT-4o-mini judge). Values are the mean fraction of binary rubric criteria satisfied within each axis. **Terminology Accessibility is the lowest axis for every model** (0.123–0.206), followed by Context Awareness (0.226–0.264), so the dominant failure mode is lexical (untranslated medical shorthand, abbreviations, drug names) rather than broadly cultural. **Bold** and underline mark the best and second-best model per axis column.

Model	Accuracy	Completeness	Context Awareness	Communication	Terminology Accessibility
GPT-5 Mini	0.416	0.396	0.260	**0.700**	**0.206**
GPT-4o Mini	0.357	0.349	0.248	0.644	0.148
Command A	0.381	0.368	0.250	0.676	0.160
Gemini 3 Pro	0.373	0.350	0.236	0.668	0.170
Gemini 3 Flash	0.387	0.370	0.239	0.688	0.163
Claude Haiku 4.5	**0.426**	**0.403**	**0.264**	0.689	0.193
Llama 3.3 70B	0.344	0.336	0.233	0.650	0.130
Llama 4 Maverick	0.346	0.337	0.226	0.644	0.131
Aya Expanse 32B	0.360	0.367	0.241	0.655	0.148
MedGemma 27B	0.389	0.360	0.253	0.675	0.159
MedGemma 4B	0.315	0.295	0.230	0.583	0.123
Gemma 3 27B (baseline)	0.384	0.357	0.251	0.688	0.170

```
60-80 words, maximum 100 words. Do NOT use
bullet points or lists. Always recommend
consulting a doctor for serious symptoms,
diagnoses, or uncertainties.

QUESTION: {question}

Provide only the answer text without any
tags or formatting.
```

B.2 Multilingual Response Generation Prompt

For Hindi and Marathi response generation, the following prompt was used. The `{language}` placeholder was set to "Hindi" or "Marathi" and `{question}` was replaced with the corresponding translated question.

```
You are a knowledgeable and caring
assistant trained to support pregnancy-
related health. [...]

CRITICAL LANGUAGE REQUIREMENT: You MUST
answer ONLY in {language}. DO NOT use
English. Your entire response must be in
{language} script and language. This is
absolutely mandatory.

[Same SCOPE and ANSWER FORMAT as English]

QUESTION: {question}

Remember: Your answer MUST be entirely in
{language}. Provide only the answer text
without any tags or formatting.
```

Table 6: Cost-performance frontier. Per-response generation cost (USD) paired with mean non-expert EN MQS under the GPT-4o-mini judge. MQS/USD is an efficiency indicator; higher is better. **Bold** marks the best and underline marks the second best in each column. Claude Opus 4.7 is accessed through a local Claude Code subscription at zero marginal cost, so it is excluded from the cost and MQS/USD ranking.

Model	$/resp	MQS	MQS/$
GPT-5 Mini	0.00038	0.418	1,088
GPT-4o Mini	0.00009	0.376	4,274
Command A	0.00203	0.393	194
Gemini 3 Pro	0.00185	0.387	210
Gemini 3 Flash	0.00052	0.396	759
Claude Opus 4.7	0.00000	–	–
Claude Haiku 4.5	0.00150	**0.422**	283
Llama 3.3 70B	0.00015	0.369	2,510
Llama 4 Maverick	0.00013	0.364	2,798
Aya Expanse 32B	0.00016	0.382	2,360
MedGemma 27B	0.00007	0.394	5,881
MedGemma 4B	**0.00005**	0.326	**6,520**
Gemma 3 27B (baseline)	0.00008	0.395	4,702

Table 7: Valid-response rate per (dataset, language) averaged across the three runs. Lower values reflect empty, truncated, or refused generations, which count as a distinct failure mode from low MQS on a valid response. We retain models with zero valid responses in certain cells (e.g., Gemini 3.1 Flash-Lite) so that coverage gaps are visible rather than hidden.

Model	Expert (149)			Non-expert (231)		
	EN	HI	MR	EN	HI	MR
GPT-5 Mini	1.000	1.000	1.000	1.000	1.000	1.000
GPT-4o Mini	1.000	1.000	1.000	1.000	1.000	1.000
Command A	1.000	1.000	1.000	1.000	1.000	1.000
Gemini 3 Pro	1.000	1.000	1.000	1.000	1.000	1.000
Gemini 3 Flash	1.000	1.000	1.000	1.000	1.000	1.000
Gemini 3.1 Flash-Lite	0.000	0.000	0.000	0.000	0.000	0.000
Claude Opus 4.7	1.000	1.000	0.271	0.000	0.000	0.089
Claude Haiku 4.5	1.000	1.000	1.000	1.000	1.000	1.000
Llama 3.3 70B	1.000	1.000	1.000	1.000	1.000	1.000
Llama 4 Maverick	0.996	1.000	1.000	1.000	1.000	1.000
Aya Expanse 32B	0.000	0.000	0.000	1.000	1.000	1.000
MedGemma 27B	0.000	–	–	1.000	1.000	1.000
MedGemma 4B	–	–	–	1.000	1.000	1.000
Gemma 3 27B (baseline)	1.000	1.000	1.000	1.000	1.000	1.000

B.3 LLM Rubric Judge Prompt

Rubric scoring used a zero-shot prompt with no few-shot examples, intentionally excluded to avoid biasing the judge toward specific scoring patterns and to encourage natural evaluation behavior. The judge received two messages:

System message:

```
Return valid JSON only. Use the exact
rubric text as keys.
```

User message:

```
Question:
{question}

Reference:
{reference_answer}

Response:
{model_response}

Rubrics:
1. {rubric_criterion_1}
2. {rubric_criterion_2}
...
15. {rubric_criterion_15}

Return JSON: {"scores": {"<Exact Rubric
Text>": 0 or 1}}
```

B.4 API Configuration

Table 8 documents the API parameters used across all experiments.

B.5 Dataset Generation Prompt

The Sakhi dataset was generated using a few-shot prompt with detailed user demographic context reflecting the target population: married women aged 23–33 from rural and semi-urban India, with education spanning no formal schooling (6%) to graduates (22%), limited dietary diversity, and reliance on family networks and ASHA workers for health information. Four few-shot examples spanning STD prevention, antenatal care visits, stretch marks, and menstrual health were provided to calibrate answer style. Generated entries were required to use simple language accessible to low-literacy populations, avoid medical jargon, and cover diverse maternal health themes. The complete dataset generation prompt is available in `prompts.txt` in the supplementary code repository.

Table 8: API parameters for response generation and rubric scoring. Proprietary models are accessed via OpenRouter. Reasoning-capable models (GPT-5 family, Gemini 3 family, o-series) use the higher temperature and token budget together with OpenRouter's `reasoning={effort=low, exclude=true}` setting so that reasoning tokens do not consume the response budget and are not scored. Claude Opus 4.7 is invoked through the local Claude Code CLI at zero marginal cost. MedGemma 4B and 27B run on HuggingFace Inference Endpoints. The primary rubric judge is GPT-4o-mini; the calibration subset also uses Claude Opus 4.6 and GPT-5.1 under the same prompt and judge parameters.

Parameter	**Generation**	**Judge**
Temperature	0.7 (1.0 for reasoning models)	0.0
Max tokens	400 (2,000 for reasoning)	800
Runs per question	3	1
Timeout (s)	120	60
Retry attempts	3	5