PREPRINT

Not Peer Reviewed

RESEARCH ARTICLE

Evaluating Frontline Health Workers' Responses to Patient Inquiries With and Without Large Language Model Support in Nigeria: An Observational Study

[version 1]

Maria Moosa[1], Oluwaseyi Malumi[1], Solomon Chinedu[1], Wilson Okah[1], Ayoposi Ogboye[2], Chiagozie Abiakam[1], Peter Onyenemerem[1], Imo Etuk[1], Nneka Mobisson[1]

[1]mDoc Healthcare, Lekki, Lagos, Nigeria
[2]British Institute in Eastern Africa, London, SW1Y 5AH, UK

v1 **First published:** 07 Oct 2025, **2**:324
https://doi.org/10.12688/verixiv.2125.1
Latest published: 07 Oct 2025, **2**:324
https://doi.org/10.12688/verixiv.2125.1

Abstract

Frontline health workers (FLWs) in low- and middle-income countries (LMICs) often face barriers that compromise care quality, including limited training, high patient loads, inadequate access to updated clinical guidance, and resource constraints. Working in underserved settings further exacerbates challenges to providing accurate, complete, and empathetic responses. Addressing these gaps requires innovative tools to support FLWs in real time. This study evaluates whether large language models (LLMs) can enhance the quality of FLWs' responses to health inquiries in Nigeria.

In this cross-sectional study, 36 licensed FLWs (doctors, nurses, community health workers) practicing general or primary care in Lagos generated responses to 15 patient-generated health questions covering maternal and neonatal health, family planning, and sexually transmitted infections. Each FLW produced responses using their own knowledge and resources ("human-only") and with ChatGPT-3·5 assistance ("GPT-aided"). A panel of clinicians evaluated responses using a 5-point Likert scale for accuracy, empathy, contextualization, completeness, safety, and overall preference. In addition, eight non-clinician health seekers assessed a subset of responses for trustworthiness, clarity, and perceived care quality.

Across 1,080 responses, ChatGPT-aided responses significantly outperformed human-only responses on all clinician-rated metrics ($p<0{\cdot}001$), with the largest gains in completeness, empathy, and overall preference, particularly among non-physician FLWs. Non-clinician reviewers also preferred ChatGPT-aided responses, especially for trustworthiness. The findings highlight LLMs' potential to improve quality, empathy, and trustworthiness of FLWs' responses, particularly in low-resource settings. There is need to explore safe integration of LLMs into clinical workflows, alongside capacity-building and further evaluation to ensure effectiveness, equity, and patient safety.

This study provides empirical evidence that a large language model enhances accuracy, empathy, and trustworthiness of responses by frontline health workers in Nigeria. These findings underscore the potential of LLM integration to strengthen healthcare delivery in resource-limited settings, highlighting the necessity for further investigation into safe, equitable, and effective implementation.

Keywords
ChatGPT, LLMs, LMIC, AI, Frontline Health Workers (FLWs), Digital health, Nigeria

This article is included in the Gates Foundation gateway.

Corresponding author: Oluwaseyi Malumi (seyi.malumi@mymdoc.com)

Author roles: Moosa M: Conceptualization, Data Curation, Methodology, Visualization, Writing – Review & Editing; **Malumi O**: Investigation, Project Administration, Supervision, Visualization, Writing – Review & Editing; **Chinedu S**: Investigation, Methodology, Project Administration, Resources, Supervision, Validation, Visualization, Writing – Review & Editing; **Okah W**: Data Curation, Software, Visualization, Writing – Review & Editing; **Ogboye A**: Writing – Original Draft Preparation, Writing – Review & Editing; **Abiakam C**: Writing – Review & Editing; **Onyenemerem P**: Writing – Review & Editing; **Etuk I**: Conceptualization, Funding Acquisition, Supervision, Writing – Review & Editing; **Mobisson N**: Conceptualization, Funding Acquisition, Supervision, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work was supported, in whole or in part, by the Gates Foundation [INV-062594]. The conclusions and opinions expressed in this work are those of the author(s) alone and shall not be attributed to the Foundation. Under the grant conditions of the Foundation, a Creative Commons Attribution 4.0 License has already been assigned to the Author Accepted Manuscript version that might arise from this submission

How to cite this article: Moosa M, Malumi O, Chinedu S *et al.* **Evaluating Frontline Health Workers' Responses to Patient Inquiries With and Without Large Language Model Support in Nigeria: An Observational Study [version 1]** VeriXiv 2025, **2**:324 https://doi.org/10.12688/verixiv.2125.1

First published: 07 Oct 2025, **2**:324 https://doi.org/10.12688/verixiv.2125.1

Introduction

Artificial intelligence (AI) is increasingly being explored as a tool to help healthcare workers provide services to their patients, particularly in regions where poor health systems, healthcare workforce shortages, and infrastructure limitations pose significant challenges, such as sub-Saharan Africa.[1] Several studies have objectively compared the quality of care provision at health facilities across different African countries. These studies consistently highlight persistent challenges such as inadequate infrastructure, workforce shortages, and inconsistent service delivery that undermine the quality of care.[2–4] Additionally, people often express dissatisfaction with the quality of care at health facilities; factors such as prolonged wait times, lack of equipment, among others, contribute to their dissatisfaction and the sub-par quality of care they receive.[5] Systematic reviews from Kenya and other countries highlight deficits in clinical care, patient safety, and outcomes, underscoring the urgent need for innovative approaches to strengthen service delivery.[3,4] These findings provide important context for exploring how emerging technologies like artificial intelligence (AI) can support healthcare workers in improving service delivery, particularly in resource-limited settings.

In Nigeria, structural barriers including economic instability, limited infrastructure, and a shortage of skilled healthcare workers (1·83 per 1,000 people, far below the WHO-recommended 4·45) contribute to delayed access to preventive care and high rates of avoidable morbidity.[6–9]Amidst these challenges, AI, specifically large language models (LLMs), show promise in digital health's role in potentially lessening systemic barriers.[10] LLMs are being deployed for various uses in healthcare delivery, along with their own risks and rewards.[11] Due to the vast amount of data upon which they are trained (and later, fine-tuned for specific applications), these models lead in their pattern discernment, predictive capacity, and ability to understand context as they successfully carry out tasks and prompts.[12] LLMs have aided clinical workflow through text and concept extraction, summarization and synthesis, automating communication, drafting relevant treatment plans, as a supplementary educational resource, and much more.[13] They have the potential to reduce workload, optimize performance, and fundamentally reshape clinical practice.[14] Nonetheless, several concerns regarding hallucinations, bias, and lack of specialised skills to use them, highlight the necessity of rigorous evaluation to safeguard patient safety.[15]

Therefore, its effectiveness and safety in responding to patient questions in LMICs have not been thoroughly evaluated. Despite this, ChatGPT has attracted growing interest from healthcare professionals, including those in LMICs such as Nigeria, often without clear information about its risks, benefits, or limitations.[19]

ChatGPT, one of the most widely used LLMs with over 300 million weekly active users as of January 2025,[16,17] generates near-human-quality text but was not designed for healthcare.[1,18] Its use is growing among clinicians in LMICs despite limited evidence on safety, effectiveness, and integration into frontline workflows.[19] While discourse often focuses on AI replacing healthcare workers, emerging research suggests AI may be better suited to complement and support them.[20,21]

To evaluate the potential of LLMs as a complementary tool in frontline workflows, this study compared the effectiveness of FLWs in Lagos, Nigeria, when responding to women's health-related questions—covering maternal and neonatal health, family planning, and sexually transmitted infections—using standard clinical resources versus ChatGPT-3·5. The study forms part of a broader mDoc Healthcare initiative on integrating LLMs into women-centered care.[22] Findings from this work provide early insights into how LLMs might be safely and effectively incorporated into clinical workflows in low-resource settings.

Methods

Study design

This cross-sectional study evaluated how using ChatGPT-3·5, the free public version last trained in 2022, influenced FLWs' responses to common primary care questions from health seekers in Lagos, Nigeria. Responses generated with ChatGPT support ("GPT-aided") were compared with those produced using standard clinical resources ("human-only"). The study also examined both clinicians' and non-clinicians' preferences for GPT-aided versus traditional responses. Participants accessed ChatGPT-3·5 through their own devices and OpenAI accounts. Throughout this study, "GPT" refers specifically to this version, and "participants" refers to the three FLW groups.

Ethical approval was obtained from the Lagos State Health Research Ethics Committee (LSHREC/2023/0011) and the FCT Health Research Ethics Committee (FHREC/2023/01/140/01-08-23). Written informed consent was obtained from all individuals who took part in the study.

Setting and participants

The study enrolled 36 licensed frontline health workers (FLWs) in Lagos, Nigeria: 12 medical doctors (MDs), 12 registered nurses (RNs), and 12 community health extension workers (CHEWs). All participants were actively

practicing general medicine or primary care in public or private facilities. Eligibility required professional licensure: MDs (MBBS/MBChB degrees and registration with Medical and Dental Council of Nigeria), RNs (Bachelor of Nursing Science B.N.Sc/Diploma or Ordinary National Diploma and RN certificate from the Nursing and Midwifery Council of Nigeria or) and CHEWs (National Diploma from School of Health Technology, Community Health Practitioners Registration Board of Nigeria).[23] FLWs in specialized roles were excluded. Recruitment was conducted via open calls, and the study was implemented in person in Lagos.

Question selection

In March 2024, 15 frequently asked questions were drawn from conversations between mDoc health coaches and women of reproductive age during an outreach event in Lagos. mDoc is a digital health social enterprise based in Lagos, Nigeria, that provides end-to-end, AI-enabled self-care health coaching to over 153,000 low-income members, more than 80% of whom are women. These questions spanned four health topic areas (maternal and neonatal health, family planning, and sexually transmitted infections). They reflected authentic, recurring health concerns raised during ongoing coaching interactions, ensuring that the study addressed issues directly relevant to women's lived experiences. The restriction to these health topics is in line with the health focus of the larger study within which this study is nested, and reflective of the commonly asked health queries by mDoc's members (Figure 1).

Measurements & Sample size justification

Each FLW answered the 15 questions twice: once using their standard resources ("human-only") and once with ChatGPT-3·5 assistance ("GPT-aided"), yielding 30 responses per participant and 1,080 responses overall (36 FLWs × 30 responses). Participants were instructed to ensure responses were accurate, empathetic, safe, complete, and contextually relevant to Nigeria. This resulted in 30 responses per participant and 1,080 total responses (36 FLWs × 30). The target sample size of 1,080 responses was determined based on *a priori* power calculations to detect a 10-percentage point difference in key response metrics (Accuracy, Empathy, Safety, Contextualization and Completeness, Overall Preference) using at least a 50% correct response for each metric for either human-only and GPT-aided responses.[24] Using a two-sided alpha of 0·05 and 80% power, approximately 776 responses (388 per condition) were required to identify the specified difference; our sample exceeded this, allowing robust comparisons. After completing their responses, participants filled out a post-survey to reflect on their experience, including ease of use, perceived usefulness, and trust in ChatGPT versus traditional tools. The survey used a 5-point Likert scale, with 1 representing *strongly disagree* and 5 representing *strongly agree*.

Evaluation and outcomes

To assess the quality and perceived value of FLWs' responses, with and without ChatGPT assistance, a dual-panel evaluation was conducted using clinical experts and non-clinicians (Table 1).

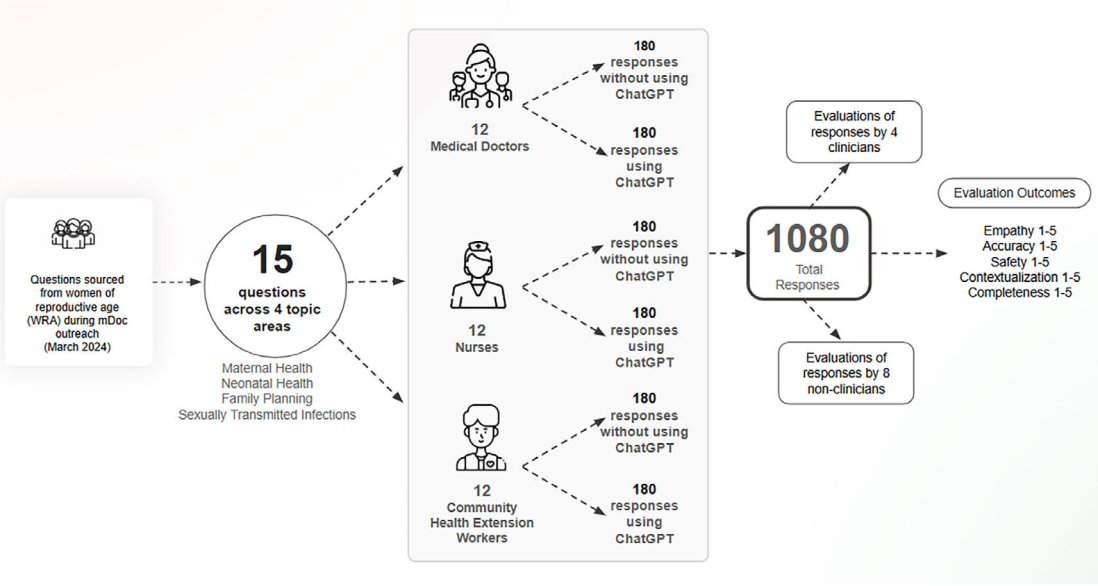

Figure 1. Process for question selection and outcome analysis.

Table 1. Demographic characteristics of evaluators.

Clinical Evaluators:		
Demographics	**Range/Category**	**Frequency (n = 4)**
Age Range	35-44	2 (50·0%)
	45-54	1 (25·0%)
	55 years and above	1 (25·0%)
What is the highest academic degree you have earned?	Bachelors (MBBS, BN.Sc, etc.)	4 (100%)
What is your specialty?	Paediatrics	1 (25·0%)
	Nursing	1 (25·0%)
	Family Medicine	1 (25·0%)
	Obstetrics and Gynecology	1 (25·0%)
Years of Experience (mean)		18 (9·07)
Non-Clinician Evaluators:		
Demographics	**Category**	**Frequency (n = 8)**
Age Range	18-24	1 (12·5%)
	25-34	5 (62·5%)
	35-44	2 (25·0%)
What is the highest academic degree you have earned?	No Degree	2 (25·0%)
	Bachelors (MBBS, BN.Sc, etc.)	6 (75·0%)

Clinical panel: This panel consisted of four licensed experts (three physicians and one nurse) who assessed the responses via a structured rubric covering six domains: Accuracy, Empathy, Contextualization, Completeness, Safety, and Overall Preference. All but Accuracy included a "Not Applicable" option. Ratings used a 5-point Likert scale with written justifications. Before evaluation, reviewers attended a virtual orientation to standardize application of the rubric. Responses were anonymized, randomized, and blinded by response type. Reviewers had five days to complete scoring.

Non-clinical panel: To capture user-centered insights, eight randomly selected non-clinicians from mDoc's member base assessed a subset of 256 responses (128 GPT-aided and 128 human-only). Using a paired-comparison format, they judged which response was more caring, trustworthy, and easier to understand. This aimed to reveal broad preference trends across FLW types (doctors, nurses, CHEWs), without testing for small effects.

Clinical scores were aggregated using ensemble scoring to reflect consensus while capturing rater variability.

Statistical analysis

Computational linguistics analyses were conducted to assess key textual features of the responses. Word counts were generated using the Python pandas package, and readability was measured with the Flesch-Kincaid Grade Level via textstat. Sentiment analysis evaluated polarity and subjectivity across response types. The demographics for the FLWs were summarized using counts and percentages for categorical variables. Chi-squared tests were used to compare the three groups on categorical variables such as age range and gender, with Fisher's exact tests applied when cell counts were below 5. Continuous metric data (Accuracy, Safety, etc.) were summarized using median and interquartile range. The decision to use non-parametric methods was based on the data distribution showing non-normality after using the Shapiro-Wilk tests and histograms. For comparisons between each metric and groups of FLWs under human-only or GPT-aided responses, the Kruskal-Wallis test was used to account for non-normality. We further visualized the data using violin plots and bar charts with standard error bars. Wilcoxon signed-rank tests were employed to assess significant differences in preferences between human-only and GPT-aided responses within each professional group. To assess the degree of agreement among multiple independent clinical raters evaluating response quality, Fleiss' Kappa statistics were computed for each evaluation metric. All analyses were conducted in Python version 3·13 and STATA version 18, with significance levels set at $p < 0·001$, and results were visualized to enhance interpretability.

Results

The majority of participants were female (69·4%), with a statistically significant difference in gender distribution across FLW groups ($p = 0·026$) (Table 2). Age distribution also varied significantly ($p = 0·030$), with most MDs (91·7%) falling

Table 2. Demographic characteristics of FLW participants.

	CHEWs	MD	RN	Total	P
N	12 (33·3%)	12 (33·3%)	12 (33·3%)	36 (100·0%)	
Age Range					0·030*
18-24	0 (0·0%)	1 (8·3%)	4 (33·3%)	5 (13·9%)	··
25-34	8 (66·7%)	11 (91·7%)	6 (50·0%)	25 (69·4%)	··
35-44	4 (33·3%)	0 (0·0%)	2 (16·7%)	6 (16·7%)	··
Gender					0·026*
Female	9 (75·0%)	5 (41·7%)	11 (91·7%)	25 (69·4%)	··
Male	3 (25·0%)	7 (58·3%)	1 (8·3%)	11 (30·6%)	··
What is the highest academic degree you have earned?					0·003**
Diploma	7 (58·3%)	0 (0·0%)	4 (33·3%)	11 (30·6%)	··
Ordinary National Diploma ND	2 (16·7%)	0 (0·0%)	3 (25%)	5 (13·9%)	··
Bachelors (MBBS, BN·Sc, etc·)	3 (25·0%)	12 (100·0%)	5 (41·7%)	20 (55·6%)	··
How many years have you been practicing?					0·057 ns
1-5	8 (66·7%)	12 (100·0%)	10 (83·3%)	30 (83·3%)	··
5-10	2 (16·7%)	0 (0·0%)	0 (0·0%)	2 (5.6%)	··
11-15	2 (16·7%)	0 (0·0%)	2 (16·7%)	4 (11·1%)	··

p-values are annotated based on significance.
*$p < 0·05$.
**$p < 0·001$, and "ns" if $p \geq 0·1$.

within the 25–34 age range, while CHEWs were more commonly aged 35–44 years. Educational attainment differed significantly between groups ($p = 0·003$): 41·7% of RNs held bachelor's degrees compared to just 25% of CHEWs. Years of experience also varied, but was not statistically significant ($p = 0·057$): CHEWs had the highest mean ($5·9 \pm 3·8$), followed by RNs ($4·3 \pm 4·2$) and MDs ($2·5 \pm 1·3$). Most participants (83%) had between 1–5 years of experience, with only six exceeding five years. Overall, two-thirds of participants (66·7%, n = 24) reported having conducted formal telemedicine consultations. Additionally, half of the respondents (50·0%, n = 18) stated that they typically answer patient queries in under five minutes (Table 3).

Post-survey results revealed diverse insights into participant experiences and perceptions with ChatGPT (Table 3). Most rated themselves as skilled or very skilled in gathering clinical information, and 52·8% (n = 19) had prior LLM experience, mainly with ChatGPT (90·5%). However, only 27·8% (n = 10) had used it for clinical tasks, while 41·7% had no experience at all (n = 15). All agreed on the importance of evidence-based responses. They found ChatGPT user-friendly, easy to navigate, and understood inputs well, and that it enhanced response speed, accuracy, and completeness. Still, MDs and RNs were more likely to disagree with ChatGPT responses compared to CHEWs ($p = 0·037$) and trusted their usual clinical resources more ($p = 0·008$). While most FLWs were open to using ChatGPT, the findings suggest participants balanced the tool's benefits with caution and reliance on established practices, reflecting measured enthusiasm for integrating AI into their workflows.

Clinician evaluator scores analysis

Figure 2 shows the score distributions for GPT-aided (green) and human-only (orange) responses across six evaluation metrics. For Accuracy and Safety, both GPT-aided and human-only were rated highly, with scores clustering around 4 and 5; however, GPT-aided responses show a slightly wider distribution toward the upper end, suggesting greater preference. Empathy scores were widely spread for both conditions, with GPT-aided responses being slightly more concentrated at a higher rate. Contextualization and Completeness GPT-aided and human-only responses scored similarly around 4, though GPT-aided responses display a wider spread towards 5, suggesting a slight preference.

GPT-aided responses were consistently rated higher overall, with notable improvements in Empathy, Completeness, Contextualization, and Overall Preference (Table 4). Safety remained the highest-rated metric across all groups and conditions (median = 4·00, IQR = 1·00), showing minimal change with GPT support, suggesting it was already a strength among respondents. In contrast, Completeness, Contextualization, and Empathy saw the greatest improvements with

Table 3. **Participants' experiences and perceptions of the use of ChatGPT.**

	CHW	MD	RN	Total	P
N	12 (33·3%)	12 (33·3%)	12 (33·3%)	36 (100·0%)	
Have you previously ever used a large language model?					0·343
I don't know	1 (8·3%)	1 (8·3%)	2 (16·7%)	4 (11·1%)	··
No	5 (41·7%)	2 (16·7%)	6 (50·0%)	13 (36·1%)	··
Yes	6 (50·0%)	9 (75·0%)	4 (33·3%)	19 (52·8%)	··
If yes, which large language models have you used?					0·314
ChatGPT	5 (83·3%)	9 (100·0%)	5 (83·3%)	19 (90·5%)	··
Speech recognition	1 (16·7%)	0 (0·0%)	0 (0·0%)	1 (4·8%)	··
Igbo/Hausa	0 (0·0%)	0 (0·0%)	1 (16·7%)	1 (16·7%)	
If you have ever used large language models (e·g· ChatGPT) have you ever used them for medical or clinical purposes?					0·028
I have never used Large Language Models	7 (58·3%)	1 (8·3%)	7 (58·3%)	15 (41·7%)	··
ChatGPT but not for medical purposes	4 (33·3%)	6 (50·0%)	1 (8·3%)	11 (30·6%)	··
ChatGPT for medical purposes	1 (8·3%)	5 (41·7%)	4 (33·3%)	10 (27·8%)	··
If yes, how many years have you been practicing?	5·917 (3·825)	2·467 (1·305)	4·283 (4·240)	4·222 (3·581)	0·057
Telemedicine is the use of electronic and telecommunications equipment for two-way					0·223
No	6 (50·0%)	2 (16·7%)	4 (33·3%)	12 (33·3%)	··
Yes	6 (50·0%)	10 (83·3%)	8 (66·7%)	24 (66·7%)	··
On average, how much time does it take you to answer a patient's question?					0·570
0-5 minutes	6 (50·0%)	7 (58·3%)	5 (41·7%)	18 (50·0%)	··
11-15 minutes	1 (8·3%)	1 (8·3%)	0 (0·0%)	2 (5·6%)	··
16-20 minutes	2 (16·7%)	0 (0·0%)	1 (8·3%)	3 (8·3%)	··
6-10 minutes	2 (16·7%)	3 (25·0%)	6 (50·0%)	11 (30·6%)	··
More than 20 minutes	1 (8·3%)	1 (8·3%)	0 (0·0%)	2 (5·6%)	··
Do you answer health questions for health seekers/patients through digital channels (e.g., answering questions via SMS/WhatsApp/Telegram/ Email, etc·)?	4·000 (1·348)	3·583 (1·311)	4·083 (0·793)	3·889 (1·166)	0·544
How important do you feel it is for licensed health care professionals to refer to evidence often when answering patient questions?	4·667 (0·888)	4·667 (0·778)	4·167 (1·267)	4·500 (1·000)	0·379

Table 3. *Continued*

	CHW	MD	RN	Total	P
On average, how often do you refer to external resources before answering patient/health seekers' questions via digital channels?	3·083 (1·379)	3·000 (0·739)	3·417 (0·669)	3·167 (0·971)	0·552
Please rate your level of skill in obtaining/gathering clinical information	4·417 (0·996)	4·333 (0·651)	4·083 (0·515)	4·278 (0·741)	0·532
Based on your experience, how common were the questions asked in the survey?	4·583 (0·669)	4·250 (0·754)	3·917 (1·084)	4·250 (0·874)	0·177
Based on your experience, how difficult were the questions in the survey?	4·583 (0·515)	3·500 (1·087)	3·833 (0·718)	3·972 (0·910)	0·008
ChatGPT was easy to navigate	4·667 (0·888)	4·750 (0·452)	4·417 (0·996)	4·611 (0·803)	0·585
When using ChatGPT, it would be easy to get confused	1·333 (0·888)	1·250 (0·452)	1·750 (0·965)	1·444 (0·809)	0·275
ChatGPT understood my inputs well	4·333 (1·231)	4·500 (0·674)	4·417 (0·793)	4·417 (0·906)	0·909
ChatGPT failed to recognize a lot my inputs and prompts	1·667 (1·231)	1·333 (0·492)	1·750 (0·866)	1·583 (0·906)	0·505
ChatGPT's responses were useful	4·583 (1·165)	4·583 (0·515)	4·583 (0·669)	4·583 (0·806)	1·000
ChatGPT's responses were irrelevant	1·750 (1·545)	1·750 (1·138)	1·417 (0·793)	1·639 (1·175)	0·736
ChatGPT coped well with any errors or mistakes	3·750 (1·422)	4·000 (0·739)	4·167 (0·718)	3·972 (1·000)	0·603
ChatGPT's responses helped make my responses more accurate to the best of my knowledge	4·333 (1·231)	3·833 (1·030)	4·333 (0·651)	4·167 (1·000)	0·379
ChatGPT's responses helped make my responses more complete to the best of my knowledge	4·250 (1·288)	4·000 (0·853)	4·333 (0·651)	4·194 (0·951)	0·683
ChatGPT helped me be faster in generating my response	4·250 (1·288)	4·250 (0·754)	4·500 (0·674)	4·333 (0·926)	0·758
ChatGPT had little effect on my response	2·667 (1·670)	2·417 (0·996)	2·083 (1·240)	2·389 (1·315)	0·566
I sometimes disagreed with some or all of ChatGPT's response	1·500 (1·168)	2·667 (0·985)	2·250 (1·055)	2·139 (1·150)	0·037
I am willing to use the responses provided by ChatGPT	4·250 (1·215)	4·167 (0·718)	4·583 (0·515)	4·333 (0·862)	0·469
I am more comfortable using my usual resources instead of ChatGPT	2·083 (1·165)	2·583 (0·996)	2·167 (1·030)	2·278 (1·059)	0·477
I trust my usual resources more than ChatGPT for clinical work	2·167 (1·193)	3·500 (0·905)	2·250 (1·138)	2·639 (1·222)	0·008

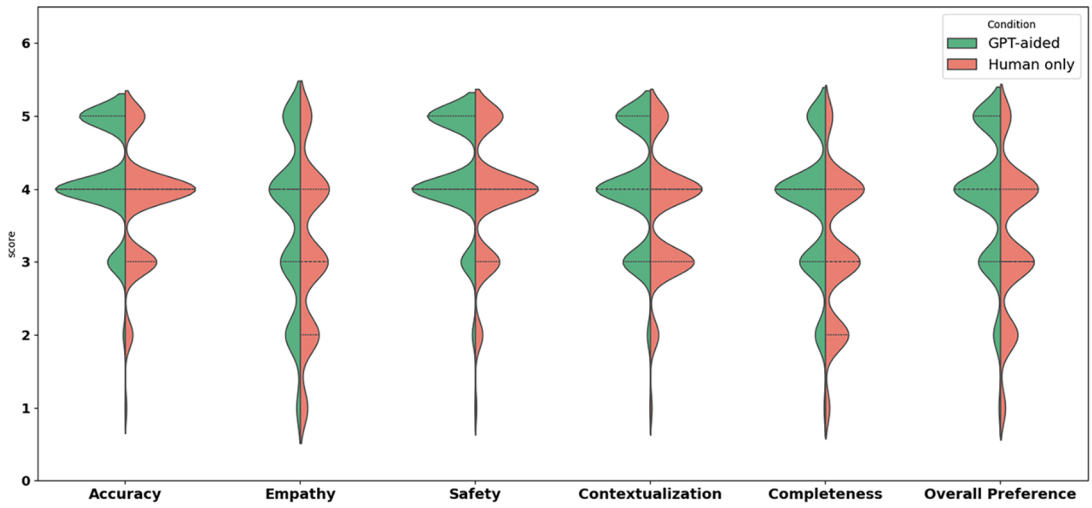

Figure 2. Violin plots displaying the distributions of various performance metrics when responses are GPT-aided (green) and Human-only (orange). Each violin represents the distribution of scores for a given metric by a rater, showing not only central tendencies but also the range and shape of the data's underlying distribution.

Table 4. Multi-Group comparison (MD, RN, CHW) between GPT-aided and FLW-only for each metric.

	GPT-Aided	GPT-Aided	GPT-Aided	GPT-Aided	Human-Only	Human-Only	Human-Only	Human-Only
Metric	**CHW**	**MD**	**RN**	**P**	**CHW**	**MD**	**RN**	**P**
Accuracy	4·00 (1·00)	4·00 (1·00)	4·00 (1·00)	0·009	4·00 (1·00)	4·00 (0·00)	4·00 (1·00)	<0·001 (***)
Empathy	4·00 (2·00)	4·00 (1·00)	4·00 (1·00)	0·002	3·00 (2·00)	4·00 (1·00)	3·00 (2·00)	<0·001 (***)
Safety	4·00 (1·00)	4·00 (1·00)	4·00 (1·00)	0·005	4·00 (1·00)	4·00 (1·00)	4·00 (1·00)	<0·001 (***)
Contextualization	4·00 (2·00)	4·00 (1·00)	4·00 (2·00)	0·015	3·00 (1·00)	4·00 (1·00)	4·00 (1·00)	<0·001 (***)
Completeness	4·00 (1·00)	4·00 (1·00)	4·00 (1·00)	0·001	3·00 (2·00)	4·00 (1·00)	3·00 (1·00)	<0·001 (***)
Overall Preference	4·00 (1·00)	4·00 (1·00)	4·00 (1·00)	<0·001	3·00 (2·00)	4·00 (1·00)	3·00 (1·00)	<0·001 (***)

p-values are annotated based on significance (*p < 0·05, ***p < 0·001, and "ns" if p ≥ 0·1). Medians and interquartile range (IQR) are reported for each metric.

GPT, particularly in CHEWs and RNs, with median scores rising from 3·00 in the human-only responses to 4·00 with GPT support, and reduced variability in scores indicating more consistent quality. These findings suggest GPT notably enhanced response quality, especially for non-MD FLWs, and helped narrow gaps in performance

Inter-rater reliability of clinical evaluations across metrics: GPT-aided vs. human-only responses

The consistency of clinical ratings across six evaluation metrics was assessed using Fleiss' Kappa, which quantifies inter-rater agreement for categorical ratings by four raters (Table 5). Kappa values range from 0 to 1, with values between 0·41–0·60 indicating moderate agreement and 0·61–0·80 indicating substantial agreement. Among all metrics, overall preference showed the highest inter-rater reliability, with GPT-aided responses achieving a Fleiss' Kappa of 0·77 (95% CI, 0·73–0·81), slightly higher than human-only responses at 0·72 (CI: 0·61–0·79). Similarly, the Completeness metric demonstrated substantial agreement, with GPT-aided responses at 0·71 (CI: 0·64–0·76) and human-only at 0·66 (CI: 0·53–0·72). Accuracy, Empathy, and Contextualization all showed moderate consistency, while the Safety metric showed the lowest inter-rater agreement in both groups.

Table 5. Consistency of clinical ratings across evaluation metrics.

	GPT-Aided	GPT-Aided	Human-Only	Human-Only
Metric	**Fleiss Kappa**	**95% CI**	**Fleiss Kappa**	**95% CI**
Accuracy	0·60	0·52 – 0·66	0·62	0·53 – 0·69
Empathy	0·56	0·30 – 0·70	0·61	0·38 – 0·73
Safety	0·50	0·41 – 0·57	0·44	0·35 – 0·58
Contextualization	0·58	0·47 – 0·66	0·55	0·43 – 0·64
Completeness	0·71	0·64 – 0·76	0·66	0·53 – 0·72
Overall Preference	0·77	0·73 – 0·81	0·72	0.61 – 0·79

Computational linguistics

The GPT-aided responses were 60% longer than human-only responses, and this difference was statistically significant (mean [95% CI] word count, ~120 [111·63–128·29] vs ~75 [69·90–79·95]; t = 9·09; p < ·001; difference, 60%). Moreover, the GPT-aided responses demanded a higher level of education to comprehend (mean [SD] Flesch-Kincaid grade level, 13·42 vs 11·77; t = 4·48; p < ·001; difference, 14%). Their polarity scores were comparable (mean [SD], 0·13 [0·13] vs 0·14 [0·18]; p = ·20), whereas the GPT-aided responses were more subjective (0·47 vs 0·45; t = 2·28; p = ·02; difference, 4%) (Figure 2).

Non-Clinician evaluation

Out of 256 responses reviewed by non-clinicians, GPT-aided responses were generally preferred across all themes: Caring (58·01%), Ease of Understanding (57·62%), and Trust (59·86%), with trust showing the strongest preference. This trend was most pronounced among CHEWs—75·00% found the responses more caring, 71·88% easier to understand, and 70·31% more trustworthy. Preferences on GPT-aided responses were lower but still consistent for RNs and MDs. For RNs, ChatGPT-aided responses were seen as more caring (52·86%), easier to understand (51·82%), and more trustworthy (54·95%). MD responses aided by ChatGPT were rated more trustworthy (57·81%), easier to understand (51·82%), and more caring (51·82%).

A binomial test confirmed statistical significance for Trust (p = 0·027), while Caring and Ease of Understanding did not reach significance (p = 0·093). Wilcoxon tests showed significant differences for CHEWs (p < 0·001) and MDs (p = 0·016), but not RNs (p = 0·112).

Discussion

This study examined FLWs' response quality and perceptions in Lagos, Nigeria, when responding to women's health questions using their usual human-only responses versus ChatGPT-aided responses. As one of the first African studies to assess healthcare professionals' use of ChatGPT, it examined efficacy, user insights, and patient perceptions of AI-supported healthcare communication.

The ChatGPT-aided responses consistently outperformed the human-only responses across all evaluated metrics: Accuracy, Completeness, Contextualization, Empathy, Safety, and overall quality (Figure 1). These findings align with other studies that have compared ChatGPT's performance to that of human experts in domains such as oncology-related patient care[25] or ophthalmology[26] and even physician responses on social media.[24] Notably, our data indicate that ChatGPT's advantages remained consistent regardless of the FLW's group level (MD, RN, or CHW), even though scores were highest for MDs in both human-only and GPT-aided responses (Table 4). These results point to ChatGPT as a potentially powerful supplementary resource in low-resource settings, potentially capable of improving the quality of responses to patient's health inquiries.

Previous studies have highlighted the potential of AI to improve efficiency and quality of care by automating administrative tasks, providing decision support, and enhancing patient engagement.[9,10] Our results empirically support these claims, showing AI can bridge knowledge and resource gaps in FLWs' knowledge, resources, or skill capacity common in low and middle-income countries (LMICs).[27,28] While MDs maintained the highest scores, ChatGPT notably elevated CHEWs' and RNs' responses near MD levels (Table 4). Beyond accuracy, ChatGPT improved empathy and safety scores, demonstrating its ability to adopt a patient-friendly tone despite being designed for broader use.[15] This underscores LLMs' capacity to reduce performance disparities in virtual coaching and general health Q&A in LMICs.

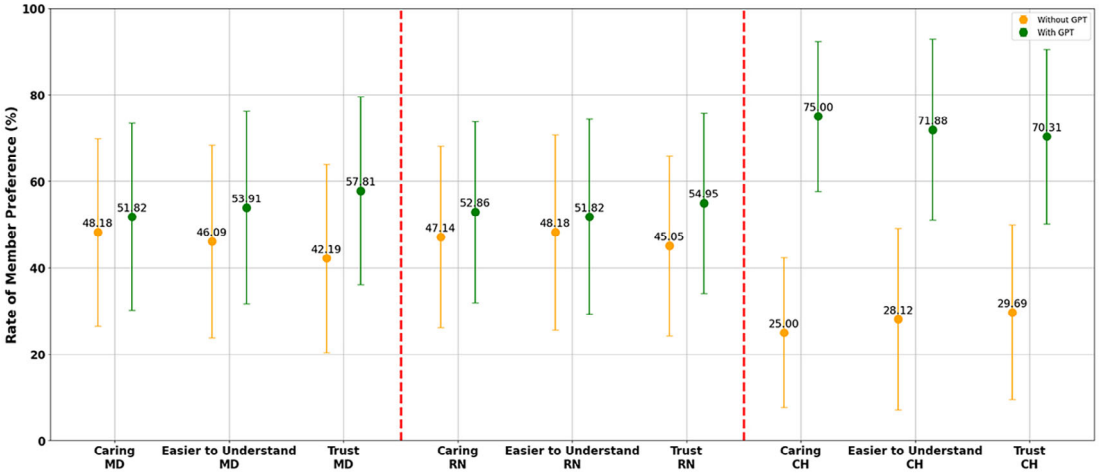

Figure 3. Comparative evaluation of member preferences across groups and themes.

Non-clinician reviewers' feedback revealed strong preferences for ChatGPT-aided responses in terms of Care, Understanding, and Trust (Figure 3), reflecting patients' prioritization of these qualities in health communication. They trusted GPT-aided responses more and found them more caring and understandable. This aligns with findings that people find ChatGPT's responses more competent and trustworthy.[25] Trust is critical for technology acceptance in healthcare, and it is an influential factor in users' acceptance of technology such as ChatGPT for their healthcare needs.[29] This was evident particularly for CHEWs who are often the first or sole healthcare contact in LMICs, where ChatGPT's ability to enhance trust and engagement is significant.

Most FLWs found ChatGPT user-friendly, efficient, and helpful in improving response speed and quality (Table 3). This suggests that integrating LLMs into clinical workflows can optimize clinician time and reduce cognitive load from repetitive research tasks. The findings highlight opportunities to develop specialized LLM health chatbots, fine-tuned with up-to-date clinical guidelines and localized data, to address misinformation and tailor outputs for safety. Such models could further narrow quality gaps among healthcare workers providing accurate and contextualized advice, especially in resource-constrained LMIC settings.

Despite promising results, caution is warranted. LLMs can generate misinformation (hallucinations)[30,31] and reflect biases in training data. Although clinicians rated GPT-aided responses as safer, rigorous validation and oversight remain essential, especially for high-risk clinical scenarios involving critical diagnoses and treatments. Ongoing efforts to detect and mitigate hallucinations must be integrated into clinical practice to address any inacuracies.[32] Additionally, while contributing to completeness, GPT-aided responses' greater length and complexity (Table 4) can risk overwhelming low-literacy individuals, necessitating design strategies for clear, concise communication.

Finally, FLWs' skills and readiness to adopt AI are critical. While two-thirds had telemedicine experience and found ChatGPT intuitive, many were new to LLMs, and nearly one-third still trusted traditional resources more. Strengthening their capacity to engage with AI in clinical decision-making will be critical for successful integration into routine care. Targeted AI training and investments in appropriate human-in-the-loop oversight can facilitate safe adoption, balancing risk mitigation with the potential to enhance frontline healthcare delivery.

In conclusion, LLM-aided responses were rated superior by clinicians and preferred by non-clinicians, highlighting opportunities to expand AI-assisted healthcare communication. Further research is needed to optimize integration, ensure safety, and evaluate impact in diverse medical fields and marginalized contexts.

Limitations

The study has some limitations. The questions were sourced from those posed by women of reproductive health at a single outreach event in Lagos on a limited number of health topics. The questions may not entirely reflect the full diversity of people's concerns or the entirety of queries encountered during coaching sessions. However, the number of question-and-answer pairs analysed in this study was similar to other studies on similar topics.[12,29] Further studies can build on this to analyse a diverse set of responses from a more proportionally representative sample. In addition, the number of FLWs, though professionally diverse, was relatively small and drawn only from Lagos State. Most participants were also relatively early in their careers, with fewer than five years of practice. As a result, the findings may not fully capture the

perspectives of more experienced providers or be generalizable to the wider population of FLWs in Nigeria or other low-resource settings. Also, participants and reviewers described their respective tasks as cumbersome. Bias may have been introduced by design-type, 'learning effect' or 'fatigue effect.' The latter could have affected the quality of participants' responses and reviewer scores. The order of questions to answer (participants) and responses to rate (evaluators) were scrambled to avert bias.

Data availability

The data supporting the findings of this study are available upon reasonable request. The data are shared in accordance with applicable ethical and privacy guidelines and are intended solely for legitimate academic and scientific purposes. Access to the data will require a clear statement of research objectives and a signed data use agreement to prevent misuse or unauthorized redistribution.

Acknowledgement

The authors would like to thank all the evaluators (clinicians and non-clinicians), the OEM Group, the mDoc team for their contributions, and, most importantly, the women who posed the questions

References

1. Abimbola S, *et al.*: **Artificial intelligence in healthcare in Africa: opportunities and challenges.** *BMJ Glob. Health.* 2020; **5**(12): e003390.
2. Kruk ME, Chukwuma A, Mbaruku G, *et al.*: **Variation in quality of primary-care services in Kenya, Malawi, Namibia, Rwanda, Senegal, Uganda and the United Republic of Tanzania.** *Bull. World Health Organ.* 2017 May 9; **95**(6): 408–418. PubMed Abstract | Publisher Full Text | Free Full Text
3. Leslie HH, Sun Z, Kruk ME: **Association between infrastructure and observed quality of care in 4 healthcare services: A cross-sectional study of 4,300 facilities in 8 countries.** Persson LÅ, editor. *PLoS Med.* 2017 Dec 12; **14**(12): e1002464.
4. Moses MW, Korir J, Zeng W, *et al.*: **Performance assessment of the county healthcare systems in Kenya: a mixed-methods analysis.** *BMJ Glob. Health.* 2021 Jun 1; **6**(6): e004707. PubMed Abstract | Publisher Full Text | Free Full Text | Reference Source
5. Ameh S, Akeem BO, Ochimana C, *et al.*: **A qualitative inquiry of access to and quality of primary healthcare in seven communities in East and West Africa (SevenCEWA): perspectives of stakeholders, healthcare providers and users.** *BMC Fam. Pract.* 2021 Feb 25; **22**(1): 45. PubMed Abstract | Publisher Full Text | Free Full Text
6. Etuk I, Iwuala A, Njoku K, *et al.*: **Barriers to health in women of reproductive age living with or at risk of non-communicable diseases in Nigeria: a Photovoice study.** *BMC Womens Health.* 2023 Jan 2; **23**(1): 3. PubMed Abstract | Publisher Full Text | Free Full Text
7. Ogboye A, Akpakli JK, Iwuala A, *et al.*: **Prevalence of non-communicable diseases and risk factors of pre-eclampsia/eclampsia in four local government areas in Nigeria: a cross-sectional study.** *BMJ Open.* 2023 Oct 1 [cited 2024 Jul 1]; **13**(10): e071652. PubMed Abstract | Publisher Full Text | Free Full Text | Reference Source
8. Olatunji G, Aderinto N, Kokori E, *et al.*: **Nigeria's new policy: solution for the health-care workforce crisis?** *Lancet.* 2024 Oct; **404**(10460): 1303–1304. PubMed Abstract | Publisher Full Text
9. Lawal L, Lawal AO, Amosu OP, *et al.*: **The COVID-19 pandemic and health workforce brain drain in Nigeria.** *Int. J. Equity Health.* 2022 Dec 5; **21**(1): 174. PubMed Abstract | Publisher Full Text | Free Full Text
10. Chen S, Yu J, Chamouni S, *et al.*: **Integrating machine learning and artificial intelligence in life-course epidemiology: pathways to innovative public health solutions.** *BMC Med.* 2024 Sep 2; **22**(1): 354. PubMed Abstract | Publisher Full Text | Free Full Text
11. Laka M, Carter D, Merlin T: **Evaluating clinical decision support software (CDSS): challenges for robust evidence generation.** *Int. J. Technol. Assess. Health Care.* 2024 Jan 1; **40**(1): e16. PubMed Abstract | Publisher Full Text | Free Full Text | Reference Source
12. Hart SN, Hoffman NG, Gershkovich P, *et al.*: **Organizational preparedness for the use of large language models in pathology informatics.** *Journal of Pathology Informatics.* 2023 Oct 1; **14**: 100338. PubMed Abstract | Publisher Full Text | Free Full Text | Reference Source
13. Bedi S, Jain SS, Shah NH: **Evaluating the clinical benefits of LLMs.** *Nat. Med.* 2024 Jul 26 [cited 2024 Aug 13]; **30**: 2409–2410. Publisher Full Text | Reference Source
14. Alowais SA, Alghamdi SS, Alsuhebany N, *et al.*: **Revolutionizing healthcare: the Role of Artificial Intelligence in Clinical Practice.** *BMC Med. Educ.* 2023 Sep 22; **23**(1): 1–15.
15. van Dis EAM, Bollen J, Zuidema W, *et al.*: **ChatGPT: Five Priorities for Research.** *Nature.* 2023 Feb 3; **614**(7947): 224–226. PubMed Abstract | Publisher Full Text
16. OpenAI.: **Introducing ChatGPT.** *OpenAI.* 2022. Reference Source
17. Deng J, Heybati K, Park YJ, *et al.*: **Artificial intelligence in clinical practice: A look at ChatGPT.** *Cleve. Clin. J. Med.* 2024 Mar 1; **91**(3): 173–180. PubMed Abstract | Publisher Full Text
18. Mu Y, He D: **The Potential Applications and Challenges of ChatGPT in the Medical Field.** *International Journal of General Medicine.* 2024 Mar 5; **17**: 817–826. PubMed Abstract | Publisher Full Text | Free Full Text | Reference Source
19. Kolata G: **When Doctors Use a Chatbot to Improve Their Bedside Manner.** *The New York Times.* 2023 Jun 12. Reference Source
20. Agarwal N, Moehring A, Rajpurkar P, *et al.*: *Combining Human Expertise with Artificial Intelligence: Experimental Evidence from Radiology.* National Bureau of Economic Research; 2023. Reference Source
21. Plesner LL, Müller FC, Brejnebøl MW, *et al.*: **Using AI to Identify Unremarkable Chest Radiographs for Automatic Reporting.** *PubMed.* 2024 Aug 1; **312**(2): e240272.
22. Lobach D, Sanders GD, Bright TJ, *et al.*: **Enabling health care decisionmaking through clinical decision support and knowledge management.** *Evid. Rep. Technol. Assess.* 2012 Apr [cited 2025 Feb 7]; (203): 1. Reference Source
23. Ajisegiri WS, Abimbola S, Tesema AG, *et al.*: **"We just have to help": Community health workers' informal task-shifting and task-sharing practices for hypertension and diabetes care in Nigeria.** *Front. Public Health.* 2023 Jan 26; **11**. PubMed Abstract | Publisher Full Text | Free Full Text
24. Ayers JW, Poliak A, Dredze M, *et al.*: **Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum.** *JAMA Intern. Med.* 2023 Apr 28; **183**(6): 589–596. PubMed Abstract | Publisher Full Text | Free Full Text | Reference Source

25. Yalamanchili A, Sengupta B, Song J, *et al.*: **Quality of Large Language Model Responses to Radiation Oncology Patient Care Questions.** *JAMA Netw. Open.* 2024 Apr [cited 2024 Apr 3]; **7**(4): e244630–0.
PubMed Abstract | **Publisher Full Text** | **Free Full Text** | **Reference Source**

26. Bernstein IA, Zhang Y, Govil D, *et al.*: **Comparison of Ophthalmologist and Large Language Model Chatbot Responses to Online Patient Eye Care Questions.** *JAMA Netw. Open.* 2023 Aug 22; **6**(8): e2330320–0.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

27. Boro E, Stoll B: **Barriers to COVID-19 Health Products in Low-and Middle-Income Countries During the COVID-19 Pandemic: A Rapid Systematic Review and Evidence Synthesis.** *Front. Public Health.* 2022 Jul 22; **10**.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

28. Olayiwola JN, Udenyi ED, Yusuf G, *et al.*: **Leveraging Electronic Consultations to Address Severe Subspecialty Care Access Gaps in Nigeria.** *J. Natl. Med. Assoc.* 2020 Feb; **112**(1): 97–102.
PubMed Abstract | **Publisher Full Text**

29. Choudhury A, Elkefi S, Tounsi A: **Exploring factors influencing user perspective of ChatGPT as a technology that assists in healthcare decision making: A cross sectional survey study.** *PloS one.* 2024 Mar 8; **19**(3): e0296151.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

30. Ji Z, Lee N, Frieske R, *et al.*: **Survey of Hallucination in Natural Language Generation.** *ACM Comput. Surv.* 2022 Nov 17; **55**(12).

31. Shen Y, Heacock L, Elias J, *et al.*: **ChatGPT and Other Large Language Models Are Double-edged Swords.** *Radiology.* 2023 Jan 26; **307**(2): e230163.
PubMed Abstract | **Publisher Full Text**

32. Farquhar S, Kossen J, Kuhn L, *et al.*: **Detecting hallucinations in large language models using semantic entropy.** *Nature.* 2024 Jun 1; **630**(8017): 625–630.
PubMed Abstract | **Publisher Full Text** | **Free Full Text** | **Reference Source**