WHBench: A Women's Health Benchmark for Evaluating Frontier LLMs with Expert-in-the-Loop Validation

Sneha Maurya
Columbia University
sm5755@columbia.edu

Pragya Saboo
Rubric AI
pragya@therubric.ai

Girish Kumar
Rubric AI
girish@therubric.ai

Abstract

Large language models are increasingly consulted for medical information, yet to our knowledge no widely adopted benchmark evaluates their performance on women's health, a domain where clinical guidelines shift frequently, treatment decisions hinge on individual patient factors, and the medical literature itself reflects a well-documented gender data gap. We introduce the **Women's Health Benchmark (WHBench)**: 47 expert-crafted clinical scenarios across 10 women's health topics, each targeting a specific LLM failure mode such as outdated guidelines, dosage errors, or missed health disparities. Reference answers were authored by board-certified clinicians including OB/GYN specialists, a gynecologic oncologist, general medicine physicians, and fertility nursing specialists. We evaluated 22 models that span frontier, reasoning, and open-source categories using a 23-criterion rubric across 8 clinical dimensions, with asymmetric penalties that weigh safety failures more heavily. Across 3,100 scored responses, no model's mean score exceeded 75% overall; the top model reached 72.1%, and the frontier tier remained tightly clustered, suggesting a capability ceiling not visible in saturated benchmarks. Performance was also uneven: even the best model achieved only 35.5% fully correct responses meeting the 80% correctness threshold, and harm rates varied substantially across otherwise strong systems. Inter-rater reliability was modest at the final label level ($\kappa = 0.238$) but much stronger for model ranking ($\rho = 0.916$), indicating that while single-response judgments remain noisy and require expert oversight, the benchmark is stable for comparing systems overall. We release WHBench as a public resource for evaluating clinical AI in women's health.

1 Introduction

A patient asks an AI system whether it is safe to receive a COVID-19 vaccine before starting an IVF cycle. Getting this right demands current ASRM guidelines, understanding of reproductive timelines, and sensitivity to the anxiety that surrounds fertility treatment. If the model draws on outdated information, the patient may delay a time-sensitive procedure.

This scenario is not hypothetical. Large language models are increasingly used to seek health information [Singhal et al., 2023], including on women's health topics such as fertility, contraception, pregnancy, and menopause. Yet these are the domains where models are most prone to error. Clinical guidelines evolve continuously: ATA thyroid thresholds in pregnancy have shifted multiple times in the past decade alone [Alexander et al., 2017]. Treatment decisions depend on intersecting patient factors like age, BMI, reproductive history, and race, requiring integrative reasoning rather than pattern matching. And the medical literature itself suffers from a well-documented gender data gap [Criado-Perez, 2019], meaning models trained on it inherit existing blind spots.

Existing benchmarks do not capture these risks. MedQA [Jin et al., 2021], MedMCQA [Pal et al., 2022], and MMLU [Hendrycks et al., 2021] test medical knowledge through multiple-choice questions that reward recognition over generation. HealthBench [OpenAI, 2025] moves toward open-ended health evaluation but does not target women's health failure modes, include clinician-authored reference answers, or assess health equity. To our knowledge, no benchmark today measures whether models account for health disparities, a gap with real consequences when AI-generated advice reaches diverse populations.

We introduce the **Women's Health Benchmark (WHBench)** with four contributions:

1. **Failure-mode-targeted scenarios.** 47 open-ended questions across 10 women's health topics, each designed to expose a specific LLM failure mode, paired with 4–6 independent expert reference answers per question.
2. **A multi-dimensional rubric.** 23 criteria organized into 8 categories, including what we believe is the first dedicated *Equity & Inclusivity* dimension in a medical LLM benchmark, with asymmetric severity-weighted penalties.
3. **Large-scale evaluation.** 22 models × 47 questions × 3 runs = 3,102 attempted responses 3100 scored, evaluated by two frontier LLM judges—**Claude Sonnet 4.6** as primary and **GPT-5.4** as secondary—with multi-judge inter-rater reliability analysis.
4. **Documented capability gaps.** Even the best and the latest model (Claude Opus 4.6, 72.1%) leaves substantial room for improvement, and every model tested performs poorly on social determinants of health.

2 Related Work

Medical LLM Benchmarks. MedQA [Jin et al., 2021] and MedMCQA [Pal et al., 2022] draw on licensing exams, but multiple-choice formats cannot assess the integrative reasoning that real patient queries demand. MultiMedQA [Singhal et al., 2023] aggregates several medical QA datasets without targeting gender-specific clinical domains. HealthBench [OpenAI, 2025] takes an important step toward open-ended evaluation with physician-designed rubrics, yet it lacks failure-mode targeting and health equity criteria. Ritchie et al. [Ritchie et al., 2026] demonstrated that domain-specific evaluation on realistic workplace tasks surfaces capability gaps invisible to general benchmarks; we carry this insight into clinical medicine.

LLM-as-Judge. Using LLMs to judge open-ended outputs has gained traction [Zheng et al., 2023], although leniency bias, particularly toward verbose responses, remains a known limitation. AdvancedIF [He et al., 2025] showed that compositional rubrics with per-item criteria catch failure modes that holistic scoring misses. We build on these insights with a "default to fail" judging philosophy, per-question clinical checklists authored by domain experts, and server-side score recalculation that prevents judge arithmetic errors from propagating into final scores.

Gender Bias and Health Equity in AI. Gender data gaps are prevalent in medical imaging [Larrazabal et al., 2020], clinical research [Criado-Perez, 2019], and NLP systems [Sun et al., 2019]. Obermeyer et al. [Obermeyer et al., 2019] showed that a widely deployed healthcare algorithm systematically disadvantaged Black patients by using cost as a proxy for need. Despite this body of work, we are not aware of any benchmark that evaluates LLM performance on women's health specifically or includes dedicated equity criteria. WHBench addresses both gaps.

3 Benchmark Design

3.1 Question Design

We designed 47 clinical scenarios around three guiding principles.

1. **Clinical realism:** questions mirror the kinds of queries real patients bring to providers, from straightforward factual questions ("What are the Rotterdam criteria for PCOS?") to complex multi-factor cases involving fertility with comorbidities or post-abortion contraception counseling.
2. **Failure-mode targeting:** each question is mapped to one of six error categories (Table 1), so that when a model fails we know not just *that* it failed but *how*.
3. **Difficulty calibration:** questions span four levels, from factual recall (Level 2, n=3) through frontier reasoning with conflicting evidence (Level 5, n=4), with intermediate and advanced scenarios making up the bulk.

Questions were developed in collaboration with board-certified OB/GYNs, reproductive endocrinologists, fertility specialists and our partner Dandi Fertility (`https://dandifertility.com`), a healthcare technology company whose network of registered fertility nurses spans all 50 U.S. states. Their clinical team contributed real-world patient scenario patterns that informed both scenario design and difficulty calibration. The full question set is listed in Appendix A.4.

Table 1: Failure mode taxonomy. Each WHBench question targets one primary failure mode.

Failure Mode	n	**Example**
Missing information	14	Omitting follow-up for chronic anovulation
Factual / outdated	12	Pre-2017 TSH thresholds in pregnancy
Health equity gaps	4	Ignoring racial disparities in fibroids
Incorrect treatment	4	Surgery over IVF for bilateral tubal disease
Contraindication / dosage	6	Wrong folic acid dose with valproate
Other (urgency, dx, recs)	7	Not flagging post-retrieval fever

Questions span 10 clinical topics: Fertility (n=10), Hormonal Health/HRT (n=7), Pregnancy (n=6), PCOS (n=5), Contraception (n=5), Endometriosis (n=4), Cancer Screening (n=4), Vaginal Health (n=3), Mental Health (n=2), and Bone Health (n=1).

3.2 Expert Reference Answers

Each question was answered independently by 4–6 members of our expert panel (mean 4.5 per question; full credentials in Appendix A.3). The panel includes multiple experts from OB/GYN (MBBS, MS; 20 years of experience), orthopedic surgeons with general medicine training (MBBS, MS; 10 years of experience), fertility nursing specialists from Dandi Fertility (BSN, RN; 8 years of experience), gynecologic surgeons (MBBS, MS; 11 years of experience), and gynecologic oncologists (MD; 10 years of experience). Experts had no access to each other's responses. Their answers average 94 words, reflecting the concise style of practicing clinicians and notably shorter than typical model outputs of 200 to 500 words.

3.3 Scoring Rubric

Table 2 presents the 23-criterion rubric across 8 categories. Three design choices set it apart.

1. **Asymmetric penalties:** safety criterion C9a carries -5 for failure vs. $+6$ for passing, while formatting criterion E17 carries only -1, ensuring a well-formatted but clinically dangerous response cannot score well.
2. **Equity evaluation:** F18a (social determinants) and F18b (bias avoidance) form what we believe is the first dedicated health equity dimension in a medical LLM benchmark.
3. **Ordinal depth:** four criteria (A3, B5, B6, B7) use 3-level ordinal scales, capturing partial credit.

Raw scores range from -58 to $+92$, normalized to 0–100% via (raw+58)/150×100. Responses are classified as *Correct* (≥80%), *Partially Correct* (45–79%), or *Incorrect* (<45%).

Table 2: WHBench v3.0 rubric: 23 criteria across 8 categories. † = ordinal (3-level). Point values reflect clinical severity.

Category	Criterion	Type	+	−
A: Clinical Accuracy	A1 Core clinical conclusion	Bin	6	3
	A2 Numerical precision	Bin	5	2
	A3 Guideline alignment†	Ord	4	2
	A4 Factual error absence	Bin	4	3
B: Completeness	B5 Clinical considerations†	Ord	5	2
	B6 Differential diagnosis†	Ord	3	1
	B7 Follow-up monitoring†	Ord	3	2
	B8 Patient-specific factors	Bin	3	2
C: Safety	C9a No unsafe commission	Bin	6	5
	C9b No unsafe omission	Bin	5	4
	C10 Urgency recognition	Bin	5	4
	C11 Contraindications	Bin	4	2
	C12 Dosage accuracy	Bin	4	3
D: Communication Quality	D13 Certainty calibration	Bin	3	2
	D14 Evolving evidence handling	Bin	3	2
	D15 Internal consistency	Bin	3	2
E: Instruction Follow	E16 Answers the question asked	Bin	6	2
	E17 Zero-shot compliance	Bin	2	1
F: Equity	F18a Social determinants	Bin	3	2
	F18b Bias avoidance	Bin	3	3
U: Uncertainty	U19 Appropriate uncertainty	Bin	4	3
	U20 Escalation and referral	Bin	5	4
G: Guideline Adherence	G21 Citation / guideline groundedness	Bin	3	2

4 Experiments

Models. We evaluate 22 models across four categories: *frontier* (GPT-5.4, GPT-4.1, GPT-4o, Claude Opus 4.6, Claude Sonnet 4.6, Claude Opus 4, Claude Sonnet 4, Gemini 3 Flash Preview, Gemini 2.5 Pro, Gemini 2.5 Flash, DeepSeek V3.2, Mistral Large, Grok 4, Grok 3, Grok 3 Mini), *reasoning-specialized* (OpenAI o3, DeepSeek-R1), and *open-source* (Llama 3.1 405B, Llama 3.3 70B, Llama 4 Maverick, Llama 4 Scout, Nemotron 70B). API-accessible models were queried through OpenRouter; self-hosted models (Llama 3.1 405B, Llama 3.3 70B) ran on NVIDIA A100 80GB GPUs provisioned through Vast.ai using vLLM. Full API identifiers appear in Appendix A.2.

Protocol. Each model receives all 47 questions in a **zero-shot, closed-book** setting with a standardized system prompt (Appendix A.1) instructing it to respond as a board-certified physician specializing in women's health. We set **temperature** = 0 and collect **3 independent runs** per model (22 × 47 × 3 = 3,102 total attempted responses, 3,100 scored).

Judging pipeline. Responses are scored by **Claude Sonnet 4.6** as primary judge, operating under a "default to fail" philosophy: each criterion starts at *Fail* unless the response clearly and explicitly meets the requirement. The judge receives the question, all expert reference answers, the model response, the targeted failure mode, and a per-question clinical checklist, but **not the model name**, ensuring blinded evaluation. Crucially, raw scores are **recalculated server-side** from individual pass/fail verdicts using fixed criterion weights, so that judge arithmetic errors cannot affect final scores. For inter-rater reliability, **GPT-5.4** independently scores all the models as a secondary judge.

5 Results

5.1 Overall Performance

Table 3 presents the WHBench v3.0 leaderboard and Figure 1 visualizes the score distribution.

Table 3: WHBench v3.0 leaderboard. Mean normalized score (%) across 3 runs with 95% bootstrap CI (n=10,000). C/P/I = Correct / Partially Correct / Incorrect rates. Harm = percentage of responses where either C9a (unsafe commission) or C9b (unsafe omission) failed.

#	Model	Score	95% CI	C%	P%	I%	Harm%
1	Claude Opus 4.6	72.1	[69.6, 74.4]	35.5	58.2	6.4	12.8
2	Claude Sonnet 4.6	67.1	[64.5, 69.6]	22.7	67.4	9.9	27.0
3	GPT-5.4	66.8	[64.5, 69.2]	21.3	67.4	11.3	47.5
4	Gemini 3 Flash Preview	64.7	[61.7, 67.7]	25.5	62.4	12.1	32.6
5	OpenAI o3	63.6	[61.3, 65.9]	15.0	76.4	8.6	38.6
6	DeepSeek V3.2	61.3	[58.6, 63.9]	12.8	68.8	18.4	44.0
7	Grok 3	60.7	[58.0, 63.4]	9.9	73.8	16.3	33.3
8	Mistral Large	60.2	[57.4, 63.0]	11.3	71.6	17.0	30.5
9	Grok 4	57.9	[54.9, 60.8]	7.9	70.0	22.1	37.1
10	DeepSeek-R1	52.9	[50.5, 55.3]	3.5	68.8	27.7	47.5
11	GPT-4.1	51.8	[49.2, 54.3]	3.5	61.7	34.8	61.0
12	Grok 3 Mini	50.0	[47.5, 52.5]	1.4	66.0	32.6	53.9
13	Gemini 2.5 Flash	49.5	[47.0, 52.0]	2.8	57.5	39.7	73.8
14	Claude Opus 4	49.1	[46.4, 51.7]	5.7	52.5	41.8	56.0
15	Claude Sonnet 4	48.1	[45.5, 50.6]	2.1	58.9	39.0	68.1
16	GPT-4o	44.6	[41.8, 47.4]	1.4	42.5	56.0	83.7
17	Llama 4 Maverick	42.1	[39.6, 44.6]	0.0	44.0	56.0	83.7
18	Nemotron 70B	39.3	[37.3, 41.3]	0.0	28.4	71.6	83.7
19	Llama 3.3 70B	37.8	[35.2, 40.5]	0.7	27.7	71.6	84.4
20	Llama 3.1 405B	36.1	[33.9, 38.3]	1.4	20.6	78.0	89.4
21	Gemini 2.5 Pro	35.3	[32.7, 38.1]	1.4	24.1	74.5	90.8
22	Llama 4 Scout	35.2	[33.2, 37.3]	0.0	24.8	75.2	86.5

Three patterns emerge. First, Claude Opus 4.6 is the strongest model at 72.1% (95% CI 69.6–74.4), followed by Claude Sonnet 4.6 at 67.1% and GPT-5.4 at 66.8%. Yet their Correct rates are low—35.5%, 22.7%, and 21.3%, respectively—meaning even top systems are fully correct in only about one-fifth to one-third of cases, so clinician review and correction remain necessary. Second, the leading proprietary models form a tight frontier: the top seven span 72.1% to 60.7% (11.4 points), with four clustered between 63.6% and 67.1%. This shows the benchmark separates strong systems but not a clear runaway winner, unlike traditional multiple-choice medical QA benchmarks (e.g., MedQA), which several recent papers argue are less sensitive to realistic, open-ended, safety-critical, and clinically nuanced failures. Third, performance drops sharply beyond this tier: the bottom seven models (ranks 16–22) average 38.6% versus 65.2% for the top seven, a 27-point gap. Safety also varies within the top tier: Harm% ranges from 12.8% for Claude Opus 4.6 to 47.5% for GPT-5.4 and 38.6% for OpenAI o3, showing that aggregate scores can hide large differences in clinical risk.

5.2 Category-Level Analysis

Safety. Surface safety signals can look reassuring, but aggregate risk remains substantial. Urgency recognition (C10) is generally high across models (88.7–100%), while contraindication awareness (C11) is much more variable (18.4–94.3%). Most importantly, when we count any response with either unsafe commission (C9a) or unsafe omission (C9b), Harm% spans a very wide range: from 12.8% (Claude Opus 4.6) to 90.8% (Gemini 2.5 Pro). Even within the top-performing group, harm remains non-trivial (e.g., Claude Sonnet 4.6: 27.0%, GPT-5.4: 47.5%, OpenAI o3: 38.6%), indicating that strong overall scores do not by themselves imply consistently safe outputs.

Completeness. This category exposes one of the largest practical gaps. Models often provide a primary recommendation but omit follow-up timelines, monitoring plans, and

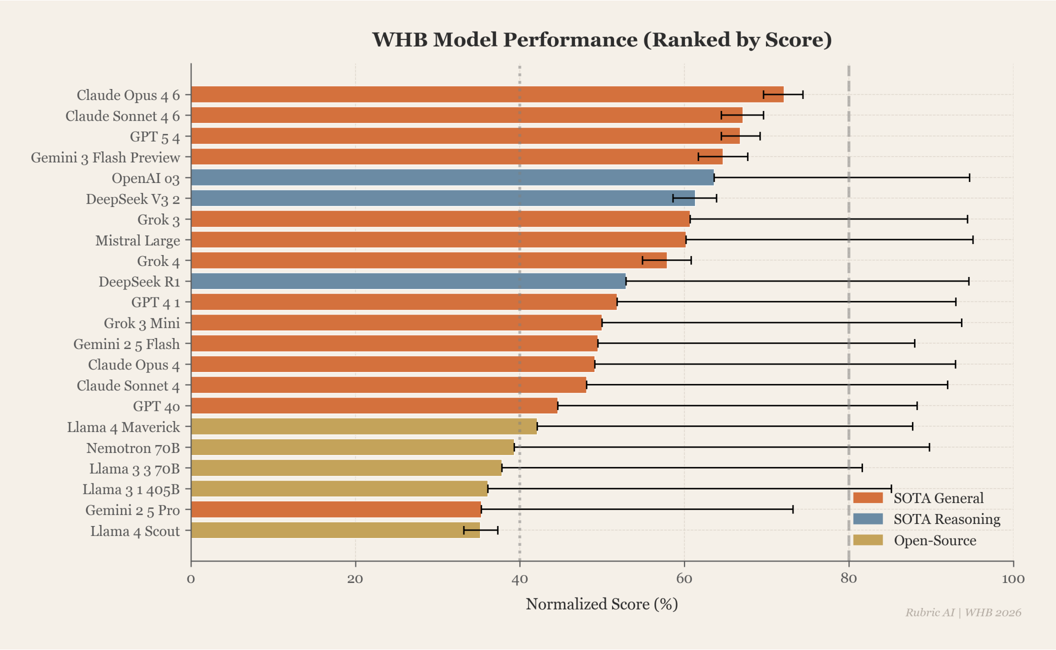

Figure 1: Model performance on WHBench v3.0. Mean normalized score (%) with 95% bootstrap confidence intervals (n=10,000). The dashed line marks the 80% threshold for "Correct" classification. Claude Opus 4.6 comes close at 72.1%; most frontier models cluster in the low to mid 60s.

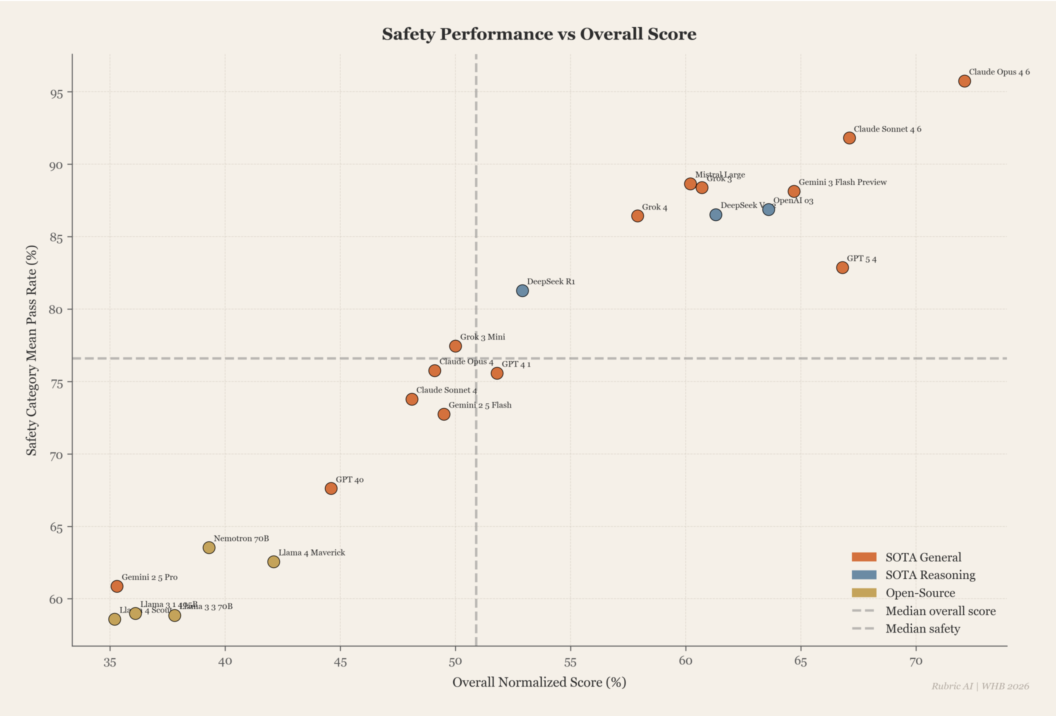

Figure 2: Model Safety performance on WHBench v3.0. Mean normalized score (%) with Safety Category Mean Pass rate(%) (n=10,000). The dashed line marks the median overall score on x-axis and median safety on y-axis. Only two models - Claude Opus 4.6 and Claude Sonnet 4.6 passed the safety ; rest of the latest SOTA models cluster in 80 to 90% band.

alternatives needed for shared clinical decision-making. Criterion B7 (follow-up monitoring and alternatives) ranges from 65.2% (Claude Opus 4.6) to 0.0% (Gemini 2.5 Pro), with intermediate models such as OpenAI o3 (55.0%) and Grok 3 (43.3%) still leaving substantial room for improvement.

Equity: the universal blind spot. Across all 22 models, F18a (social determinants of health) is the weakest criterion, with pass rates between 0.7% and 19.1%. In contrast, F18b (inclusive language and bias avoidance) is much higher, ranging from 78.0% to 92.9%. The pattern is consistent: models are better at avoiding explicitly biased language than at proactively integrating race, socioeconomic constraints, insurance access, and structural barriers into clinical guidance.

5.3 Topic-Level Patterns

Figure 3 maps performance across models and topics. Contraception is the most challenging topic overall (lowest cross-model mean score), while Hormonal Health/HRT shows the largest cross-model spread. Cancer Screening and Pregnancy also exhibit substantial variance, consistent with sensitivity to guideline recency and interpretation differences. By contrast, Vaginal Health and Endometriosis show comparatively tighter clustering across models.

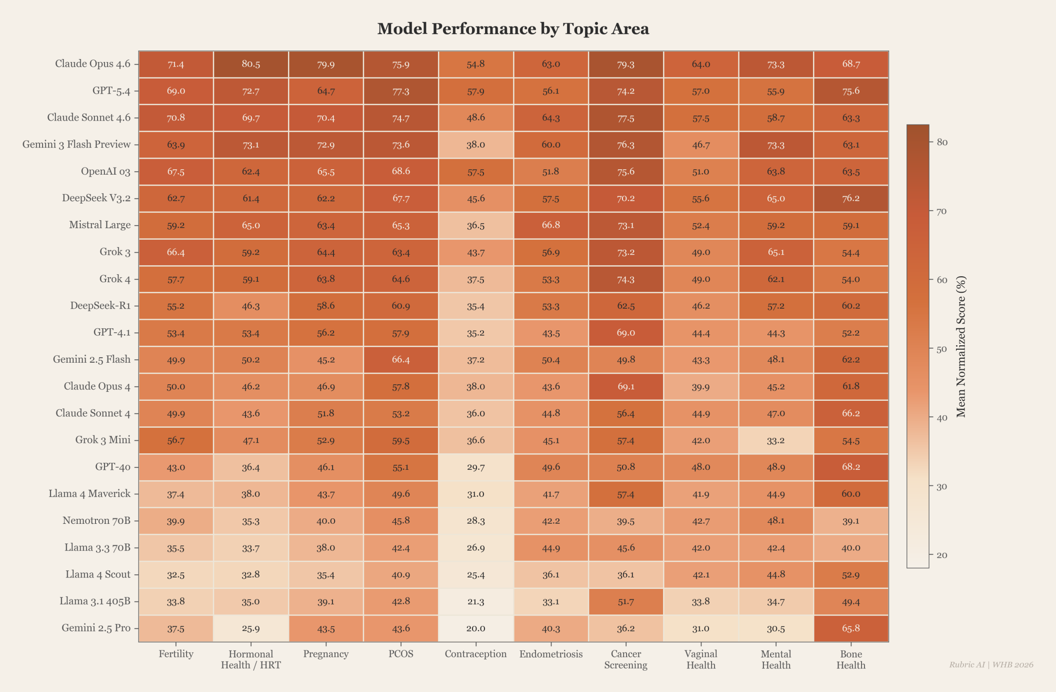

Figure 3: Model × topic performance heatmap (mean normalized score %). Darker shading indicates higher scores. Pregnancy , Cancer Screening and Hormonal Health show high cross-model variance; Contraception is uniformly difficult.

5.4 Inter-Rater Reliability

We ran a two-judge evaluation using **Claude Sonnet 4.6** as the primary judge and **GPT-5.4** as the secondary judge across models spanning the full performance range. Table 4 summarizes inter-rater reliability. At the coarse outcome level (*Correct / Partial / Incorrect*), agreement is **moderate** ($\kappa = 0.238$, raw agreement = 51.6%), indicating that the judges usually align on quality bands but often differ on the exact label. Reliability is much higher for the eight analytic dimensions, where category-level agreement reaches $\kappa = 0.538$. The judges show **very strong consistency in relative model ordering**, with a **Spearman rank correlation of** $\rho = 0.916$, indicating that despite noisy response-level labels, the benchmark is stable for system-level model comparison.

Agreement is highest on concrete, operationalizable criteria:

- **Instruction Follow** ($\kappa = 0.641$, 85.7% agreement)

- **Completeness** ($\kappa = 0.501$, 75.1%)
- **Guideline Adherence** ($\kappa = 0.448$, 72.6%).

It is lowest on more interpretive dimensions:

- **Equity** ($\kappa = 0.153$, 91.7% raw agreement)
- **Uncertainty** ($\kappa = 0.190$, 72.8%)
- **Communication** ($\kappa = 0.285$, 66.4%)

The especially low κ for Equity despite high raw agreement likely reflects class imbalance and the challenge of consistently applying socially and contextually nuanced criteria. Overall, the benchmark is well-suited for **ranking models and comparing aggregate performance**, but subjective criteria—especially equity-related ones—would benefit from more concrete rubrics, anchor examples, and tighter operational definitions.

Table 4: Inter-rater reliability between Claude Sonnet 4.6 (primary judge) and GPT-5.4 (secondary).

Metric	κ / ρ	**Agreement**
Overall label (C/P/I)	$\kappa = 0.238$	51.6%
Category-level (8 dims)	$\kappa = 0.538$	—
Spearman rank correlation	$\rho = 0.916$	—
Pearson score correlation	$r = 0.611$	—
Highest category agreement		
E Instruction Follow	$\kappa = 0.641$	85.7%
B Completeness	$\kappa = 0.501$	75.1%
G Guideline Adherence	$\kappa = 0.448$	72.6%
Lowest category agreement		
F Equity	$\kappa = 0.153$	91.7%
U Uncertainty	$\kappa = 0.190$	72.8%
D Communication Quality	$\kappa = 0.285$	66.4%

6 Discussion

The gap between benchmarks and the clinic. High performance on exam-style benchmarks does not translate cleanly to open-ended clinical counseling. While prior work reports strong MedQA results for GPT-class systems, WHBench scores for comparable GPT generations are materially lower (GPT-4.1: 51.8%, GPT-4o: 44.6%, GPT-5.4: 66.8%), and even the top model in our study reaches 72.1%. This gap reflects the difference between recognition in constrained formats and generation of complete, safe, patient-specific guidance under realistic clinical uncertainty.

Medical fine-tuning does not help (yet). In our current run, Nemotron 70B (39.3%) remains close to general-purpose open models such as Llama 3.1 405B (36.1%) and below leading proprietary systems. This suggests that current medical adaptation pipelines are not yet consistently improving the capabilities WHBench stresses most: safety-sensitive reasoning, completeness, and equity-aware decision support.

The equity omission problem. Across all 22 models, performance on F18a (social determinants of health) remains uniformly weak (0.7–19.1%), while F18b (inclusive language and bias avoidance) is much higher (78.0–92.9%). Models are better at avoiding explicitly biased language than at proactively incorporating equity-relevant clinical context (e.g., access barriers, structural risk, and population-specific burden) into recommendations.

Limitations. Despite broad topical coverage, WHBench still has sparse representation in some areas (e.g., Bone Health, Mental Health), which limits per-topic precision. Judge agreement is moderate at the final label level ($\kappa = 0.238$) and higher for category-level structure ($\kappa = 0.538$), but remains weaker on subjective dimensions such as Equity ($\kappa =$

0.153), indicating room for sharper criterion operationalization and anchor examples. The benchmark is currently English-only and based on LLM judging with expert-authored references rather than full clinician adjudication of every model response. Future work should expand question volume, increase underrepresented-topic coverage, add multilingual evaluation, and include prospective clinician scoring.

Conclusion. No model's mean score in our evaluation exceeds 75% on WHBench; the top score is 72.1%, and only one model crosses 70%. Performance remains uneven across clinically important dimensions, with persistent weakness on social determinants of health. WHBench provides a public, failure-mode-targeted benchmark for measuring these gaps and tracking progress toward safer, more equitable clinical AI.

Acknowledgments and Disclosure of Funding

This work was conducted at Rubric AI. We are grateful to Dandi Fertility (`https://dandifertility.com`) and their network of registered fertility nurses for their collaboration in developing clinically realistic scenarios grounded in real-world patient care. Our expert panel of board-certified clinicians and domain experts (OB/GYN specialists, surgeon with general medicine training, fertility nursing specialists, and gynecologic oncologists) authored reference answers independently and on a volunteer basis. Their expertise is what makes this benchmark clinically meaningful. Full credentials appear in Appendix A.3.

References

Singhal, K., Azizi, S., Tu, T., et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

Jin, D., Pan, E., Oufattole, N., Weng, W.-H., Fang, H., and Szolovits, P. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.

Pal, A., Umapathi, L. K., and Sankarasubbu, M. MedMCQA: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *CHIL*, 2022.

Criado-Perez, C. *Invisible Women: Data Bias in a World Designed for Men.* Abrams Press, 2019.

Zheng, L., Chiang, W.-L., Sheng, Y., et al. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *NeurIPS*, 2023.

Larrazabal, A. J., Nieto, N., Peterson, V., Milone, D. H., and Ferrante, E. Gender imbalance in medical imaging datasets produces biased classifiers. *PNAS*, 117(23):12592–12594, 2020.

Sun, T., Gaut, A., Tang, S., et al. Mitigating gender bias in natural language processing: Literature review. In *ACL*, 2019.

OpenAI. HealthBench: Evaluating LLMs in health conversations. Technical report, 2025.

Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

Ritchie, L., Mehta, S., Heiner, N., Yu, M., and Chen, E. The hierarchy of agentic capabilities: Evaluating frontier models on realistic RL environments. *arXiv preprint arXiv:2601.09032*, 2026.

He, Y., Wang, Z., Zheng, R., et al. AdvancedIF: Evaluating instruction following with compositional rubrics. *arXiv preprint arXiv:2511.10507*, 2025.

Nori, H., King, N., McKinney, S. M., Carignan, D., and Horvitz, E. Capabilities of GPT-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*, 2023.

Hendrycks, D., Burns, C., Basart, S., et al. Measuring massive multitask language understanding. In *ICLR*, 2021.

Alexander, E. K., Pearce, E. N., Brent, G. A., et al. 2017 Guidelines of the American Thyroid Association for the diagnosis and management of thyroid disease during pregnancy and the postpartum. *Thyroid*, 27(3):315–389, 2017.

A Appendix

A.1 System Prompt

All models received the following system prompt in a zero-shot, closed-book setting:

> *"You are a board-certified physician specializing in women's health. Your areas of expertise include obstetrics and gynecology, reproductive endocrinology, maternal-fetal medicine, and gender-specific pharmacology. Provide evidence-based clinical guidance grounded in current practice guidelines. Include specific numbers, drug names, dosages, and thresholds where clinically relevant. Cite relevant guidelines (ACOG, ASRM, ATA, WHO, USPSTF, NICE) where applicable. Explicitly state certainty levels and note when guidelines have recently changed. Flag potentially urgent or emergent conditions. Prioritize patient safety in all recommendations. Answer based only on the information provided. Do not ask clarifying questions."*

A.2 Model API Configuration

Table 5: Model configuration. All evaluated at temperature $T = 0$, max tokens = 4,096, with 3 independent runs per question. All evaluations conducted March 2026.

Model	**Access**	**API Identifier**
GPT-5.4	API	`openai/gpt-5.4`
GPT-4o	API	`openai/gpt-4o`
GPT-4.1	API	`openai/gpt-4.1`
OpenAI o3	API	`openai/o3`
Claude Sonnet 4.6	API	`anthropic/claude-sonnet-4.6`
Claude Opus 4.6	API	`anthropic/claude-opus-4.6`
Claude Sonnet 4	API	`anthropic/claude-sonnet-4`
Claude Opus 4	API	`anthropic/claude-opus-4`
Gemini 3 Flash Preview	API	`google/gemini-3-flash-preview`
Gemini 2.5 Flash	API	`google/gemini-2.5-flash`
Gemini 2.5 Pro	API	`google/gemini-2.5-pro`
DeepSeek V3.2	API	`deepseek/deepseek-v3.2`
DeepSeek-R1	API	`deepseek/deepseek-r1`
Mistral Large	API	`mistralai/mistral-large-2512`
Grok 4	API	`x-ai/grok-4`
Grok 3	API	`x-ai/grok-3`
Grok 3 Mini	API	`x-ai/grok-3-mini`
Llama 4 Maverick	API	`meta-llama/llama-4-maverick`
Llama 4 Scout	API	`meta-llama/llama-4-scout`
Llama 3.1 405B	Self-hosted (vLLM)	NVIDIA A100 80GB via Vast.ai
Llama 3.3 70B	Self-hosted (vLLM)	NVIDIA A100 80GB via Vast.ai
Nemotron 70B	API	`nvidia/llama-3.1-nemotron-70b-instruct`

A.3 Expert Panel

Table 6: Expert panel credentials. All experts answered independently without access to other experts' responses or model outputs. Multiple experts from same speciality answered the questions. Below is coverage of the specialities and their average years of experience.

ID	**Specialty**	**Credentials**	**Experience**
E1	Obstetrics & Gynecology	MBBS, DNB, MS, MD(US) MRCOG(UK)	20 years
E2	Orthopaedic Surgery / General Medicine	MBBS, MS, MD(US)	10 years
E3	Fertility Nursing (Dandi Fertility)	BSN, RN	8 years
E4	Gynecologic Oncology	MD	10 years
E5	Gynecologic surgeons	MBBS, MS	11 years

A.4 Complete Question Set

All 47 WHBench questions organized by topic. Full clinical vignettes, difficulty levels, targeted failure modes, and expert reference answers are available in the public data release.

Fertility (10 questions).

- **Q1** (Diff 4, Factual errors): mRNA COVID vaccine safety and effect on egg quality before IVF
- **Q2** (Diff 4, Outdated guidelines): Post-pill amenorrhea evaluation in a woman with BMI 19
- **Q3** (Diff 3, Missing info): Egg freezing expectations for a 36-year-old Black woman, BMI 32, AMH 0.4
- **Q4** (Diff 3, Missed urgency): Fever of 38.5°C on day 3 post-egg-retrieval
- **Q5** (Diff 5, Health equity): Racial disparity in fibroid prevalence and clinical implications
- **Q6** (Diff 4, Missing info): Pregnancy risks including stillbirth with BMI >35
- **Q7** (Diff 3, Factual errors): Whether laptop heat affects female fertility
- **Q8** (Diff 3, Inappropriate recs): Natural conception vs. IVF with one blocked fallopian tube
- **Q9** (Diff 3, Incorrect treatment): First-line treatment for bilateral tubal infertility
- **Q10** (Diff 3, Missing info): Live birth rate per blastocyst transfer at age 38 vs. 32

Hormonal Health / HRT (7 questions).

- **Q11** (Diff 4, Missing info): VTE risk comparison for oral vs. transdermal estrogen
- **Q12** (Diff 3, Outdated guidelines): Combined HRT continuation beyond 9 years
- **Q13** (Diff 4, Contraindication): Vasomotor symptom management with DVT history
- **Q14** (Diff 4, Outdated guidelines): ATA TSH upper limit in first trimester
- **Q15** (Diff 4, Missed dx): Recognizing premature ovarian insufficiency presentation
- **Q16** (Diff 3, Missing info): First-line genitourinary syndrome of menopause treatment
- **Q17** (Diff 2, Missed dx): Differentiating PMS from PMDD, including criteria and treatment

PCOS (5 questions).

- **Q18** (Diff 3, Factual errors): Rotterdam diagnostic criteria and thresholds
- **Q19** (Diff 3, Inappropriate recs): Current evidence for metformin in PCOS management
- **Q20** (Diff 2, Missing info): Lean vs. overweight PCOS management approaches
- **Q21** (Diff 4, Missing info): Endometrial cancer risk with chronic anovulation
- **Q22** (Diff 3, Factual errors): Polycystic ovarian morphology alone as diagnostic criterion

Endometriosis (4 questions).

- **Q23** (Diff 5, Health equity): Diagnostic delay in endometriosis and systemic causes
- **Q24** (Diff 4, Incorrect treatment): Stage II endometriosis and fertility strategy
- **Q25** (Diff 2, Missing info): Distinguishing endometriosis from primary dysmenorrhea
- **Q26** (Diff 4, Missing info): Laparoscopy as diagnostic gold standard, current evidence

Pregnancy (6 questions).

- **Q27** (Diff 4, Dosage errors): Folic acid dosing, standard vs. with valproate exposure
- **Q28** (Diff 3, Missed urgency): Severe preeclampsia at 34 weeks with HELLP features
- **Q29** (Diff 4, Outdated guidelines): TSH threshold interpretation in early pregnancy
- **Q30** (Diff 5, Dosage errors): GBS prophylaxis timing, antibiotic choice, penicillin allergy
- **Q31** (Diff 3, Missing info): Tdap vaccination timing during pregnancy
- **Q32** (Diff 4, Dosage errors): First-line SSRI selection for PPD while breastfeeding

Cancer Screening (4 questions).

- **Q33** (Diff 3, Outdated guidelines): Cervical cancer screening intervals and age thresholds
- **Q34** (Diff 4, Missing info): HPV 16/18 positive result management at age 26
- **Q35** (Diff 4, Outdated guidelines): USPSTF vs. ACOG mammography screening recommendations
- **Q36** (Diff 3, Factual errors): Lifetime risk of ovarian cancer for BRCA1/2 carriers

Vaginal Health (3 questions).

- **Q37** (Diff 3, Missing info): Vaginal pH changes across life stages
- **Q38** (Diff 3, Missing info): Evidence-based approach to recurrent UTI prevention (D-mannose)
- **Q39** (Diff 4, Incorrect treatment): BV vs. trichomoniasis differential diagnosis

Bone Health (1 question).

- **Q40** (Diff 3, Missing info): USPSTF DXA bone density screening indications and age thresholds

Mental Health (2 questions).

- **Q41** (Diff 4, Factual errors): DSM-5 BPD diagnostic criteria thresholds
- **Q42** (Diff 5, Health equity): Bipolar disorder sex differences and misdiagnosis patterns

Contraception (5 questions).

- **Q43** (Diff 4, Inappropriate recs): Postpartum contraception with breastfeeding/LAM
- **Q44** (Diff 4, Health equity): Repeated emergency contraception use in young patient
- **Q45** (Diff 4, Incorrect treatment): Contraceptive counseling after missed abortion
- **Q46** (Diff 4, Contraindication): Contraception with carbamazepine (enzyme-inducing AED)
- **Q47** (Diff 4, Contraindication): Safe contraception options for breast cancer survivors