Supplementary Materials

Contents

1 Study Protocol **2**

2 Vignette and Rubric Details **8**

3 Kenya Sub-Group Analysis **16**

4 GPT-4o performance **20**

5 Sensitivity Analyses **21**

6 Vignette Analysis **29**

7 Prompting Instructions **31**

8 CONSORT Checklist **32**

Supplement 1: Study Protocol

Title

How Large Language Models Can Affect Clinical Reasoning: A Randomized Controlled Trial on Physicians' Use of Large Language Models

Study Design

- **Design:** Parallel-group randomized controlled trial using a superiority framework
- **Setting:** Clinical vignettes administered in computer labs at universities in Kenya, Indonesia, and the Netherlands
- **Procedure:** Physicians completed clinical vignettes—hypothetical but viable patient scenarios—via an online Qualtrics survey. Half of the participants were randomly assigned an LLM assistant (ChatGPT-4o via OpenAI API integrated into Qualtrics temperature of 1.0). The other half received no LLM assistance

Vignette Methodology

Participants were each presented with up to five fictional patient vignettes, each measuring quality of care using Peabody's (2000; 2004a; 2004b) framework. This framework provides a validated approach for evaluating healthcare quality using fictional patient vignettes that combine the high internal validity of experiments with the high external validity of surveys to unravel predictors of clinicians' decision-making (Evans et al., 2015; Sheringham, Kuhn, & Burt, 2021). The vignettes were identical in the three countries. The standard vignettes were chosen as the final differential diagnoses have a high prevalence worldwide; they do not require sophisticated technology or highly specialized care; and all had diagnostic strategies and treatment that were affordable and effective in all nations. They presented patient conditions that related to five different disease groups [cardiovascular, respiratory, musculoskeletal disease, fatigue and infectious disease and rare genetic disorders].

The fifth was chosen to be a 'white raven', where the final diagnosis is a low probability condition. The vignette was designed to be more complicated than appeared at first glance. Furthermore, the condition was not covered in single guidelines, but rather with advice scattered throughout various separate guidelines. We used only the four comparable vignettes.

Clinical Vignette Structure

1. Initial differential diagnosis (limited information)

2. Patient history
3. Additional differential diagnoses
4. Physical examination
5. Third differential diagnosis
6. Further investigations
7. Final differential diagnosis
8. Treatment or medication
9. Follow-up advice

Vignette Content

See Supplement 2

Trial Registration

- **Registry:** AEA Registry
- **Trial ID:** AEARCTR-0013399
- **Title:** The Big Unknown: A Journey into Generative AI's Transformative Effect on Professions, starting with Medical Practitioners
- **Registration Link:** `https://www.socialscienceregistry.org/trials/13399`
- **Initial Registration:** April 17, 2024
- **First Published:** April 25, 2024, 11:44 AM EDT
- **Registry**: NCT Registry
- **Trial ID**: NCT07374926
- **Title:** The Big Unknown: A Journey Into Generative AI's Transformative Effect on Medical Professions
- **Registration Link:** `https://clinicaltrials.gov/study/NCT07374926`
- **Published:** January 21, 2026

Ethical Approval

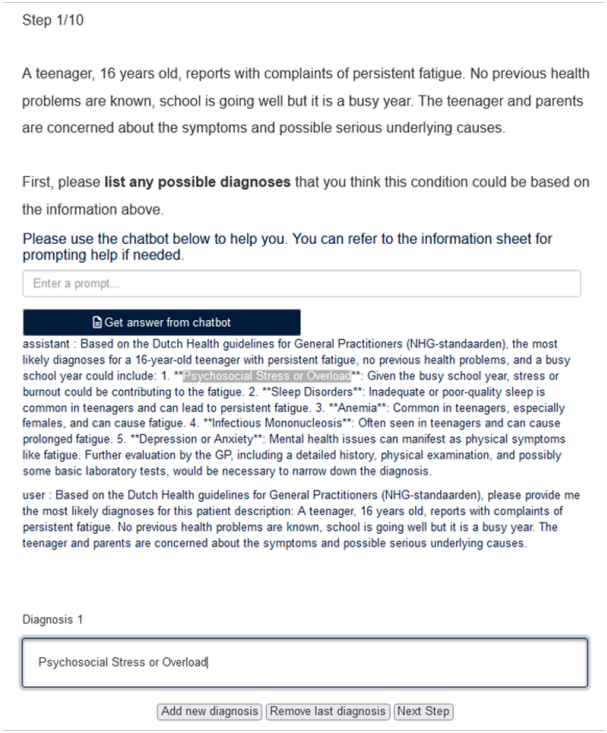
The study was reviewed and approved by review boards of the participating universities (University of Indonesia in Jakarta, Aga Kahn University Nairobi, and Maastricht University). Written informed consent was obtained from participants preceding enrolment and randomization. Participants were not compensated for participating in this study.

Study Design Details

- **Type:** Interventional (Clinical Trial)
- **Model:** Parallel group design, 1:1 allocation, superiority framework
- **Randomization:** Simple randomization using Qualtrics' randomizer (Evenly Present Elements)
- **Blinding:** Participants not blinded; graders were blinded to condition and identity
- **Setting:** University computer labs in Kenya, Indonesia, and the Netherlands
- **Eligibility:** Resident physicians (internal/family medicine), first-year residents in other specialties, and post-internship, pre-residency physicians with knowledge of primary care
- **Recruitment:** Participants were recruited via university networks of the medical teaching faculties of the three participating universities and were predominantly residents in internal and family medicine
- **Sample Size:** Indonesia: 81; Kenya: 60; the Netherlands: 108
- **Intervention:** Access to ChatGPT-4o via custom Qualtrics interface
- **Control:** No LLM assistant
- **Study start date data collection:** August 2024
- **Study end date data collection:** February 2025
- **Data collection by country:** Data was collected in August and October 2024 in the Netherlands; November 2024 in Indonesia; and January 2025 in Kenya

Study Arms

- **Active Comparator:** ChatGPT-4o API access (gpt-4o-2024-05-13 & gpt-4o-2024-08-06; temperature 0.5) via custom-built interface embedded in Qualtrics (iFrame), see Supplement Figure 1.1
- **No Intervention:** No additional resources



Supplement Figure 1.1: Example of GPT-4o integration into clinical vignettes

Outcome Measures

Grading: In each country, two independent graders assessed responses against a rubric. Graders were randomly assigned from a pool (Kenya: 4, Indonesia: 8, Netherlands: 11). Disagreements were resolved by a third reviewer. All final scores required agreement between at least two reviewers. Resolved disagreements are used for sensitivity analyses.

Primary Outcome: Percent correct score (rubric-based). Calculated as the weighted number of rubric items marked as present, divided by the total possible.

Note on Step 3: Many participants left Step 3 blank as the question was asked as "List any additional diagnoses". To adjust, rubric items marked in Step 1 were counted toward Step 3 if relevant. Sensitivity analyses will test robustness by removing Step 3 entirely.

Power Analysis

A two mean power analysis was conducted using Stata/SE 17.0 to determine the appropriate sample size for detecting meaningful effects in the randomized controlled trial. Using vignette parameters derived from the literature (Peabody et al, 2000), 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, treating each participant as a cluster and each vignette as an observation. 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%, based on literature estimates of

LLM's impact on physician performance papers (Bien et al., 2018; Han et al., 2020; Jain et al., 2021; Jussupow et al., 2021)

Statistical Analysis

Descriptive analyses were performed using Stata/SE 17.0. Initial descriptive results compare the group with LLM access, the group without the LLM access and the overall. 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 Table 5.5). Model assumptions (i.e. normalcy, homoskedasticity) were verified. Models were estimated using Stata/SE 17.0 package reghdfe, suitable for running linear regressions by implementing the estimator of Correia (2015). In the primary analysis no covariates were included (adding gender, GenAI use and experience as controls led to marginal differences in estimates but did not lead to substantive differences). 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.

We visually presented outcomes with boxplots that show the distribution of average participant vignette scores across the three countries, grouped by Without and With LLM access. The boxplots indicated score distributions: the boxes span the interquartile range (IQR), and whiskers extend to the minimum or maximum values within 1.5×IQR.

To evaluate cross-national differences in performance we ran linear OLS regressions (reghdfe) 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.

As an additional analysis that was not part of a prior analysis plan, we investigated the average use of the LLM per participant. We used the average of the percentage of steps per vignette for which the participant made use of the LLM. For Indonesia the median is 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). To distinguish between high and low use of the LLM we used for each country the median as a threshold. We then compared the impact of the LLM on the scores of high and low use participants.

Preplanned Sensitivity Analyses

To evaluate the empirical consequences of our modelling decisions we ran multiple sensitivity analyses. First, to evaluate the sensitivity of results to the reviewer process we reran all analyses on a dataset in which the dependent variable (quality of care score) was evaluated by the third expert reviewer (Supplement Table 5.1). Second, to assess the impact of cross-national differences in rubric items and weights (caused by cross-national difference in prevalence of conditions and differences in national guidelines), we reran all analyses on a dataset in which the dependent variable (quality of care score) was based on only rubric items that were present in all three countries (Supplement Table 5.2). Third, to assess the consequences of a necessary post hoc corrections of scores on the third vignette step, we reran all analyses on a dataset from which all scores on the third step were removed (Supplement Table 5.3). Fourth, to assess the consequences of our decision to correct for potential correlation of error terms across observations by clustering at the participant level we reran all analyses at the participant vignette level and estimated a mixed effect model with participant and vignette intercepts (Supplement Table 5.4). Finally, to assess the impact of reviewer disagreement we reran all analyses at the participant vignette level and removed all instances where the difference in the dependent variable was larger than 10 percentage points (Supplement Table 5.6).

Supplement 2: Vignette and Rubric Details

Supplement Table 2.1: Vignette 1: Respiratory Disease - Clinical Information Provided and Questions Asked

Step	Information Provided	Question Asked
1	A male, 58 years old, reports to GP with complaints of palpitations, described as "irregular and rapid".	First, please list any possible diagnoses that you think this condition could be based on the information above.
2	A male, 58 years old, reports to GP with complaints of palpitations, described as "irregular and rapid".	Please list the questions you would ask the patient, to begin your investigations.
3	The patient reports that the complaints have existed for several days, with no obvious onset. He does not experience chest pain, but reports fatigue and shortness of breath on exertion. He has a history of hypertension for which he takes hydrochlorothiazide.	Using this extra information, please list any additional possible diagnoses you think this could be. You can skip this question if you have no more.
4	*(Same as previous step)*	Given your differential diagnoses at this stage, please list the physical exams (if any) you would conduct to continue your investigation?
5	The results of physical tests performed are below: Blood pressure: 150/90 mmHg. Heart rate: Irregular, 110 beats per minute. Auscultation: Irregular heart sounds, no souffles. Lung auscultation: Slight crepitations at the base of both lungs, no wheezing. Physical examination: ankle oedema present. History of hyper tension.	What is your differential diagnosis now?
6	*(Same as previous step)*	Given your differential diagnosis, please list any additional research you would conduct.
7	The results of extra research are outlined below: ECG: confirms atrial fibrillation without further abnormalities. Additional blood tests for thyroid function, kidney function, iron status and electrolytes show no abnormalities. A chest radiograph to evaluate pulmonary status also showed no abnormalities.	Please list your final differential diagnoses, if you have more than one please put the most likely at the top.
8	*(Same as previous step)*	Would you prescribe any medication? If so, please list them.
9	*(Same as previous step)*	Would you discuss any further topics with this patient? If so, please list them.

Note: This table presents the clinical information provided to physicians and questions asked at each step of Vignette 1. Information marked as "(Same as previous step)" indicates that no new clinical information was provided at that step, and physicians were asked to respond based on previously presented information.

Supplement Table 2.2: Vignette 1: Respiratory Disease - Sample Rubric Items

Step	Sample Expert Answers	IND	KEN	NED
1	Other arrhythmia	1.00	1.00	1.00
	Atrial fibrillation	1.00	1.00	1.00
	Anxiety disorder	0.50	0.50	1.00
	Hyperthyroidism	0.50	1.00	0.50
	Cardiac ischemia	0.50	0.50	0.50
	...and 1 more items			
2	Do you have any other symptoms such as: chest pain, fatigue, or sho...	1.00	1.00	1.00
	Since when exactly have you been experiencing these symptoms?	1.00	1.00	1.00
	Have you been taking your blood pressure medication regularly?	1.00	1.00	1.00
	Are you suffering from a fever?	0.50	0.50	0.50
	Do you have other complaints on other tract such as coughing?	0.50	0.50	0.50
	...and 5 more items			
3	Other arrhythmia	1.00	1.00	1.00
	Atrial fibrillation	1.00	1.00	1.00
	Anxiety disorder	0.50	0.50	1.00
	Cardiac ischemia	0.50	0.50	0.50
	Hyperthyroidism	0.50	0.50	0.50
	...and 1 more items			
4	Signs of oedema	1.00	1.00	1.00
	Vital parameters (blood pressure, pulse rate)	1.00	1.00	1.00
	Inspection (signs of respiratory distress)	1.00	1.00	1.00
	Heart auscultation	0.50	1.00	0.50
	Lung auscultation	0.50	1.00	0.50
	...and 3 more items			
5	Other arrhythmia	1.00	1.00	1.00
	Atrial fibrillation	1.00	1.00	1.00
	Anxiety disorder	0.50	0.50	1.00
	Cardiac ischemia	0.50	0.50	0.50
	Hyperthyroidism	0.50	0.50	0.50
	...and 1 more items			
6	Consideration of a chest radiograph to evaluate pulmonary status	1.00	1.00	1.00
	ECG	1.00	1.00	1.00
	Additional blood tests for thyroid function, kidney function and el...	0.50	1.00	0.50
	Echocardiogram	0.50	0.50	0.50
	NTproBNP	—	1.00	—
7	Atrial fibrillation with possible heart failure	1.00	1.00	1.00
	Hypertension as a contributing factor	0.50	0.50	0.50
8	Medication: Start anticoagulant, preferably a NOAC, to prevent thro...	1.00	1.00	1.00
	Adding a beta-blocker to control heart rate.	1.00	1.00	1.00
	If needed, additional medication such as diuretics for symptom cont...	1.00	1.00	0.50
	Consider addition of an SGLT2 inhibitor for further improvement of ...	0.33	1.00	0.33
	If heart failure is confirmed on ultrasound: Start an ACE inhibitor...	0.50	0.50	0.50
	...and 2 more items			
9	Diet, exercise counselling	1.00	1.00	1.00
	Follow-up: Schedule follow-up after two weeks	0.50	0.50	0.50

Note: This table shows sample rubric items (acceptable expert answers) for Vignette 1. Only the first five items per step are shown for brevity. Scores indicate the weight assigned to each item in each country's rubric (IND=Indonesia, KEN=Kenya, NED=Netherlands). Higher scores reflect greater clinical importance.

Supplement Table 2.3: Vignette 2: Musculoskeletal Disease - Clinical Information Provided and Questions Asked

Step	Information Provided	Question Asked
1	A woman, 62 years old, reports with persistent low back pain for several weeks.	First, please list any possible diagnoses that you think this condition could be based on the information above.
2	A woman, 62 years old, reports with persistent low back pain for several weeks.	Please list the questions you would ask the patient, to begin your investigations.
3	The patient describes the pain as severe and continuous, worsening with movement and not completely disappearing at rest. No recent trauma is known, but she reports that the pain started after lifting a heavy object. She has also noticed that she seems to have become slightly smaller.	Using this extra information, please list any additional possible diagnoses you think this could be. You can skip this question if you have no more.
4	*(Same as previous step)*	Given your differential diagnoses at this stage, please list the physical exams (if any) you would conduct to continue your investigation?
5	The results of physical tests performed are below: Obvious pain on palpation over affected spinal regions. Limited mobility due to pain. Neurological examination: no obvious failure, reflexes normal.	What is your differential diagnosis now?
6	*(Same as previous step)*	Given your differential diagnosis, please list any additional research you would conduct.
7	The results of extra research are outlined below: The findings of the x-ray suggest a collapse fracture of a vertebra due to osteoporosis.	Please list your final differential diagnoses, if you have more than one please put the most likely at the top.
8	*(Same as previous step)*	Would you prescribe any medication? If so, please list them.
9	*(Same as previous step)*	Would you discuss any further topics with this patient? If so, please list them.

Note: This table presents the clinical information provided to physicians and questions asked at each step of Vignette 2. Information marked as "(Same as previous step)" indicates that no new clinical information was provided at that step, and physicians were asked to respond based on previously presented information.

Supplement Table 2.4: Vignette 2: Musculoskeletal Disease - Sample Rubric Items

Step	Sample Expert Answers	IND	KEN	NED
1	Collapse fracture of a vertebra due to osteoporosis	1.00	0.33	1.00
	Malignancy involving bone metastases	0.50	0.50	0.50
	Non-specific low back pain	0.50	0.50	0.50
	Lumbar hernia	0.50	0.50	0.33
2	Is there any history of trauma/fall?	1.00	1.00	1.00
	Pain characteristics	1.00	1.00	1.00
	Have you lost weight recently?	1.00	1.00	1.00
	Have you had any fractures in the past?	1.00	0.50	1.00
	Are you taking corticosteroids or other medications that may affect...	0.50	1.00	0.50
	...and 2 more items			
3	Collapse fracture of a vertebra due to osteoporosis	1.00	1.00	1.00
	Non-specific low back pain	0.50	0.50	0.50
	Malignancy involving bone metastases	0.50	0.50	0.50
	Lumbar hernia	0.50	0.50	0.33
4	Palpation	1.00	1.00	1.00
	Inspection back	1.00	1.00	1.00
	Neurological testsing	1.00	1.00	1.00
5	Collapse fracture due to osteoporosis	1.00	1.00	1.00
	Non-specific low back pain	1.00	0.33	1.00
	Malignancy involving bone metastases	0.50	1.00	0.50
	Bone TB	0.50	—	—
	TB spine	—	0.50	—
6	X-ray of lumbar spine	1.00	1.00	1.00
	In case of demonstrable collapse fracture Dexa scan to diagnose ost...	0.50	1.00	0.50
	Blood tests to rule out evidence of malignancy in which BSE, alkali...	0.50	1.00	0.50
	MRI lumbar spine	0.50	1.00	—
	PCR TB	0.50	—	—
7	The findings suggest a collapse fracture of a vertebra due to osteo...	1.00	1.00	1.00
	Non-specific low back pain	1.00	0.33	1.00
	Malignancy involving bone metastases	0.50	0.50	0.50
8	Supplements: Advice to take calcium and vitamin D supplements daily.	1.00	1.00	1.00
	Medication: Start a bisphosphonate to prevent further bone breakdow...	1.00	1.00	1.00
	Analgesics	1.00	1.00	1.00
	Surgical intervetion for pain management	0.50	0.50	—
9	Follow-up: Regular follow-up to monitor treatment effectiveness and...	1.00	1.00	1.00
	Lifestyle advice: Encourage weight-bearing and muscle-strengthening...	1.00	1.00	1.00
	Preventive measures: Possibly evaluate the home environment for ris...	1.00	1.00	1.00

Note: This table shows sample rubric items (acceptable expert answers) for Vignette 2. Only the first five items per step are shown for brevity. Scores indicate the weight assigned to each item in each country's rubric (IND=Indonesia, KEN=Kenya, NED=Netherlands). Higher scores reflect greater clinical importance.

Supplement Table 2.5: Vignette 3: Cardiovascular Disease - Clinical Information Provided and Questions Asked

Step	Information Provided	Question Asked
1	A male, 65 years old, reports with complaints of shortness of breath and wheezing. He is a known smoker.	First, please list any possible diagnoses that you think this condition could be based on the information above.
2	A male, 65 years old, reports with complaints of shortness of breath and wheezing. He is a known smoker.	Please list the questions you would ask the patient, to begin your investigations.
3	The patient has a history of smoking and has so far used only salbutamol for respiratory problems. The symptoms have recently worsened, which he relates to a recent respiratory infection.	Using this extra information, please list any additional possible diagnoses you think this could be. You can skip this question if you have no more.
4	*(Same as previous step)*	Given your differential diagnoses at this stage, please list the physical exams (if any) you would conduct to continue your investigation?
5	The results of physical tests performed are below: General appearance: Appears to be in respiratory distress. respiratory rate 25 breaths/min, SaO2 94%, T36.2 degrees Celsius. Auscultation of the lungs: Wheezing bilaterally, decreased breath sounds.	What is your differential diagnosis now?
6	*(Same as previous step)*	Given your differential diagnosis, please list any additional research you would conduct.
7	The results of extra research are outlined below: Spirometry does not show reversibility but does show persistent airway obstruction.	Please list your final differential diagnoses, if you have more than one please put the most likely at the top.
8	*(Same as previous step)*	Would you prescribe any medication? If so, please list them.
9	*(Same as previous step)*	Would you discuss any further topics with this patient? If so, please list them.

Note: This table presents the clinical information provided to physicians and questions asked at each step of Vignette 3. Information marked as "(Same as previous step)" indicates that no new clinical information was provided at that step, and physicians were asked to respond based on previously presented information.

Supplement Table 2.6: Vignette 3: Cardiovascular Disease - Sample Rubric Items

Step	Sample Expert Answers	IND	KEN	NED
1	COPD with exacerbation due to respiratory infection	1.00	1.00	1.00
	Lower respiratory tract infection/pneumonia	1.00	1.00	1.00
	Upper respiratory tract infection	1.00	0.50	1.00
	Malignancy	0.50	1.00	0.50
	Bronchiectasis	0.33	1.00	0.33
	...and 3 more items			
2	How long have you been experiencing these symptoms?	1.00	1.00	1.00
	Weight loss?	1.00	1.00	1.00
	Do you have a history of allergies or other respiratory conditions?	1.00	1.00	1.00
	Do you have a fever or productive cough? If so, what is the colour ...	1.00	1.00	0.33
	Have you recently been in contact with tuberculosis patients or bee...	1.00	1.00	—
	...and 2 more items			
3	COPD with exacerbation due to respiratory infection	1.00	1.00	1.00
	Upper respiratory tract infection	1.00	1.00	1.00
	Lower respiratory tract infection	1.00	1.00	1.00
	Malignancy	0.50	0.50	0.50
	Tuberculosis	1.00	0.50	—
	...and 3 more items			
4	Inspection (signs of respiratory distress)	1.00	1.00	1.00
	Vital parameters (blood pressure, pulse rate)	1.00	1.00	1.00
	Signs of oedema	0.50	1.00	1.00
	Lung auscultation	1.00	1.00	0.50
	Heart auscultation	0.50	1.00	0.50
	...and 1 more items			
5	Most likely COPD with exacerbation.	1.00	1.00	1.00
	Asthma, given wheezing and possible reaction to salbutamol.	0.50	0.50	0.50
	Lung malignancy	0.50	0.50	0.50
	Bronchiectasis and tuberculosis remain differential risk factors an...	0.50	0.50	—
6	X-ray	1.00	1.00	1.00
	Spirometry to measure the reversibility of airway obstruction and t...	1.00	0.50	1.00
	HRCT chest	1.00	1.00	—
	C-Reactive protein	0.50	0.50	0.50
	Sputum gene xpert	0.33	0.33	—
7	The symptoms and findings suggest an exacerbation of COPD, possibly...	1.00	1.00	1.00
8	Continuation of salbutamol for immediate relief of symptoms.	1.00	1.00	1.00
	Start a long-acting bronchodilator such as a LABA (long-acting beta...	0.50	1.00	0.50
	Antibiotics and/or corticosteroids orally if exacerbation is associ...	0.50	1.00	0.50
	Pulmonary rehabilitation	1.00	—	—
	Oxygen therapy	1.00	—	—
	...and 1 more items			
9	Strongly advise smoking cessation and offer support resources for t...	1.00	1.00	1.00
	Regular monitoring of airway symptoms and adjustment of treatment b...	0.50	0.50	0.50
	Regular follow-up with spirometry to monitor progression and respon...	0.50	0.50	0.50
	Promote physical activity adapted to the patient's capacity.	0.50	0.50	0.50
	Education on recognising exacerbations and the importance of early ...	0.33	0.33	0.33
	...and 1 more items			

Note: This table shows sample rubric items (acceptable expert answers) for Vignette 3. Only the first five items per step are shown for brevity. Scores indicate the weight assigned to each item in each country's rubric (IND=Indonesia, KEN=Kenya, NED=Netherlands). Higher scores reflect greater clinical importance.

Supplement Table 2.7: Vignette 4: Viral Infection - Clinical Information Provided and Questions Asked

Step	Information Provided	Question Asked
1	A teenager, 16 years old, reports with complaints of persistent fatigue. No previous health problems are known, school is going well but it is a busy year. The teenager and parents are concerned about the symptoms and possible serious underlying causes.	First, please list any possible diagnoses that you think this condition could be based on the information above.
2	A teenager, 16 years old, reports with complaints of persistent fatigue. No previous health problems are known, school is going well but it is a busy year. The teenager and parents are concerned about the symptoms and possible serious underlying causes.	Please list the questions you would ask the patient, to begin your investigations.
3	When you ask to specify the weight loss he tells you he lost 3-4kg in 2 months time. The symptoms started around 3 months ago and it is difficult for him to stay awake in classes at school. After school he needs to take a nap of 45 minutes and he goes to sleep at 10 PM after which he falls asleep rapidly.	Using this extra information, please list any additional possible diagnoses you think this could be. You can skip this question if you have no more.
4	*(Same as previous step)*	Given your differential diagnoses at this stage, please list the physical exams (if any) you would conduct to continue your investigation?
5	The results of the physical test performed are below: General impression: Appears tired, pale, possibly emaciated, respiratory rate 22breaths/min, SaO2 99%, p90bpm, RR 106/80mmHg, T37.0 degrees Celsius. Palpation of thyroid is not abnormal, inspection of skin no peculiarities, oropharyngeal examination not abnormal. Pulmonary: bilateral normal vesicular breath sounds, no other lung sounds. Cardiac S1, S2 without murmur, rub or gallop. Abdomen: normal peristalsis, smooth, no abnormalities palpable and no pain. Lymph node stations: some mobile glands in the neck, further stations without abnormalities. Exploratory neurological examination without abnormalities. He felt feverish and has had a sore throat a couple of weeks ago and a loss of appetite but no other specific symptoms like muscle aches, night sweats or febrile episodes. He did lose 3-4kg of weight in 2 months time.	What is your differential diagnosis now?
6	*(Same as previous step)*	Given your differential diagnosis, please list any additional research you would conduct.
7	The results of extra research are outlined below: Blood tests show no abnormalities except slightly elevated sedimentation rate.	Please list your final differential diagnoses, if you have more than one please put the most likely at the top.
8	*(Same as previous step)*	Would you prescribe any medication? If so, please list them.
9	*(Same as previous step)*	Would you discuss any further topics with this patient? If so, please list them.

Note: This table presents the clinical information provided to physicians and questions asked at each step of Vignette 4. Information marked as "(Same as previous step)" indicates that no new clinical information was provided at that step, and physicians were asked to respond based on previously presented information.

Supplement Table 2.8: Vignette 4: Viral Infection - Sample Rubric Items

Step	Sample Expert Answers	IND	KEN	NED
1	Anemia	1.00	1.00	1.00
	Benign viral infection	0.50	1.00	1.00
	Malignancy (such as leukemia or lymphoma)	1.00	0.50	1.00
	Depression	0.50	0.50	0.50
	Hypocortisolism	0.33	0.50	0.33
	...and 4 more items			
2	How long have the complaints existed exactly?	1.00	1.00	1.00
	Are there other symptoms such as fever, night sweats or lymph node ...	1.00	1.00	1.00
	Is the patient taking medication or drugs?	1.00	1.00	1.00
	Are there any mood swings or depressive symptoms?	0.50	1.00	0.50
	Are there any changes in appetite or diet?	0.50	0.50	0.50
	...and 3 more items			
3	Viral infection remains a possibility.	1.00	0.50	1.00
	Anemia	0.50	0.50	0.50
	Hypothyroidism given pallor and fatigue.	0.50	0.50	0.50
	Malignancy should be ruled out given the symptoms of weight loss an...	0.50	0.50	0.50
	Psychological factors such as depression, given the age profile.	0.50	0.50	0.50
	...and 1 more items			
4	Abdominal examination	1.00	1.00	1.00
	Lymph node palpation	1.00	1.00	1.00
	Heart and lung auscultation	1.00	1.00	1.00
	Vital parameters	1.00	1.00	1.00
	General inspection (bleeding/hematoma)	1.00	1.00	1.00
	...and 1 more items			
5	Viral infection remains a possibility.	1.00	1.00	1.00
	Anemia	0.50	1.00	0.50
	Hypothyroidism given pallor and fatigue.	0.50	1.00	0.50
	Malignancy should be ruled out given the symptoms of weight loss an...	0.50	0.50	0.50
	Psychological factors such as depression, given the age profile.	0.50	0.50	0.50
	...and 1 more items			
6	Thyroid function tests (TSH and free T4) to rule out hypothyroidism.	1.00	1.00	1.00
	Complete blood count (CBC) to identify anemia or signs of malignanc...	1.00	1.00	1.00
	Rate of erythrocyte sedimentation (ESR) or C-reactive protein (CRP)...	0.50	1.00	0.50
	Liver and kidney function tests.	0.50	0.50	0.50
	Possible additional examination such as a chest X-ray or abdominal ...	0.33	0.33	0.33
7	The history and physical examination findings support the hypothesi...	1.00	0.50	1.00
8	Medication and treatment: No specific medication; supportive care a...	1.00	1.00	1.00
9	Well-being and stress management	0.50	0.50	0.50
	Mental health evaluation	0.50	0.50	0.50
	Counselling	0.50	0.50	0.50
	Lifestyle advice: Encourage adequate rest, nutritious diet and hydr...	0.33	0.50	0.33

Note: This table shows sample rubric items (acceptable expert answers) for Vignette 4. Only the first five items per step are shown for brevity. Scores indicate the weight assigned to each item in each country's rubric (IND=Indonesia, KEN=Kenya, NED=Netherlands). Higher scores reflect greater clinical importance.

Supplement 3: Kenya Sub-Group Analysis

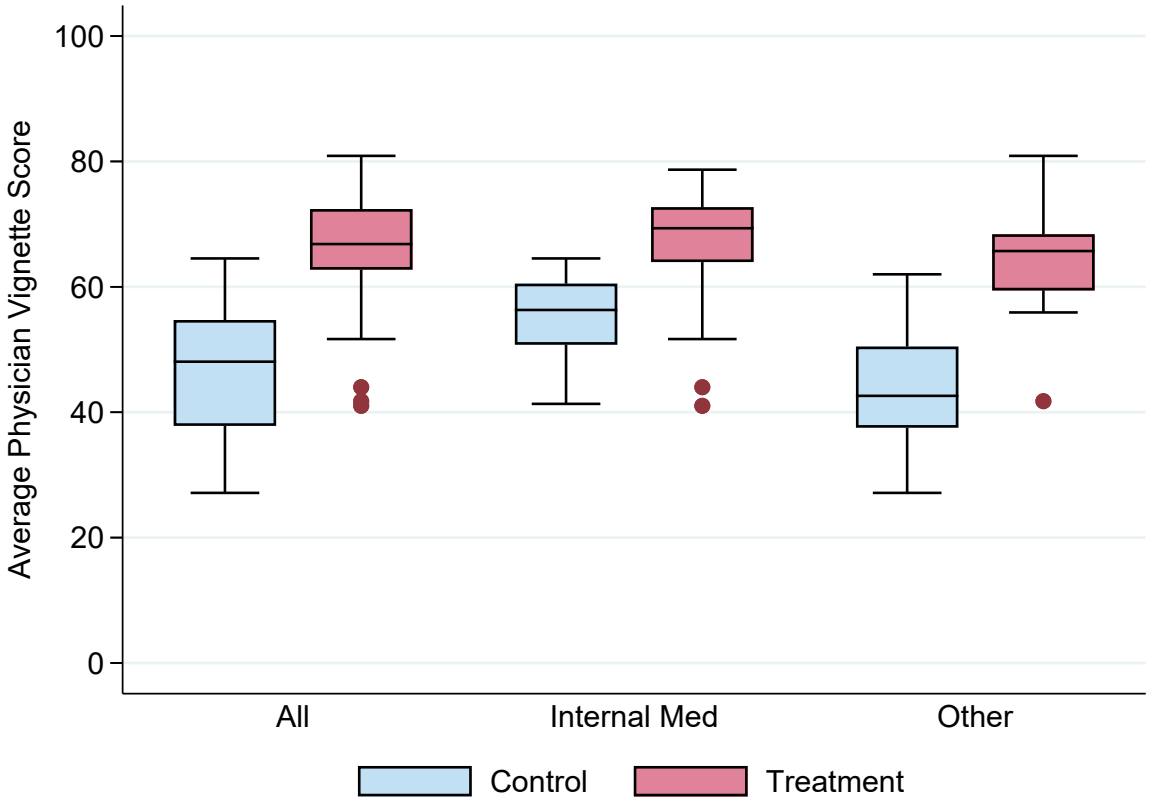
In this supplement, we assess the impact on our main findings for Kenya of the unbalanced sample distribution between two subgroups in our data: the Internal Medicine residents and residents specialising in other areas, or pre-residency physicians (referred to as non-specialists). We do so because the unbalanced sample distribution across intervention and control is potentially confounding for a number of conclusions we draw in the paper. This is driven by the fact that internal medicine residents may have been exposed to training on guidelines that cover the four patient conditions used in this study.

We explore the impact in Supplement Figure 3.1. In this study, we argue that LLM access contributed to increased homogeneity in the sample of Kenyan physicians with access to an LLM. The evidence we present in this supplement suggests that within the intervention group Internal Medicine specialists and non-specialists exhibit similar performance outcomes, with comparable medians and interquartile ranges (IQRs). By contrast, the group without LLM access show substantial differences between the two subgroups of Internal Medicine specialists and non-specialists, This indicates that LLM access may reduce performance differences/variability between Internal Medicine specialists and non-specialists.

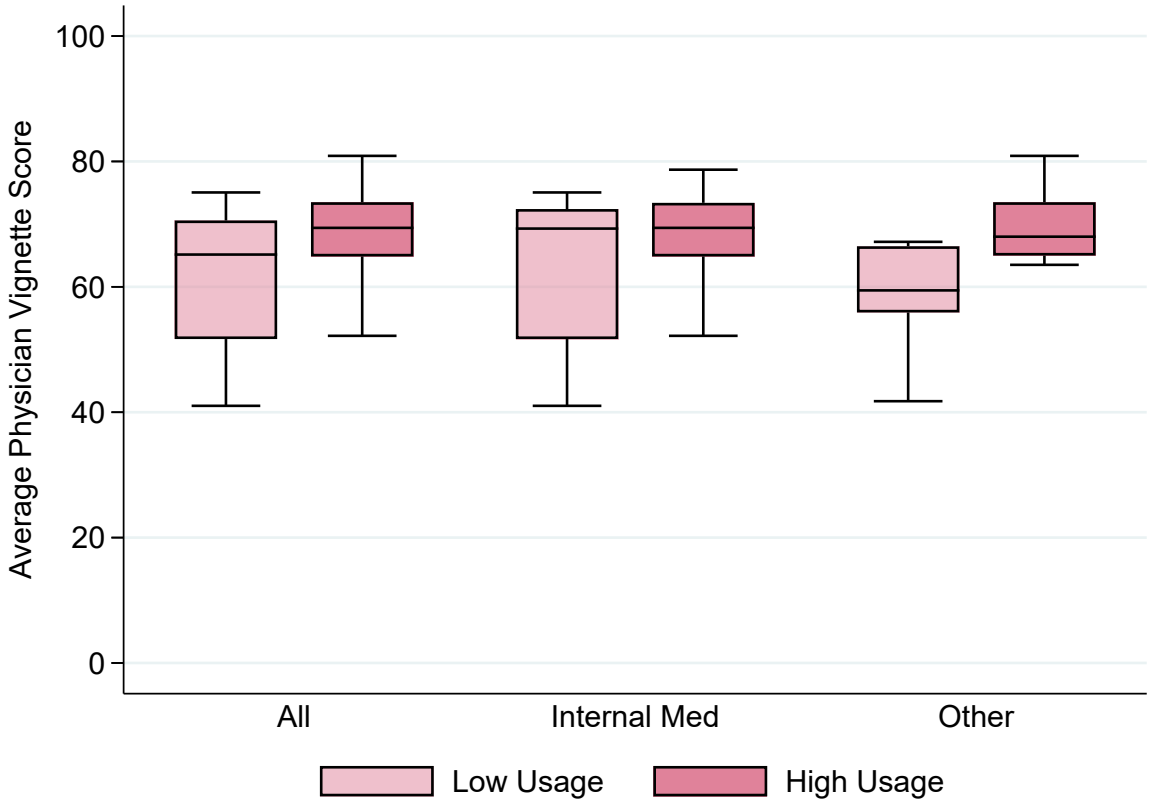
Supplement Table 3.1 reports the estimates of Table 1 in the main manuscript. Panel A reports on effects across three groups: the full sample, the Internal Medicine Specialists, and the non-specialist subgroup. As is to be expected, the intervention effect is larger for those in the non-specialist sub group, however, we still see a significant effect for the Internal Medicine subgroup. The estimated mean difference for the Internal Medicine subgroup in Kenya is 10.6% (95% CI:3.4-17.9, p=0.006) which is similar in magnitude to the effect found in Indonesia of 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).

Panel B reports the difference across the two subgroups. This shows that the gap between the two subgroups for physicians with access is 9.8% (95% CI: 0-19.7, p=0.004) smaller than for those without. Further, Kolomgorov-Smirnov tests confirm that the distributions of physicians specialising in Internal Medicine and the non-specialist subgroup are statistically significantly different for physicians without access to an LLM, while no such difference can be detected across the subgroup for the group with access . Taken together, these results are consistent with the hypothesis that LLM access reduces performance variability among physicians in Kenya.

We explore this further in Supplement Figure 3.2, which disaggregates the Kenyan intervention group into Internal Medicine and non-specialised subgroups and further stratifies each by high and low LLM usage. The figure shows that most of the outliers belong to the low-usage group across both subgroups. This pattern further supports our hypothesis that LLM access decreases performance variability in the Kenya sample. The increased homogeneity in the intervention group is driven by active usage of the LLM.



Supplement Figure 3.1: Comparison of Average Physician Vignette Scores Across Sub-Samples With and Without LLM Access



Supplement Figure 3.2: Comparison of Average Physician Vignette Scores Across Sub-Samples for High Usage and Low Usage

Supplement Table 3.1: Comparison of Average Physician Vignette Scores Across Specialisation With and Without LLM access

	Without Access	With Access	Mean Differences (CI 95%)
Panel A-Within Sample Effects			
All			
Observations	30	30	
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)	
Internal Med			
Observations	9	18	
Mean(SD)	54.9(7.1)	65.6(11.1)	10.6(3.4-17.9), p=0.006
Median(IQR)	56.3(50.8-60.5)	69.3(64-72.7)	
Other			
Observations	21	12	
Mean(SD)	43.6(9.1)	64.1(9.8)	20.5(13.5-27.5), p<0.001
Median(IQR)	42.6(37.5-50.5)	65.7(59.4-68.4)	
Panel B - Cross-Sample Effects			
Internal Med - Other			
Mean (95% CI)	-11.3(-17.6–5) p<0.001	-1.5(-9.4-6.4) p=0.70	9.8(0-19.7) p=0.004

Supplement 4: GPT-4o performance

As an additional analysis, we also provided our vignettes to GPT-4o exactly once using zero-shot prompting with a system prompt asking GPT-4o to be as concise as possible and to list the answers. The answers provided by GPT-4o were then also given to the graders to assess alongside the participant answers. Whilst LLMs are stochastic and we may expect some variation in responses, this exercise should provide a reasonable approximation of GPT-4o's performance on our vignettes. Supplement Table 4.1 demonstrates that GPT-4o outperformed the average physician score across all of our vignettes. Importantly, we also see that in each country the highest scoring physician outperformed the LLM. While those physicians may been provided responses that the LLM we used did not provide, it aligns with findings in other studies which show that the best performing physicians outperformed LLMs (Goh et al, 2025).

Supplement Table 4.1: Comparison of physicians With Access, Without Access and GPT-4o across Vignettes

	Average	Vignette 1	Vignette 2	Vignette 3	Vignette 4
Panel A: Indonesia					
Without Access	39.2	46.4	39.1	40.7	30.6
With Access	49.9	51.6	54.4	48.0	45.5
GPT-4o	66.7	67.9	80.2	63.4	55.3
Panel B: Kenya					
Without Access	47.0	45.8	50.1	54.6	37.5
With Access	65.0	63.9	72.2	64.9	59.1
GPT-4o	70.2	66.1	77.3	68.0	69.5
Panel C: Netherlands					
Without Access	54.2	49.1	54.5	59.0	54.2
With Access	61.4	57.7	61.2	65.0	61.6
GPT-4o	74.2	72.7	79.4	75.4	69.3

Note: The table presents average vignette scores (% correct) for physicians without LLM access (control group), physicians with LLM access (intervention group), and GPT-4o alone across three countries and four individual vignettes. GPT-4o scores represent the performance of the LLM when provided with vignette cases exactly as shown to physicians, using zero-shot prompting with a system prompt instructing GPT-4o to be concise and list answers. The Average column shows mean scores across all four vignettes. Vignettes covered globally prevalent primary care conditions: respiratory disease (Vignette 1), musculoskeletal disease (Vignette 2), cardiovascular disease (Vignette 3), and viral infection (Vignette 4). Scores were calculated as the weighted sum of correct rubric items divided by the weighted sum of total possible items. Panel A includes all participants in Indonesia (n=81); Panel B includes all participants in Kenya (n=60); Panel C includes all participants in the Netherlands (n=108).

Supplement 5: Sensitivity Analyses

Supplement Table 5.1: Primary outcome is with decisions of third expert reviewer

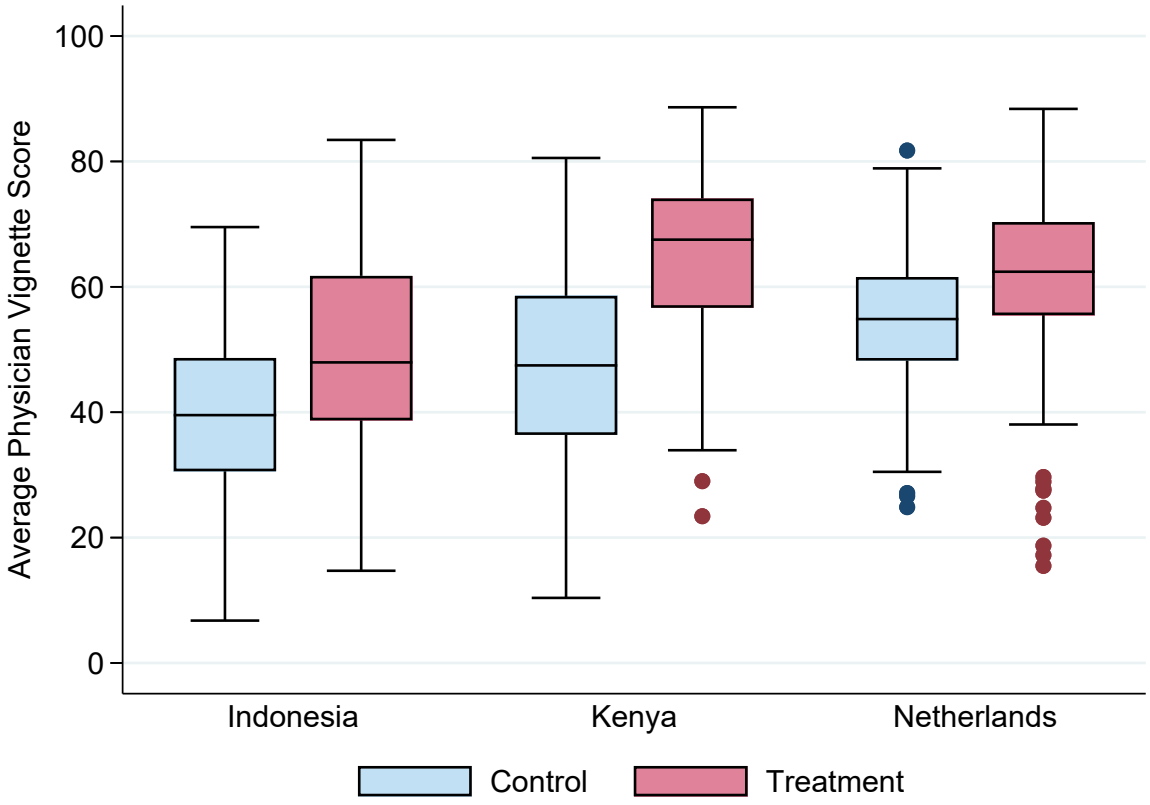
	Without Access	With Access	Mean Differences (CI 95%)
Panel A-Within Country Effects			
Indonesia			
Observations	40	41	
Mean(SD)	40.9(9.9)	52.7(14)	11.8(6.5-17.2) p<0.001
Median(IQR)	40.7(35.8-47.6)	52.6(42.6-60.4)	
Kenya			
Observations	30	30	
Mean(SD)	46.7(10)	65.5(11)	18.8(13.3-24.2) p<0.001
Median(IQR)	47.9(38.1-54.7)	67.3(62.3-71.8)	
Netherlands			
Observations	58	50	
Mean(SD)	54.8(6.7)	63.1(11.5)	8.3(4.6-12) p<0.001
Median(IQR)	56.4(51.1-59.8)	64(59.1-70.9)	
Panel B - Cross-National Effects			
Indonesia - Kenya			
Mean (95% CI)	5.8(1-10.6) p=0.02	12.8(6.8-18.7) p<0.001	6.9(-0.6-14.5) p=0.07
Indonesia - Netherlands			
Mean (95% CI)	13.9(10.3-17.4) p<0.001	10.4(5-15.8) p<0.001	-3.5(-9.9-2.9) p=0.28
Kenya - Netherlands			
Mean (95% CI)	8.1(4.1-12.1) p<0.001	-2.4(-7.5-2.8) p=0.36	-10.4(-16.9–4) p=0.002

Supplement Table 5.2: Results with consistent rubric across countries

	Without Access	With Access	Mean Differences (CI 95%)
Panel A-Within Country Effects			
Indonesia			
Observations	40	41	
Mean(SD)	41.7(10.2)	52.6(13.5)	10.9(5.6-16.2) p<0.001
Median(IQR)	41.2(36.6-47.2)	52.8(42.9-60.7)	
Kenya			
Observations	30	30	
Mean(SD)	49.2(10.2)	67.6(10.7)	18.4(13-23.8) p<0.001
Median(IQR)	50.1(39.7-57.9)	69.9(64.7-74.9)	
Netherlands			
Observations	58	50	
Mean(SD)	54.4(6.5)	61.6(10.9)	7.2(3.7-10.7) p<0.001
Median(IQR)	55.7(50.9-59.7)	62.2(58.2-69)	
Panel B - Cross-National Effects			
Indonesia - Kenya			
Mean (95% CI)	7.5(2.6-12.4) p=0.003	15(9.2-20.7) p<0.001	7.5(0-15) p=0.05
Indonesia - Netherlands			
Mean (95% CI)	12.7(9.1-16.3) p<0.001	9(3.8-14.2) p<0.001	-3.7(-10-2.6) p=0.25
Kenya - Netherlands			
Mean (95% CI)	5.2(1.2-9.3) p=0.01	-6(-10.9–1) p=0.02	-11.2(-17.6–4.9) p<0.001

Supplement Table 5.3: Results with Vignette Step 3 removed

	Without Access	With Access	Mean Differences (CI 95%)
Panel A-Within Country Effects			
Indonesia			
Observations	40	41	
Mean(SD)	38.4(10)	49.2(13.4)	10.7(5.5-16) p<0.001
Median(IQR)	38.4(33.4-43.5)	48.2(38.3-56.6)	
Kenya			
Observations	30	30	
Mean(SD)	46.4(10.1)	64.6(10.8)	18.2(12.8-23.6) p<0.001
Median(IQR)	46.9(37.4-54.4)	65.6(62.4-72.1)	
Netherlands			
Observations	58	50	
Mean(SD)	52.4(6.6)	60(11.1)	7.6(4.1-11.2) p<0.001
Median(IQR)	53.3(48.8-57.9)	60.4(56.3-67)	
Panel B - Cross-National Effects			
Indonesia - Kenya			
Mean (95% CI)	8(3.1-12.8) p=0.002	15.4(9.7-21.1) p<0.001	7.4(0-14.9) p=0.05
Indonesia - Netherlands			
Mean (95% CI)	14(10.4-17.6) p<0.001	10.9(5.7-16.1) p<0.001	-3.1(-9.4-3.2) p=0.33
Kenya - Netherlands			
Mean (95% CI)	6(2-10) p=0.004	-4.5(-9.5-0.5) p=0.08	-10.5(-16.9–4.2) p=0.001



Supplement Figure 5.1: Boxplots at Vignette Level

Supplement Table 5.4: Results at the Vignette Level

	Without Access	With Access	Mean Differences (CI 95%)
Panel A-Within Country Effects			
Indonesia			
Observations	160	164	
Mean(SD)	39.2(13.3)	49.9(15.3)	10.7(5.7-15.7) p<0.001
Median(IQR)	39.5(30.6-48.7)	47.9(38.7-61.8)	
Kenya			
Observations	120	120	
Mean(SD)	47(13.6)	65(13.1)	18(12.7-23.2) p<0.001
Median(IQR)	47.5(36.4-58.6)	67.5(56.7-74.1)	
Netherlands			
Observations	232	200	
Mean(SD)	54.2(10.1)	61.4(13.2)	7.2(3.7-10.7) p<0.001
Median(IQR)	54.9(48.2-61.6)	62.4(55.5-70.3)	
Panel B - Cross-National Effects			
Indonesia - Kenya			
Mean (95% CI)	7.8(3.1-12.5) p=0.001	15.1(9.6-20.6) p<0.001	7.3(0.1-14.4) p=0.05
Indonesia - Netherlands			
Mean (95% CI)	15(11.6-18.4) p<0.001	11.5(6.4-16.5) p<0.001	-3.5(-9.5-2.5) p=0.25
Kenya - Netherlands			
Mean (95% CI)	7.2(3.2-11.1) p<0.001	-3.6(-8.5-1.2) p=0.14	-10.8(-17–4.6) p<0.001

Supplement Table 5.5: Results using Mixed Effects model

	Without Access	With Access	Mean Differences (CI 95%)
Panel A-Within Country Effects			
Indonesia			
Observations	160	164	
Mean(SD)	39.2(13.3)	49.9(15.3)	10.7(5.8-15.6) p<0.001
Median(IQR)	39.5(30.6-48.7)	47.9(38.7-61.8)	
Kenya			
Observations	120	120	
Mean(SD)	47(13.6)	65(13.1)	18(12.9-23.1) p<0.001
Median(IQR)	47.5(36.4-58.6)	67.5(56.7-74.1)	
Netherlands			
Observations	232	200	
Mean(SD)	54.2(10.1)	61.4(13.2)	7.2(3.9-10.5) p<0.001
Median(IQR)	54.9(48.2-61.6)	62.4(55.5-70.3)	
Panel B - Cross-National Effects			
Indonesia - Kenya			
Mean (95% CI)	7.8(3.3-12.4) p<0.001	15.1(9.6-20.6) p<0.001	7.3(0.1-14.5) p=0.05
Indonesia - Netherlands			
Mean (95% CI)	15(11.9-18.1) p<0.001	11.5(6.7-16.3) p<0.001	-3.5(-9.2-2.2) p=0.23
Kenya - Netherlands			
Mean (95% CI)	7.2(3.8-10.6) p<0.001	-3.6(-8.4-1.2) p=0.14	-10.8(-16.6–5) p<0.001

Supplement Table 5.6: Results at the Vignette Level with vignettes with more than 10% difference between reviewers removed

	Without Access	With Access	Mean Differences (CI 95%)
Panel A-Within Country Effects			
Indonesia			
Observations	135	135	
Mean(SD)	39.4(13.3)	49.6(15.4)	10.5(5.3-15.6) p<0.001
Median(IQR)	39.6(30.7-49)	47.7(38.7-62.1)	
Kenya			
Observations	105	106	
Mean(SD)	46.8(13.9)	64.9(13.5)	17.9(12.3-23.6) p<0.001
Median(IQR)	47(35.9-58.1)	67.4(55.5-74.6)	
Netherlands			
Observations	212	180	
Mean(SD)	54(10)	61.3(13.3)	7.5(4-11.1) p<0.001
Median(IQR)	54.8(48-61.1)	62.4(55.5-70.3)	
Panel B - Cross-National Effects			
Indonesia - Kenya			
Mean (95% CI)	7.5(2.6-12.5) p=0.003	14.4(8.4-20.3) p<0.001	7.4(-0.1-15) p=0.05
Indonesia - Netherlands			
Mean (95% CI)	14.7(11.1-18.2) p<0.001	11.5(6.2-16.8) p<0.001	-2.9(-9.2-3.3) p=0.36
Kenya - Netherlands			
Mean (95% CI)	7.3(3.1-11.4) p<0.001	-3.3(-8.6-1.9) p=0.21	-10.5(-17.1–3.9) p=0.002

Supplement Table 5.7: Comparison of Average Physician Vignette Scores Across Countries With and Without LLM access controlling for covariates (gender, clinical experience, AI use and specialisation)

	Without Access	With Access	Mean Differences (CI 95%)
Panel A-Within Country Effects			
Indonesia			
Observations	40	41	
Mean(SD)	39.2(9.5)	49.9(12.9)	10.6(5.5-15.7) p<0.001
Median(IQR)	38.9(34.4-44.5)	49(39.7-57.4)	
Kenya			
Observations	30	30	
Mean(SD)	47(10)	65(10.4)	16.2(11-21.4) p<0.001
Median(IQR)	48.1(37.8-54.7)	66.8(62.7-72.4)	
Netherlands			
Observations	58	50	
Mean(SD)	54.2(6.5)	61.4(10.9)	7.4(4.2-10.6) 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)	15.7(10.2-21.3) p<0.001	14.8(7.3-22.4) p<0.001	5.6(-1.6-12.8) p=0.37
Indonesia - Netherlands			
Mean (95% CI)	14.3(10.5-18.1) p<0.001	11.1(5.8-16.4) p<0.001	-3.4(-9.4-2.7) p=0.81
Kenya - Netherlands			
Mean (95% CI)	-1.1(-6.8-4.7) p>0.99	0.8(-6.5-8.2) p>0.99	-9(-15.2- -2.8) p=0.01

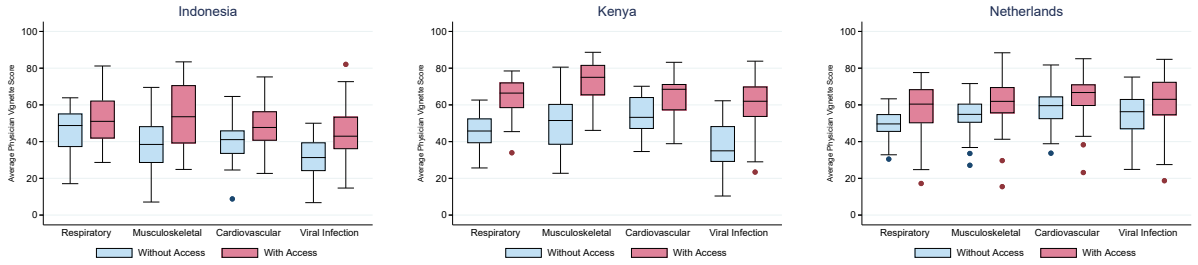
Supplement 6: Vignette Analysis

In this section we present analyses per vignette to examine whether particular vignettes drive the results observed across countries in our sample. Supplement Figure 6.1 displays the distribution of physician scores for each of the four vignettes, separated by treatment group within each country. The boxplots reveal variation in LLM effects across different clinical scenarios.

Supplement Table 6.1 presents the mean differences in percentage point scores between physicians with and without LLM access for each vignette-country combination. In Kenya, LLM access produced the largest improvements across all vignettes, with mean differences ranging from 10.3% (cardiovascular) to 22.0% (musculoskeletal). Indonesia showed substantial gains in musculoskeletal (15.3 points) and viral infection cases (14.9%), while the respiratory vignette showed a modest, statistically insignificant difference (5.3%, 95% CI: [-0.2, 10.7]). The Netherlands exhibited more uniform but smaller improvements across vignettes, ranging from 6.1% to 8.6%, with all confidence intervals excluding zero.

These results suggest that LLM effectiveness varies by clinical context and country setting, with musculoskeletal and viral infection cases showing particularly strong effects in lower-resource settings, while benefits the Netherlands were more modest but consistent across case types.

Supplement Figure 6.1: Average Physician Vignette Scores per Vignette by Country and LLM Access



(a) Indonesia (b) Kenya (c) Netherlands

Note: Each panel shows the distribution of physician scores for each of the four vignettes grouped by Without and With LLM access: respiratory disease (Vignette 1), musculoskeletal disease (Vignette 2), cardiovascular disease (Vignette 3), and viral infection (Vignette 4). 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.

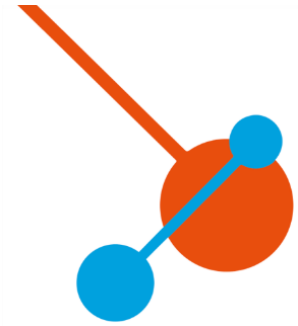
Supplement Table 6.1: Mean Difference in Physician Vignette Scores by Country and Vignette Type (With vs. Without LLM Access)

	Respiratory (Vignette 1)	**Musculoskeletal (Vignette 2)**	**Cardiovascular (Vignette 3)**	**Viral (Vignette 4)**
Indonesia				
Mean Difference	5.3	15.3	7.3	14.9
95% CI	[-0.2, 10.7]	[8.3, 22.3]	[2.1, 12.6]	[8.6, 21.2]
Kenya				
Mean Difference	18.0	22.0	10.3	21.6
95% CI	[12.7, 23.4]	[15.0, 29.1]	[4.8, 15.7]	[14.3, 28.8]
Netherlands				
Mean Difference	8.6	6.7	6.1	7.4
95% CI	[4.4, 12.8]	[2.4, 11.0]	[2.1, 10.0]	[2.2, 12.5]

Note: Mean differences represent the percentage point increase in correct scores for physicians with LLM access compared to those without, calculated separately for each vignette type within each country. 95% confidence intervals are shown in brackets.

Supplement 7: Prompting Instructions

Briefing Prompting AI in Primary Care



Huisartsenopleiding x ROA/ SBE

An Event for General Practitioners

Dear Participants,

Thank you for joining this experiment. In this study, we aim to evaluate the impact of AI on your diagnostic reasoning. During the session, you are allowed to use a chatbot to help you.

Tips on how to interact with the chatbot:

- Ask clear and specific questions.
- Include how you want the chatbot to present its answer (e.g., create a list that contains...).
- Specify how long the answer should be; this will influence the level of detail you will receive in the chatbot's answer.
- Give the chatbot one task or question at a time.
- You can copy and paste from the chatbot.
- The safeguards of the chatbot don't always allow you to get answers to specific questions, such as asking for medication advice, unless you prompt it creatively by e.g. describing your role and medical experience.

Tips on how NOT to interact with the chatbot:

- Avoid ambiguity and leading questions.
- Don't overload with information but specify the steps you require the chatbot to take.
- Don't repeat your questions if the answer was not right but build on your previous prompts.
- Don't expect perfection, always be critical of its answers and correct the chatbot when it makes a mistake.
- And very important: never feed the chatbot any personal data.

There are no wrong answers; please respond to the best of your knowledge. For the quality of the research, we also request that you refrain from using your phone or any other form of communication during the experiment. Additionally, please remain seated at your computer and avoid using any functionalities other than the browser tab provided for the experiment.

Thank you for your cooperation!

Supplement 8: CONSORT Checklist

Section/topic	No	CONSORT 2025 checklist item description	Reported on page no.
Title and abstract			
Title and structured abstract	1a	Identification as a randomised trial	1
	1b	Structured summary of the trial design, methods, results, and conclusions	1
Open science			1
Trial registration	2	Name of trial registry, identifying number (with URL) and date of registration	
Protocol and statistical analysis plan	3	Where the trial protocol and statistical analysis plan can be accessed	6
Data sharing	4	Where and how the individual de-identified participant data (including data dictionary), statistical code and any other materials can be accessed	8
Funding and conflicts of interest	5a	Sources of funding and other support (eg, supply of drugs), and role of funders in the design, conduct, analysis and reporting of the trial	1
	5b	Financial and other conflicts of interest of the manuscript authors	9
Introduction			
Background and rationale	6	Scientific background and rationale	2
Objectives	7	Specific objectives related to benefits and harms	2
Methods			
Patient and public involvement	8	Details of patient or public involvement in the design, conduct and reporting of the trial	n.a.
Trial design	9	Description of trial design including type of trial (eg, parallel group, crossover), allocation ratio, and framework (eg, superiority, equivalence, non-inferiority, exploratory)	6
Changes to trial protocol	10	Important changes to the trial after it commenced including any outcomes or analyses that were not prespecified, with reason	3
Trial setting	11	Settings (eg, community, hospital) and locations (eg, countries, sites) where the trial was conducted	5
Eligibility criteria	12a	Eligibility criteria for participants	5
	12b	If applicable, eligibility criteria for sites and for individuals delivering the interventions (eg, surgeons, physiotherapists)	n.a.
Intervention and comparator	13	Intervention and comparator with sufficient details to allow replication. If relevant, where additional materials describing the intervention and comparator (eg, intervention manual) can be accessed	6
Outcomes	14	Prespecified primary and secondary outcomes, including the specific measurement variable (eg, systolic blood pressure), analysis metric (eg, change from baseline, final value, time to event), method of aggregation (eg, median, proportion), and time point for each outcome	6-7
Harms	15	How harms were defined and assessed (eg, systematically, non-systematically)	1
Sample size	16a	How sample size was determined, including all assumptions supporting the sample size calculation	5-6
	16b	Explanation of any interim analyses and stopping guidelines	n.a.
Randomisation:			5
Sequence generation	17a	Who generated the random allocation sequence and the method used	
	17b	Type of randomisation and details of any restriction (eg, stratification, blocking and block size)	5

Allocation concealment mechanism	18	Mechanism used to implement the random allocation sequence (eg, central computer/telephone; sequentially numbered, opaque, sealed containers), describing any steps to conceal the sequence until interventions were assigned	5
Implementation	19	Whether the personnel who enrolled and those who assigned participants to the interventions had access to the random allocation sequence	5
Blinding	20a	Who was blinded after assignment to interventions (eg, participants, care providers, outcome assessors, data analysts)	5,7
	20b	If blinded, how blinding was achieved and description of the similarity of interventions	n.a.
Statistical methods	21a	Statistical methods used to compare groups for primary and secondary outcomes, including harms	7
	21b	Definition of who is included in each analysis (eg, all randomised participants), and in which group	5
	21c	How missing data were handled in the analysis	7
	21d	Methods for any additional analyses (eg, subgroup and sensitivity analyses), distinguishing prespecified from post hoc	3
Results			
Participant flow, including flow diagram	22a	For each group, the numbers of participants who were randomly assigned, received intended intervention, and were analysed for the primary outcome	5
	22b	For each group, losses and exclusions after randomisation, together with reasons	n.a.
Recruitment	23a	Dates defining the periods of recruitment and follow-up for outcomes of benefits and harms	5
	23b	If relevant, why the trial ended or was stopped	n.a.
Intervention and comparator delivery	24a	Intervention and comparator as they were actually administered (eg, where appropriate, who delivered the intervention/comparator, how participants adhered, whether they were delivered as intended (fidelity))	6
	24b	Concomitant care received during the trial for each group	n.a.
Baseline data	25	A table showing baseline demographic and clinical characteristics for each group	12
Numbers analysed, outcomes and estimation	26	For each primary and secondary outcome, by group: ● the number of participants included in the analysis ● the number of participants with available data at the outcome time point ● result for each group, and the estimated effect size and its precision (such as 95% confidence interval) ● for binary outcomes, presentation of both absolute and relative effect size	13
Harms	27	All harms or unintended events in each group	n.a.
Ancillary analyses	28	Any other analyses performed, including subgroup and sensitivity analyses, distinguishing pre-specified from post hoc	3
Discussion			**3-4**
Interpretation	29	Interpretation consistent with results, balancing benefits and harms, and considering other relevant evidence	**4**
Limitations	30	Trial limitations, addressing sources of potential bias, imprecision, generalisability, and, if relevant, multiplicity of analyses	

Citation: Hopewell S, Chan AW, Collins GS, Hróbjartsson A, Moher D, Schulz KF, et al. CONSORT 2025 Statement: updated guideline for reporting randomised trials. BMJ. 2025; 388:e081123. https://dx.doi.org/10.1136/bmj-2024-081123