Impact of LLM Assistance on Physician Decision-Making: A Multi-Country Randomized Controlled Trial*

Nicholas Rounding[1*], Luthfi Saiful Arif[2], Janine Berg[3], Jochen Cals[4], Diederik De Boer[5], Eefje De Bont[4], Sander Dijksman[1], Ardi Findyartini[2], Didier Fouarge[1], Marie-Christine Fregin[1], Pawel Gmyrek[3], Nadia Greviana[2], Ralph Leijenaar[4], Soraiya Manji[6], Annastacia Mbithi[6], Norah Obungu[6], Arierta Pujitresnani[2], Roselyter Rianga[6], Diantha Soemantri[2], Sairabanu Mohamed Rashid Sokwalla[6], Sanne Steens[1], Lucia Velasco[1], Ardy Wildan[7], Prasandhya Astagiri Yusuf[2], and Mark Levels[1]

[1]Research Centre for Education and the Labour Market, Maastricht University, Maastricht, Netherlands
[2]Medical Education Center, Indonesian Medical Education and Research Institute, Universitas Indonesia, Jakarta, Indonesia
[3]Research Department, International Labour Organization, Geneva, Switzerland
[4]CAPHRI Care and Public Health Research Institute, Maastricht University, Maastricht, Netherlands
[5]Maastricht School of Management, Maastricht University, Maastricht, Netherlands
[6]Aga Khan University Hospital, Nairobi, Kenya
[7]Division of Endocrinology, Metabolism, and Diabetes, Department of Internal Medicine, Universitas Indonesia, Jakarta, Indonesia
[*]Corresponding Author: n.rounding@maastrichtuniversity.nl

Abstract

Background Poor-quality healthcare contributes to preventable mortality, particularly in low- and middle-income countries. We evaluated whether large language model (LLM) access improves physician clinical performance across economically diverse settings.

Methods We conducted a parallel-group randomised controlled trial in Indonesia, Kenya, and the Netherlands. Physicians (internal/family medicine residents; supplemented with senior house officers and first-year residents in Kenya; attending physicians in the Netherlands) were recruited through university networks. Sessions were completed in computer laboratories (Indonesia, Netherlands) or consultation rooms (Kenya). Participants (N=249) were randomly assigned via simple randomisation to control or intervention (GPT-4o access). Participants and investigators were unmasked; graders were masked. Participants completed four standardised clinical vignettes covering globally prevalent primary care conditions. The primary outcome was percent correct score (weighted correct rubric items divided by total possible). All randomised participants were analysed by assigned group. Registered: AEA RCT Registry (AEARCTR-0013399), ClinicalTrials.gov (NCT07374926); complete.

Findings Between August 2024 and January 2025, 249 physicians were randomised and analysed: Indonesia 81 (40 control, 41 intervention), Kenya 60 (30, 30), Netherlands 108 (58, 50). Mean scores were 39.2% (SD 9.5) without versus 49.9% (SD 12.9) with LLM access in Indonesia (difference 10.7 percentage points, 95% CI 5.7 to 15.7, $p<0.001$); 47% (SD 10) versus 65% (SD 10.4) in Kenya (18 pp, 95% CI 12.7 to 23.2, $p<0.001$); 54.2% (SD 6.5) versus 61.4% (SD 10.9) in the Netherlands (7.2 pp, 95% CI 3.7 to 10.7, $p<0.001$). Harms were not assessed.

Interpretation LLM assistance improved physician performance across high-, middle-, and low-income settings, with the largest effects where baseline performance was lowest. These findings suggest LLMs may help reduce disparities in care quality, though results derive from controlled conditions rather than real clinical practice.

Funding INRIA on behalf of Global Partnership on Artificial Intelligence (Grant Number: INRIA 2024-1242).

Competing interests The authors declare that they have no known competing financial or non-financial interests that could have appeared to influence the work reported in this paper.

*We are deeply grateful to all physicians who participated in the experiment. We thank the support staff in Indonesia, Kenya, and the Netherlands for ensuring smooth study operations and technical support. We acknowledge Linda Colen, Jordy Frijns, Ingrid van der Heijden, Elsje Kuijper for advice, assistance, and operational support. We also thank the partnering institutions and clinical facilities in each country for providing space, infrastructure, and organizational assistance. This study was funded INRIA on behalf of the Global Partnership on Artificial Intelligence.

1 Introduction

Over the past decades many regions have expanded basic health coverage, yet disparities in the quality of care remain a pervasive global challenge [1]. Poor-quality of care is a global concern, with some studies suggesting that it contributes more to mortality than lack of access, with particularly severe consequences in low- and middle-income countries where 60% of deaths from treatable and preventable conditions arising from poor-quality care with the remainder arising from non-utilization of healthcare [2]. Inequalities in the quality of care have been recorded within and across countries [3–5]. Improving the quality of physician care could eliminate millions of unnecessary deaths every year and reduce inequities in care provision between the Global North and South, and between population groups within a country. The value of digital technologies to improve the provision of healthcare, and the quality of physician care, is increasingly being acknowledged by healthcare providers and governments [6].

Innovations in Generative AI, i.e. Large Language Models (LLMs), have been proposed as tools that can augment the provision of healthcare by aiding physicians in their work [7]. LLMs have shown potential in key clinical tasks, including clinical reasoning and generating differential diagnoses, and have demonstrated strong performance in simulated clinical environments [8–13]. Recent studies suggest that LLMs could help improve the quality of care by augmenting physician clinical decision-making, including both diagnosis and management tasks [14–17]. Whether physicians will be directly replaced by AI or not is debated, research suggests that the exposure of physicians to task replacement by LLMs is relatively low suggesting augmentation is more likely in the immediate future [18]. Although the potential of LLMs to improve physicians quality of care seems well-established, it is less clear whether these impacts are shaped by cultural differences in clinical reasoning [19], technology acceptance [20] or different regulatory and organizational contexts. To establish whether contexts indeed matter, it is important to explore whether LLMs affect physician performance differently in different countries.

We designed a parallel group randomized controlled trial to evaluate the effectiveness of an LLM (GPT-4o) in improving physician clinical performance on vignettes. We administered clinical vignettes[21] in simulated primary care scenarios across three economically and geographically diverse countries: Indonesia, Kenya, and the Netherlands. These countries represent distinct income strata, upper-middle-income, lower-middle-income, and high-income, respectively, offering a broader perspective on regional and economic differences in healthcare delivery. Clinical vignettes are commonly used to assess the augmentative abilities of LLMs in other studies [14–16]. Performance on clinical vignettes is a validated measure of the quality of physician clinical practice and care, comparable to the gold standard of standardized patients [21]. Physicians' answers in our vignettes are graded against detailed rubrics built from context specific, evidence-based, best practice guidelines. Such guidelines have been demonstrated to improve quality of care globally [22]. By comparing the outcomes of physicians with and without access to the LLM in three countries, we analyze how LLM access affects variation in physician performance within and across countries.

2 Methods

2.1 Participants

Our target group consisted of residents in family and internal medicine. Recruitment was supplemented with attending physicians in internal medicine in the Netherlands, and first-year residents in other specialties (i.e., surgery, anesthesiology, and pediatrics) or post-internship pre-residency medical officers, referred to as Senior House Officers, in Kenya. Whilst our target group were primarily family and internal medicine residents, there are differences in practice across the three countries. In the Netherlands, our target group was family medicine residents. Residents follow a 3-year structured training program, in the 1st and 3rd year they are placed in a General Practice surgeries, with regular learning days at their university. In the second year, residents are placed in mandatory internships in emergency care, mental healthcare, and a nursing home. Our target group was 1st and 3rd year residents. In Indonesia, the participants were internal medicine residents in at least their second year. The internal medicine residency program is a 4-year program in university affiliated teaching hospitals. In Kenya, internal and family medicine residents complete 4-year programs in hospitals and out-reach clinics. To supplement our primary target group, first-year residents from other specialties and Senior House Officers (SHOs) were included. SHOs are physicians who have finished undergraduate education and have undertaken 1-year of internships. They are placed across hospital departments and are responsible for patient care, including daily ward rounds, examinations, and monitoring. We assessed that they would have the required experience and knowledge to be able to complete the vignettes. Also added were first-year residents in other specialties (surgery, anesthesia, pediatrics, pathology, radiology and obstetrics/gynecology), as they will have only recently been SHOs.

Participants were recruited through the university networks of the three participating universities (Maastricht University, Universitas Indonesia, Aga Khan University). Written informed consent was obtained from participants preceding enrollment and randomization. Participants were not compensated for participating in this study. Sessions were organized in controlled environments: computer labs in Indonesia and the Netherlands, and educational consultation rooms in Kenya. Participants in both arms in Indonesia and the Netherlands completed the experiment in the same room. To avoid the chance of cross-participant contamination, participants were placed apart from each other. If participants were seated to closely to each other a divider was placed so they could not see other screens. Furthermore, all participants were monitored by members of the research team. In Kenya, participants completed the experiments in an individual room. All participants were asked not to discuss the content of the study with their colleagues until data collection was finalized. Figure 1 shows the flow of participants through each stage of the trial.

2.2 Study Design and Vignette Development

We designed clinical vignettes to simulate patient-physician consultations, an approach used extensively in the literature when attempting to assess both the performance of LLMs and their interactions with physicians [8, 12, 14–16, 23]. We follow the structure of clinical performance and value vignettes, a vignette design validated to perform well against standard patient designs [21]. This structure has previously been used to measure physician clinical performance in diverse global settings [24–26]. Our vignette design follows 9 stages (1) Presenting Problem & Initial Differential Diagnosis, (2) Asking about patient history, (3) Additional Differential Diagnosis, (4) Listing Physical Exams, (5) Differential Diagnosis, (6) Additional diagnostic (Lab) Tests, (7) Final Differential Diagnosis, (8) Medication, (9) Follow-Up/Advice. Information was provided sequentially and at each stage participants listed the actions they would take.

The provided information was static, meaning that at each step every participant saw the same information regardless of their answers in previous steps. All clinical vignettes were presented to participants in English, participants were allowed to respond either in English or their own language. All medical education is delivered in English in Kenya. In Indonesia, participants i.e residents are required to take the TOEFL test during their enrollment as resident. In the Netherlands, physicians are expected to be able to read English to at least a B2 level, in order to keep on top of developments in professional literature. See Supplement 2 for an overview of each vignette and the sequential information provided (Supplement Tables 2.1, 2.3, 2.5, and 2.7).

Each case was developed for this trial by a group of co-authors. Following the clinical performance and value vignette design, we created four cases 21. The selected conditions (cardiovascular, respiratory, musculoskeletal disease and a viral infection) were chosen as they are globally prevalent and can be diagnosed and treated in primary care without the need for expensive treatments, advanced technology, or specialized care, a design specification used previously in the literature 24. Participants in each county were presented with the same cases. To assess the validity of the case selection both geographically and for our sample of physicians, participants were asked whether the patient cases presented were representative of those typically encountered in clinical practice. Participants across all three countries reported high levels of agreement: 89% in Indonesia, 92% in Kenya, and 94% in the Netherlands.

A central rubric for each case based on evidence-based best practices was developed by an experienced vignette creator in the Netherlands (EdB), drawing on comprehensive national guidelines i.e. Dutch College of General Practitioners (NHG) and the UK's National Institute for Health and Care Excellence (NICE). These best-practice rubrics were then adapted by domain experts (SM, AW) in each country to reflect local clinical contexts. Each rubric item was applied a weighting based on its clinical significance, using a standardized scale of 0.33, 0.5, or 1. This process generated some differences, with more items being present in the rubrics. See Supplement 2 for the rubric items per vignette (Supplement Tables 2.2, 2.4, 2.6, and 2.8). Preplanned sensitivity analyses address cross-country differences in the rubric, reported in Supplement 5, Table 5.2.

Participants were randomly assigned either to the control or intervention group using simple randomization with an allocation ratio of 1:1 and asked to complete the vignettes in an online environment. Intervention group participants were given access to an LLM (GPT-4o) via the OpenAI API through an interface developed by Maastricht University. This interface was integrated into the online Qualtrics environment (Supplement Figure 1.1). There was no affiliation with OpenAI. The intervention group were instructed that they could optionally use the LLM, and were provided with prompting instructions (see Supplement 7). Due to the widespread availability of LLMs (including integration into popular search engines), control group participants were asked not to use internet search to avoid contamination. To allow us to isolate the effect of LLM usage across our samples, participants were asked not to use traditional resources. This allows us to determine that any statistically significant differences found across our samples can be attributed purely to the use of the LLM.

As a final step, to assess the baseline performance of the LLM alone, we input the vignette cases exactly as they were shown to physicians into GPT-4o once, using no fine-tuned prompting (i.e. zero shot learning). See Supplement 4 for the results of the GPT-4o responses (Supplement Table 4.1).

The study was reviewed and approved by the ethical review boards of the participating universities (University of Indonesia, Aga Khan University Nairobi, and Maastricht University). Written informed consent was obtained from participants preceding enrollment and randomization. Participants were not compensated for participating in this study. We follow the CONSORT reporting guideline for randomized trials. The study protocol is available in Supplement 1. The study design was preregistered April 17, 2024 at AEA RCT Registry (RCT ID: AEARCTR-0013399). It was further registered at the National Clinical Trials Registry submitted December 12 2025, granted January 29 2026 (NCT ID: NCT07374926).

2.3 Response Grading and Outcome Generation

Open-ended participant responses were independently graded by two reviewers using the locally adapted rubrics. Graders were recruited through university networks and included recently graduated internal medicine physicians in Indonesia (n=11) and Kenya (n=4), and final-year medical students in the Netherlands (n=8). Each rubric item was assessed as present (1) or absent (0) by the reviewers. The primary outcome was calculated as the weighted sum of present items divided by the weighted sum of total possible items, creating a percent correct score per vignette. This was then averaged across participants. Pre-planned sensitivity analyses were conducted with scores generated at the vignette level using linear regressions and mixed effects models (see Supplement 5, Table 5.3 & 5.4).

We calculated Cohen's kappa for all rubric items, which yielded a value of 0.71 (Indonesia: 0.68; Kenya: 0.67; Netherlands: 0.77), indicating substantial agreement. While the reviewers did not assign weights to the responses, the overall Cohen's kappa could be driven predominanlty by lower weighted items. Therefore, we also calcuated a Cohen's kappa for the different weighted items, showing little variation with the highest value for rubric items weighted as 1 (0.72), followed by 0.5 (0.69) and 0.33 (0.68). We also calculated a one-way random-effects intraclass correlation coefficient (ICC) for our the scores at the vignette level. The calculated ICC was 0.96 (95% CI: 0.957 to 0.966; $p<.001$) indicating excellent agreement in composite scores between graders.

To address disagreements in assessment, we recruited a third expert reviewer in each country to adjudicate on all disagreements. Sensitivity analyses were conducted including adjudicated results and are reported in Supplement 5, Table 5.1. Further, in Supplement 5, Table 5.6, we report results for models run removing vignettes with the highest distance between reviewer assessment (>10% difference between the reviewers in their score assessment). Results using a mixed effects model are reported in Supplement 5, Table 5.5. Boxplots of scores at the vignette level are shown in Supplement Figure 5.1. To account for differences across countries in the reviewers, and to test the accuracy of the selected cohort in each country a member of the author list reviewed a selection of participants responses.

As an additional post-hoc analysis, we investigated the average use of the LLM per participant. Difference in usage patterns were observed when analyzing the data, which helped to explain some of the cross country differences we observed in our results. We used the average of the percentage of steps per vignette for which the participant made use of the LLM. For Indonesia the median was 0.44 (IQR 0.19 to 0.83), for Kenya 0.93 (IQR 0.78 to 0.97), for The Netherlands 0.68 (IQR 0.44 to 0.88). We then classified participants into two subgroups, based on the median value for each country. If the participant was above the median, they were classified into a High Usage subgroup, if they used below they were classified into a Low Usage subgroup. We then compared the impact of the LLM on the scores of high and low use participants

2.4 Statistical Analysis

A power analysis was conducted to determine the appropriate sample size for detecting meaningful effects in the randomized controlled trial. Using vignette parameters derived from the literature [21], including a mean vignette score of 71 and standard deviation of 5.4, a two-means clustered power analysis was implemented using Stata/SE 17.0. With an assumed intra-cluster correlation of 0.9 and targeting a power of 0.8, the analysis suggested that a minimum of 50 participants would be sufficient to detect a lower-bound effect size of 4.8%.

Descriptive analyses were performed using Stata/SE 17.0. Initial descriptive results compared the group with LLM access and the group without the LLM access. Variables shown are participants' career stage, medical specialty, years since start of medical education, and pre-experiment generative AI use. For categorical variables we reported percentages and numbers, for continuous variables we present means, standard deviations, medians, and interquartile ranges. The analyses were conducted at the participant level for each country separately using two-sided linear ordinary least squares regressions, clustering for standard errors at the participant level to correct for hierarchical clustering (additional pre-planned sensitivity analyses with different clustering strategies were performed – see Supplement 5). Model assumptions (i.e. normalcy, homoskedasticity) were verified. Mean differences and their 95% confidence intervals and p-values of two-sided regressions were presented to evaluate mean differences in average scores for each country as a separate sample. To evaluate cross-national differences in performance we ran linear OLS regressions on all combinations of countries: Indonesia - Kenya, Indonesia - Netherlands, Kenya - Netherlands, without control variables. We presented mean differences and their 95% confidence intervals and p-values for the three combinations.

3 Results

We recruited 249 resident physicians: 81 in Indonesia, 60 in Kenya, and 108 in the Netherlands. Data were collected in August–October 2024 (Netherlands), November 2024 (Indonesia), and January 2025 (Kenya). The cases used in this paper, and the rubric did not change in between tests. Table 1 reports baseline characteristics. In Indonesia, all participants were internal medicine residents. In Kenya, 45% were internal medicine residents, while the remaining 55% were either first-year residents in other specialties (i.e., surgery, anesthesiology, and pediatrics) or post-internship pre-residency medical officers referred to as Senior House Officers in Kenya. We refer to this group collectively as the non-internal medicine subgroup throughout. In the Netherlands, all participants were family medicine specialists, 83% were residents and 17% were attending physicians. The mean number of years since beginning medical education was 13 in Indonesia, 11.8 in Kenya, and 13.9 in the Netherlands Participants were randomly assigned to either a control or intervention group, with the latter receiving access to an LLM (GPT-4o). They were administered the same 4 clinical vignettes in a randomized order, in English, via the Qualtrics survey environment.

Figure 2 displays the distributions of our samples comparing physicians with and without LLM access, showing higher medians, interquartile ranges (IQRs), and boxplot tails in each country. Table 2 presents linear regression model estimates for the effect of LLM access on physician average vignette scores. We observed the largest difference between physicians with and without LLM access in the Kenyan sample 18% (95% CI: 12.7 to 23.2, p<0.001), followed by Indonesia 10.7% (95% CI: 5.7 to 15.7, p<0.001) and the Netherlands 7.2% (95% CI: 3.7 to 10.7, p<0.001). We address the effect of the Kenyan non-internal medicine resident subgroup in Supplement 3 (Supplement Table 3.1 and Supplement Figures 3.1–3.2), which found a larger effect of 20.5% (95% CI: 13.5 to 27.5, p<0.001) for the non-internal medicine subgroup. We found a correspondingly smaller effect of 10.6% (95% CI: 3.4 to 17.9, p=0.006) for the internal medicine residents, which is similar in magnitude to that found in Indonesia and the Netherlands. We present the analysis split by vignette case in Supplement 6 (Supplement Figure 6.1 and Supplement Table 6.1), while we detect differences in effects across countries and vignettes we are unable to detect a meaningful pattern.

Investigating cross-country differences, we found that physicians without LLM access had the lowest mean score in Indonesia (39.2%), followed by those in Kenya (47%) and in the Netherlands (54.2%). For those with access, we found the lowest mean score in Indonesia (49.9%), followed by the Netherlands (61.4%), and Kenya (65%). There was a statistically significant

Participants in the intervention group were provided with LLM access, usage was encouraged but not required. Hence we present an intent-to-treat analysis. We investigate the extent to which the LLM was used throughout the experiment. To further explore the impact of usage intensity, we classified physicians in the intervention group into High Usage and Low Usage groups, using the median rate of LLM use detected in each country as the threshold. We acknowledge that usage was not randomly assigned and inferences are limited by potential self-selection. Table 3 presents the distribution of participants in each category.

Figure 3 displays performance outcomes for High and Low Usage across countries. Due to the small sample sizes, estimates may be more sensitive to outliers and should therefore be interpreted with caution. Nevertheless, average scores were consistently higher for High Usage physicians in all three countries. Table 3 reports the estimated additional effects of high use. In Indonesia, we found that the High Usage sub group scored higher than the Low Usage sub group on our vignettes by 16.1% (95% CI: 9.6 to 22.5, p<0.001). In Kenya, the estimated difference was 8.4% (95% CI: 1.1 to 15.6, p=0.08). In the Netherlands, the difference was 7.5% (95% CI: 1.6 to 13.3, p=0.04). These results indicate that higher levels of LLM use are descriptively associated with higher average performance. However, some physicians in the Low Usage and no access groups achieved higher scores than some in the High Usage group, suggesting that access to an LLM alone is not a necessary condition for high-quality performance.

4 Discussion

In three randomized controlled trials conducted in Indonesia, Kenya, and the Netherlands, we found that providing physicians with access to GPT-4o improved their clinical vignette scores by 10.7% in Indonesia, 18% in Kenya, and 7.2% in the Netherlands compared to physicians without access. This study provides the first multi-country experimental evidence that LLM augmentation may enhance physician performance across diverse economic and healthcare contexts. Below we contextualize these findings as exploratory evidence for LLM effectiveness, discuss usage patterns and distributional effects, address study limitations, and outline implications for future research.

Our primary findings extends prior single-country studies [14, 15, 17] and demonstrate that LLM support is generalizable to different countries and contexts. Medical education and clinical reasoning contexts can vary from country to country [19] and thus we cannot expect that access could directly affect physician performance in different settings. By conducting our trial in three economically and geographically diverse settings, each using the same four globally prevalent primary-care vignettes, we provide evidence that physicians can use LLMs to enhance their clinical reasoning in all three contexts. The larger treatment effect in Kenya (18%) aligns with evidence suggesting that LLMs use is more effective for lower skilled or tenured individuals, due to the large proportion of non-specialized physicians in this sample [27]. Further research should explore the skill and experience nexus with relation to LLM use. As an additional analysis, we examined differential usage patterns of LLMs, highlighting that those who used the LLM in more steps in general performed better across all countries in our sample. We also found higher overall usage in Kenya than in Indonesia or the Netherlands, in line with the higher treatment effect in Kenya. These results could imply that greater scores are driven by humans working with the LLM and producing more than the sum of their parts, alternatively users could merely copy and paste LLM results. Experimental and observational evidence suggests professionals in other settings assess LLM responses before submitting them or selectively choose the recommendations to follow [27].

Our conclusions should come with four important caveats, that follow from key features of our design. First, control participants were restricted from traditional resources (clinical guidelines, web search) to prevent LLM contamination across intervention arms and to isolate the effect of LLM access across countries from differences in traditional resources. Access to such resources was restricted for two primary reasons: to ensure a baseline comparison that was not driven by differing traditional resources in each context and to avoid contamination as LLMs were widely available on the internet and incorporated into search such as Google. However, this restriction likely overestimates LLM benefits relative to real-world practice where physicians can access guidelines. Compared to similar studies, it is likely that our estimates represent an upper bound of the LLM effect. These studies find effect sizes for LLM access ranging from 2%-12% [14–16]. In an RCT comparing LLM support to internet search, McDuff et al [10]. found that the effect of LLM support on differential diagnosis generation was 7% larger than that of search, with search providing an 8% improvement over own knowledge on diagnosis. Due to differences in testing across these studies, it is impossible to adjust our results accurately, however, a cautious interpretation could suggest that the LLM did not improve results compared to internet search in at least the Netherlands and Indonesia.

Second, we cannot assume our results will hold for all medical conditions, either in primary care or otherwise. Our case selection focused on globally prevalent conditions with well-established evidence based best practice guidelines. Differences could arise either from changing physician or LLM behaviour. Physicians tend to use a non-analytical or heuristic approach in their clinical reasoning when dealing with common cases, and will switch to a more deliberate analytical approach on more complicated cases [28]. This could then change how they interact with LLMs. For LLM behaviour, we are unable to test for bias in the LLM training data. Future work could test LLM performance on more regionally specific case selections, especially focusing on discovering biases between the Global North and South. Similarly, racial and ethnic biases in the LLM may arise that could affect cross-country variations in quality of care [23]. Further, as our cases are globally prevalent and well covered in guidelines, it can be assumed that LLMs will have codified the relevant knowledge.

Third, this study was conducted in computer laboratories, a more ideal setting that is minimally influenced by external factors. Although such designs are common practice in studies that assess causal impacts of LLMs on physicians' performance [15, 16, 23] and generate results with high internal validity, they have relatively limited external validity. In a real clinical setting, decision making could be influenced by the unavailability of diagnostic tools and treatments, insurance rules, and other types of interruption, decreasing practitioner efficiency. Fourth, we did not assess harms. In a related study, Goh et al [15] found that responses from physicians with LLM access were less likely to contain harmful suggestions. However, in other settings when provided with deliberately falsified AI advice, physicians perform worse [29, 30]. Future research should investigate cross-national differences in these undesirable effects of LLMs.

These caveats notwithstanding, the results are in line with the assumption that LLMs may affect physicians' quality of care differently in different contexts. We cannot exclude the possibility that cross-country differences could be attributed to sampling differences, differences in the rubrics and grading across Kenya, Indonesia, and the Netherlands, alongside varying English proficiency across countries. However, additional analyses suggest these explanations are unlikely. To address limitations relating to the rubric, we performed additional analyses that demonstrated that cross-country gaps are reduced but not eliminated when only using answers found in all 3 context specific rubrics, suggesting that the different tests alone did not cause differences. Across all countries inter-grader reliability was strong, suggesting that between grader variability was low. We also assessed inter-country differences in grading by having a grader from each country review the same sample of participant responses.

The explanation for cross-national differences remains an area for further study. For example, there may be cross-cultural differences in technology acceptance [33] or cultural differences in clinical reasoning [19] of LLMs that should be further explored to provide guidance to the global medical community. Given the apparent potential of LLMs to reduce global inequalities in the quality of care, these further venues for research seem as promising as they are important.

5 References

References

[1] Jishnu Das and Jeffrey Hammer. Quality of Primary Care in Low-Income Countries: Facts and Economics. *Annual Review of Economics*, 6(1):525–553, August 2014. ISSN 1941-1383, 1941-1391. doi: 10.1146/annurev-economics-080213-041350. URL `https://www.annualreviews.org/doi/10.1146/annurev-economics-080213-041350`.

[2] Margaret E Kruk, Anna D Gage, Catherine Arsenault, Keely Jordan, Hannah H Leslie, Sanam Roder-DeWan, Olusoji Adeyi, Pierre Barker, Bernadette Daelmans, Svetlana V Doubova, Mike English, Ezequiel García-Elorrio, Frederico Guanais, Oye Gureje, Lisa R Hirschhorn, Lixin Jiang, Edward Kelley, Ephrem Tekle Lemango, Jerker Liljestrand, Address Malata, Tanya Marchant, Malebona Precious Matsoso, John G Meara, Manoj Mohanan, Youssoupha Ndiaye, Ole F Norheim, K Srinath Reddy, Alexander K Rowe, Joshua A Salomon, Gagan Thapa, Nana A Y Twum-Danso, and Muhammad Pate. High-quality health systems in the Sustainable Development Goals era: time for a revolution. *The Lancet Global Health*, 6(11):e1196–e1252, November 2018. ISSN 2214-109X. doi: 10.1016/S2214-109X(18)30386-3. URL `https://www.sciencedirect.com/science/article/pii/S2214109X18303863`.

[3] Luke N Allen, Luisa M Pettigrew, Josephine Exley, Harry Collin, Shona Bates, and Michael Kidd. Global health inequity and primary care. *BJGP Open*, 8(4):BJGPO.2024.0189, December 2024. ISSN 2398-3795. doi: 10.3399/BJGPO.2024.0189. URL `http://bjgpopen.org/lookup/doi/10.3399/BJGPO.2024.0189`.

[4] Jishnu Das, Jeffrey Hammer, and Kenneth Leonard. The Quality of Medical Advice in Low-Income Countries. *Journal of Economic Perspectives*, 22(2):93–114, March 2008. ISSN 0895-3309. doi: 10.1257/jep.22.2.93. URL `https://pubs.aeaweb.org/doi/10.1257/jep.22.2.93`.

[5] J Peabody, Riti Shimhkhada, Olusoji Adeyi, Wang Huihui, Edward Broughton, and Margaret Kirk. Chapter 10: Quality of Care. In *Disease Control Priorities, Third Edition (Volume 9): Improving*

Health and Reducing Poverty. World Bank Publications, December 2017. ISBN 978-1-4648-0528-8. Google-Books-ID: KVFDDwAAQBAJ.

[6] WHO. *Global Strategy on Digital Health 2020-2025*. World Health Organization, Geneva, 1st ed edition, 2021. ISBN 978-92-4-002092-4.

[7] Mohammad R Ali, Claire A Lawson, Angela M Wood, and Kamlesh Khunti. Addressing ethnic and global health inequalities in the era of artificial intelligence healthcare models: a call for responsible implementation. *Journal of the Royal Society of Medicine*, 116(8):260–262, August 2023. ISSN 0141-0768. doi: 10.1177/01410768231187734. URL https://doi.org/10.1177/01410768231187734. Publisher: SAGE Publications.

[8] Peter G. Brodeur, Thomas A. Buckley, Zahir Kanjee, Ethan Goh, Evelyn Bin Ling, Priyank Jain, Stephanie Cabral, Raja-Elie Abdulnour, Adrian D. Haimovich, Jason A. Freed, Andrew Olson, Daniel J. Morgan, Jason Hom, Robert Gallo, Liam G. McCoy, Haadi Mombini, Christopher Lucas, Misha Fotoohi, Matthew Gwiazdon, Daniele Restifo, Daniel Restrepo, Eric Horvitz, Jonathan Chen, Arjun K. Manrai, and Adam Rodman. Superhuman performance of a large language model on the reasoning tasks of a physician, May 2025. URL http://arxiv.org/abs/2412.10849. arXiv:2412.10849 [cs].

[9] Stephanie Cabral, Daniel Restrepo, Zahir Kanjee, Philip Wilson, Byron Crowe, Raja-Elie Abdulnour, and Adam Rodman. Clinical Reasoning of a Generative Artificial Intelligence Model Compared With Physicians. *JAMA Internal Medicine*, 184(5):581–583, May 2024. ISSN 2168-6106. doi: 10.1001/jamainternmed.2024.0295. URL https://doi.org/10.1001/jamainternmed.2024.0295.

[10] Daniel McDuff, Mike Schaekermann, Tao Tu, Anil Palepu, Amy Wang, Jake Garrison, Karan Singhal, Yash Sharma, Shekoofeh Azizi, Kavita Kulkarni, Le Hou, Yong Cheng, Yun Liu, S. Sara Mahdavi, Sushant Prakash, Anupam Pathak, Christopher Semturs, Shwetak Patel, Dale R. Webster, Ewa Dominowska, Juraj Gottweis, Joelle Barral, Katherine Chou, Greg S. Corrado, Yossi Matias, Jake Sunshine, Alan Karthikesalingam, and Vivek Natarajan. Towards accurate differential diagnosis with large language models. *Nature*, pages 1–7, April 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-08869-4. URL https://www.nature.com/articles/s41586-025-08869-4. Publisher: Nature Publishing Group.

[11] Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. Capabilities of GPT-4 on Medical Challenge Problems, April 2023. URL http://arxiv.org/abs/2303.13375. arXiv:2303.13375 [cs].

[12] 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, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Agüera y Arcas, Nenad Tomašev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle K. Barral, Dale R. Webster, Greg S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. Toward expert-level medical question answering with large language models. *Nature Medicine*, pages 1–8, January 2025. ISSN 1546-170X. doi: 10.1038/s41591-024-03423-7. URL https://www.nature.com/articles/s41591-024-03423-7. Publisher: Nature Publishing Group.

[13] Tao Tu, Anil Palepu, Mike Schaekermann, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Nenad Tomasev, Shekoofeh Azizi, Karan Singhal, Yong Cheng, Le Hou, Albert Webson, Kavita Kulkarni, S. Sara Mahdavi, Christopher Semturs, Juraj Gottweis, Joelle Barral, Katherine Chou, Greg S. Corrado, Yossi Matias, Alan Karthikesalingam, and Vivek Natarajan. Towards Conversational Diagnostic AI, January 2024. URL http://arxiv.org/abs/2401.05654. arXiv:2401.05654 [cs].

[14] Selin S. Everett, Bryan J. Bunning, Priyank Jain, Ivan Lopez, Anup Agarwal, Manisha Desai, Robert Gallo, Ethan Goh, Vinay B. Kadiyala, Zahir Kanjee, Jacob M. Koshy, Andrew Olson, Adam Rodman, Kevin Schulman, Eric Strong, Jonathan H. Chen, and Eric Horvitz. From Tool to Teammate: A Randomized Controlled Trial of Clinician-AI Collaborative Workflows for Diagnosis, June 2025. URL https://www.medrxiv.org/content/10.1101/2025.06.07.25329176v1. Pages: 2025.06.07.25329176.

[15] Ethan Goh, Robert J. Gallo, Eric Strong, Yingjie Weng, Hannah Kerman, Jason A. Freed, Joséphine A. Cool, Zahir Kanjee, Kathleen P. Lane, Andrew S. Parsons, Neera Ahuja, Eric Horvitz, Daniel Yang, Arnold Milstein, Andrew P. J. Olson, Jason Hom, Jonathan H. Chen, and Adam Rodman. GPT-4 assistance for improvement of physician performance on patient care tasks: a randomized controlled trial. *Nature Medicine*, pages 1–6, February 2025. ISSN 1546-170X. doi: 10.1038/s41591-024-03456-y. URL https://www.nature.com/articles/s41591-024-03456-y. Publisher: Nature Publishing Group.

[16] Ethan Goh, Robert Gallo, Jason Hom, Eric Strong, Yingjie Weng, Hannah Kerman, Joséphine A. Cool, Zahir Kanjee, Andrew S. Parsons, Neera Ahuja, Eric Horvitz, Daniel Yang, Arnold Milstein, Andrew P. J. Olson, Adam Rodman, and Jonathan H. Chen. Large Language Model Influence on Diagnostic Reasoning: A Randomized Clinical Trial. *JAMA Network Open*, 7(10):e2440969, October 2024. ISSN 2574-3805. doi: 10.1001/jamanetworkopen.2024.40969. URL https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2825395.

[17] Robert Korom, Sarah Kiptinness, Najib Adan, Kassim Said, Catherine Ithuli, Oliver Rotich, Boniface Kimani, Irene King'ori, Stellah Kamau, Elizabeth Atemba, Muna Aden, Preston Bowman, Michael Sharman, Rebecca Soskin Hicks, Rebecca Distler, Johannes Heidecke, Rahul K. Arora, and Karan Singhal. AI-based Clinical Decision Support for Primary Care: A Real-World Study, July 2025. URL http://arxiv.org/abs/2507.16947. arXiv:2507.16947 [cs] version: 1.

[18] Pawel Gmyrek, Janine Berg, Karol Kamiński, Filip Konopczyński, Agnieszka Ładna, Balint Nafradi, Konrad Rosłaniec, Marek Troszyński, and International Labour Organization. Research Department. *Generative AI and jobs: a refined global index of occupational exposure*. ILO, Geneva, 2025. ISBN 978-92-2-042184-0. doi: 10.54394/HETP0387. URL https://researchrepository.ilo.org/esploro/outputs/encyclopediaEntry/995653520102676.

[19] Ardi Findyartini, Lesleyanne Hawthorne, Geoff McColl, and Neville Chiavaroli. How clinical reasoning is taught and learned: Cultural perspectives from the University of Melbourne and Universitas Indonesia. *BMC Medical Education*, 16(1):185, July 2016. ISSN 1472-6920. doi: 10.1186/s12909-016-0709-y. URL https://doi.org/10.1186/s12909-016-0709-y.

[20] C. Metallo, R. Agrifoglio, L. Lepore, and L. Landriani. Explaining users' technology acceptance through national cultural values in the hospital context. *BMC Health Services Research*, 22(1):84, January 2022. ISSN 1472-6963. doi: 10.1186/s12913-022-07488-3. URL https://doi.org/10.1186/s12913-022-07488-3.

[21] John W. Peabody, Jeff Luck, Peter Glassman, Timothy R. Dresselhaus, and Martin Lee. Comparison of Vignettes, Standardized Patients, and Chart AbstractionA Prospective Validation Study of 3 Methods for Measuring Quality. *JAMA*, 283(13):1715–1722, April 2000. ISSN 0098-7484. doi: 10.1001/jama.283.13.1715. URL https://doi.org/10.1001/jama.283.13.1715.

[22] J. M. Grimshaw and I. T. Russell. Effect of clinical guidelines on medical practice: a systematic review of rigorous evaluations. *The Lancet*, 342(8883):1317–1322, November 1993. ISSN 0140-6736. doi: 10.1016/0140-6736(93)92244-N. URL https://www.sciencedirect.com/science/article/pii/014067369392244N.

[23] Ethan Goh, Bryan Bunning, Elaine C. Khoong, Robert J. Gallo, Arnold Milstein, Damon Centola, and Jonathan H. Chen. Physician clinical decision modification and bias assessment in a randomized controlled trial of AI assistance. *Communications Medicine*, 5(1):59, March 2025. ISSN 2730-664X. doi: 10.1038/s43856-025-00781-2.

[24] John W. Peabody and Anli Liu. A cross-national comparison of the quality of clinical care using vignettes. *Health Policy and Planning*, 22(5):294–302, September 2007. ISSN 0268-1080. doi: 10.1093/heapol/czm020.

[25] John W. Peabody, Jeff Luck, Peter Glassman, Sharad Jain, Joyce Hansen, Maureen Spell, and Martin Lee. Measuring the Quality of Physician Practice by Using Clinical Vignettes: A Prospective Validation Study. *Annals of Internal Medicine*, 141(10):771, November 2004. ISSN 0003-4819. doi: 10.7326/0003-4819-141-10-200411160-00008. URL http://annals.org/article.aspx?doi=10.7326/0003-4819-141-10-200411160-00008.

[26] John W. Peabody, Lisa DeMaria, Owen Smith, Angela Hoth, Edmond Dragoti, and Jeff Luck. Large-Scale Evaluation of Quality of Care in 6 Countries of Eastern Europe and Central Asia Using Clinical Performance and Value Vignettes. *Global Health: Science and Practice*, 5(3):412–429, September 2017. ISSN 2169-575X. doi: 10.9745/GHSP-D-17-00044. URL `http://www.ghspjournal.org/lookup/doi/10.9745/GHSP-D-17-00044`.

[27] Shakked Noy and Whitney Zhang. Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654), 2023. doi: 10.1126/science.adh2586. URL `https://www.science.org/doi/10.1126/science.adh2586`.

[28] Dale F. Whelehan, Kevin C. Conlon, and Paul F. Ridgway. Medicine and heuristics: cognitive biases and medical decision-making. *Irish Journal of Medical Science*, 189(4):1477–1484, November 2020. ISSN 1863-4362. doi: 10.1007/s11845-020-02235-1.

[29] Ihsan Ayyub Qazi, Ayesha Ali, Asad Ullah Khawaja, Muhammad Junaid Akhtar, Ali Zafar Sheikh, and Muhammad Hamad Alizai. Automation Bias in Large Language Model Assisted Diagnostic Reasoning Among AI-Trained Physicians, September 2025. URL `https://www.medrxiv.org/content/10.1101/2025.08.23.25334280v2`. ISSN: 3067-2007 Pages: 2025.08.23.25334280.

[30] Ekaterina Jussupow, Kai Spohrer, Armin Heinzl, and joshua Gawlitza. Augmenting Medical Diagnosis Decisions? An Investigation into Physicians' Decision-Making Process with Artificial Intelligence. *Information Systems Research*, 32(3), 2021. doi: 10.1287/isre.2020.0980. URL `https://pubsonline.informs.org/doi/epdf/10.1287/isre.2020.0980`.

Author Contributions

Levels and Rounding had full access to all of the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.

Conceptualisation: All authors.

Data Curation: Rounding, Dijksman, Steens.

Formal Analysis: Rounding.

Funding Acquisition: Levels, Rounding, Berg, Gmyrek, Velasco.

Investigation: Rounding, Levels, Fregin, Steens, Dijksman, Arif, Findyartini, Soemantri, Pujitresnani, Greviana, Yusuf, Manji, Mbithi, Obungu, Rianga, Sokwalla.

Methodology: Rounding, Levels, Fregin, Steens, Dijksman, Arif, Findyartini, Wildan, Leijenaar, De Bont, Cals, Manji, Sokwalla.

Project Administration: Rounding, Levels, Fregin, Steens, Dijksman, De Boer, Arif, Findyartini, Soemantri, Pujitresnani, Greviana, Yusuf, Manji, Mbithi, Obungu, Rianga, Sokwalla, Velasco, Berg.

Resources: N/A.

Software: Rounding, Steens, Dijksman.

Supervision: Levels, Sokwalla, Findyartini, Fouarge.

Validation: Levels, Dijksman, Steens, Fregin.

Visualisation: Rounding.

Writing – Original Draft: Rounding.

Writing – Review & Editing: Levels, Cals, De Bont, Leijenaar, Arif, Findyartini, Soemantri, Pujitresnani, Manji, Gmyrek, Fouarge, Fregin, De Boer.

Data Sharing Statement

Data types and scope: De-identified participant-level data (demographics, survey response, graded vignette response, outcomes), analysis scripts (Stata17).

Documentation: Data dictionary in CSV; annotated Stata scripts.

Repository and access: All materials archived in DataverseNL (`https://dataverse.nl`), restricted access.

Timing: Available upon article publication.

Use permissions: Available for reproducibility or collaborative academic research, taking into account privacy constraints. Usage only under legal framework.

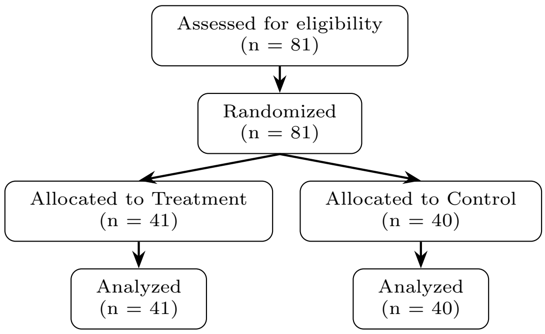
Contact: Sanne Steens (s.steens@maastrichtuniversity.nl).

Ethics Statement

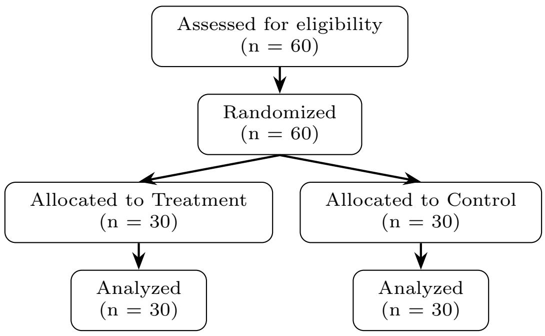
This research was conducted in accordance with the NETHICS Code of Ethics for Research in the Social and Behavioural Sciences Involving Human Participants anhd the Declaration of Helsinki. Ethical approval was sought and granted in each participating country. In Indonesia, it was provided by Komite Etik Penelitian Kesehatan Fakultas Kedokteran Universitas Indonesia (RSUPN) [Approval Number: Ket-1135/UN2.F1/ETIK/PPM.00.02/2024]. In Kenya by The Aga Khan University, Nairobi Institutional Scientific and Ethics Review Committee (ISERC) [Approval Number: 2024/ISERC-152 (v4)]. In the Netherlands by Maastricht University Ethical Review Board Inner City (UM ERCIC) [Approval Number ERCIC_572_25_04_2024]

LLM Use

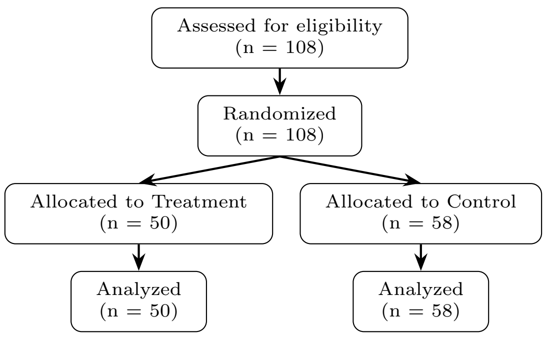
Figures and Tables



(a) Indonesia



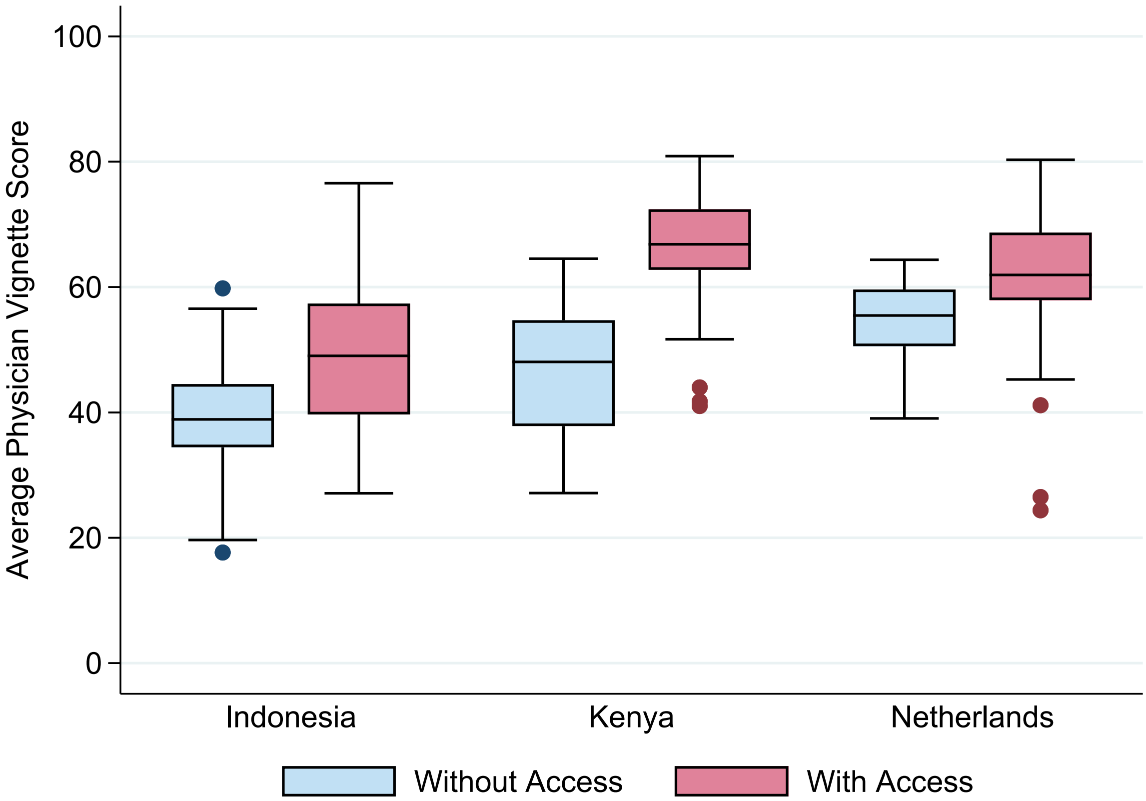
(b) Kenya



(c) The Netherlands

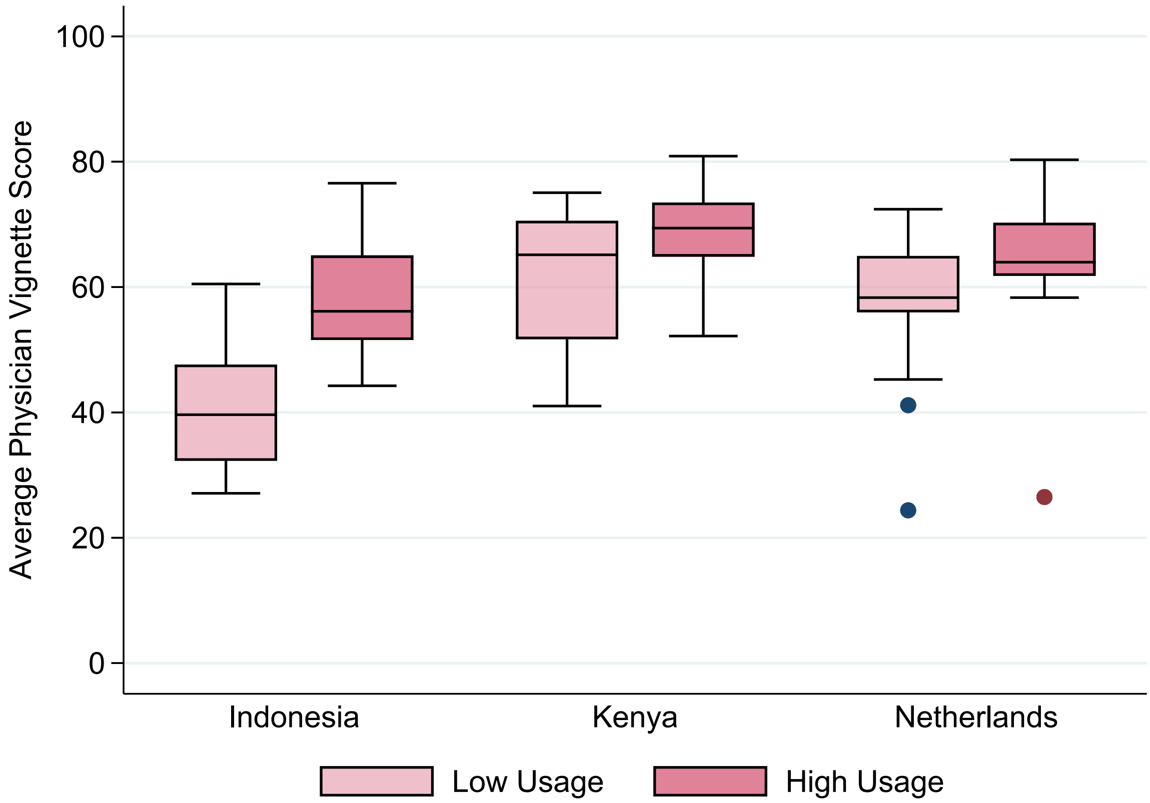
Figure 1: Participant Flow Diagrams by Country

Figure 2: Comparison of Average Physician Vignette Scores Across Countries With and Without LLM Access



Note: The figure shows the distribution of average physician vignette scores across the three countries, grouped by Without and With LLM access. Boxplots indicate score distributions: the boxes span the interquartile range (IQR), and whiskers extend to the minimum or maximum values within 1.5×IQR. Scores falling outside of the IQR are shown.

Figure 3: Comparison of Average Physician Vignette Scores Across Countries for High and Low within experiment LLM Usage



Note: The figure shows average physician vignette scores for the with access group in the three countries, grouped by low and high within experiment usage of LLMs. Boxplots indicate score distributions: the boxes span the interquartile range (IQR), and whiskers extend to the minimum or maximum values within 1.5×IQR. Scores falling outside of the IQR are shown.

Table 1: Baseline Characteristics

	Indonesia			**Kenya**			**The Netherlands**		
	Overall	Without Access	With Access	Overall	Without Access	With Access	Overall	Without Access	With Access
Observations	81	40	41	60	30	30	108	58	50
Career Stage									
Attending	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	0(0%)	17(16%)	9(16%)	8(16%)
Resident	81(100%)	40(100%)	41(100%)	60(100%)	30(100%)	30(100%)	91(84%)	49(84%)	42(84%)
Medical Specialty									
Internal/Family	81(100%)	40(100%)	41(100%)	27(45%)	9(30%)	18(60%)	108(100%)	58(100%)	50(100%)
Non-Internal	0(0%)	0(0%)	0(0%)	33(55%)	21(70%)	12(40%)	0(0%)	0(0%)	0(0%)
Years Since Start of Medical Education									
Mean	13(3.2)	13(3.5)	12.9(3)	11.8(3.5)	11.8(3.9)	11.8(3)	13.9(8.4)	13.3(8)	14.6(8.8)
Median	13(12-15)	13(11.5-14)	13(12-15)	11(10-13)	11(9-13)	11(10-13)	11(9-14)	11(9-13)	11(9-15)
Previous GenAI Use									
Daily	18(22%)	10(25%)	8(20%)	12(20%)	6(20%)	6(20%)	2(2%)	1(2%)	1(2%)
Weekly	29(36%)	16(40%)	13(32%)	27(45%)	10(33%)	17(57%)	12(11%)	7(12%)	5(10%)
Monthly	14(17%)	5(13%)	9(22%)	13(22%)	9(30%)	4(13%)	13(12%)	9(16%)	4(8%)
Rarely	12(15%)	5(13%)	7(17%)	4(7%)	3(10%)	1(3%)	40(37%)	17(29%)	23(46%)
Never	8(10%)	4(10%)	4(10%)	4(7%)	2(7%)	2(7%)	41(38%)	24(41%)	17(34%)

Table 2: Comparison of Average Physician Vignette Scores Across Countries With and Without LLM access

	Without Access	With Access	Mean Differences (CI 95%)
Panel A-Within Country Effects			
Indonesia			
Observations	40	41	
Raw Score(SD)	49.7(11.9)	62.6(16)	12.9(6.7-19.1), p<0.001
Mean%(SD)	39.2(9.5)	49.9(12.9)	10.7(5.7-15.7), p<0.001
Median%(IQR)	38.9(34.4-44.5)	49(39.7-57.4)	
Kenya			
Observations	30	30	
Raw Score(SD)	62.4(12.9)	85.8(13.7)	23.4(16.5-30.3), p<0.001
Mean%(SD)	47(10)	65(10.4)	18(12.7-23.2), p<0.001
Median%(IQR)	48.1(37.8-54.7)	66.8(62.7-72.4)	
Netherlands			
Observations	58	50	
Raw Score(SD)	62.4(7.4)	70.7(12.6)	8.30(4.3-12.4), p<0.001
Mean%(SD)	54.2(6.5)	61.4(10.9)	7.2(3.7-10.7), p<0.001
Median%(IQR)	55.5(50.6-59.6)	61.9(57.9-68.7)	
Panel B - Cross-National Effects			
Indonesia - Kenya			
Mean (95% CI)	7.8(3.1-12.5) p=0.004	15.1(9.6-20.6) p<0.001	7.3(0.1-14.5) p=0.14
Indonesia - Netherlands			
Mean (95% CI)	15(11.6-18.4) p<0.001	11.5(6.4-16.5) p<0.001	-3.5(-9.6-2.5) p=0.76
Kenya - Netherlands			
Mean (95% CI)	7.2(3.2-11.1) p=0.002	-3.6(-8.5-1.3) p=0.43	-10.8(-17–4.6) p=0.002

Note: Panel A presents within-country comparisons of the average physician vignette scores (raw and % correct) between control (Without Access) and intervention (With Access to GPT-4o) groups. Scores are reported as mean (standard deviation), median (interquartile range), and min-max difference (minimum-maximum values). Mean differences are estimated using two-sided linear ordinary least squares regressions with standard errors clustered at the participant level, reported with 95% confidence intervals and Bonferroni corrected p-values. Panel B presents cross-national comparisons of mean scores using linear OLS regressions without control variables. All statistical tests are two-sided.

Table 3: Comparison of Average Physician Vignette Scores Across Countries with High and Low Usage

		Low Usage	High Usage	Mean Differences (CI 95%)
Indonesia				
	Observations(%)	22(54%)	19(46%)	
	Mean(SD)	42.4(10)	58.5(10.4)	16.1(9.6-22.5), p<0.001
	Median(IQR)	40.6(32.6-49)	56.1(51.6-69.4)	
Kenya				
	Observations(%)	15(50%)	15(50%)	
	Mean(SD)	60.8(11.7)	69.2(7.2)	8.4(1.1-15.6), p0.08
	Median(IQR)	65.1(51.7-70.6)	69.4(64.9-73.5)	
Netherlands				
	Observations(%)	25(50%)	25(50%)	
	Mean(SD)	57.6(10.3)	65.1(10.3)	7.5(1.6-13.3), p0.04
	Median(IQR)	58.3(56-65)	64(61.8-70.3)	

Note: The table presents average physician percent correct vignette scores for participants in the intervention group (With Access to GPT-4o) stratified by usage intensity. Participants were classified as High Usage if their LLM usage rate exceeded the country-specific median, and Low Usage otherwise. Country-specific median usage rates were: Indonesia 0.44, Kenya 0.93, Netherlands 0.68 (measured as the proportion of vignette steps in which the LLM was consulted). Scores are reported as mean (standard deviation) and median (interquartile range). Mean differences are estimated using two-sided linear ordinary least squares regressions with standard errors clustered at the participant level, reported with 95