Can Feedback from a Large Language Model Improve Health Care Quality?

Anja Sautmann

with Jason Abaluck, Robert Pless, Nirmal Ravi, Aaron Schwartz

AI & the Future of Human Capital in the Global South

September 29, 2025

Introduction

Introduction

Question: Can an LLM provide attending-physician style feedback to frontline health workers "out of the box"?

Introduction

Question: Can an LLM provide attending-physician style feedback to frontline health workers "out of the box"?

- Key for scalability across countries/settings
  - Range of (often low-skill) providers from public clinics to CHWs to informal care; frequently no documentation, no medical record keeping, and no training data
- Mimics private use by providers/patients

Introduction

Question: Can an LLM provide attending-physician style feedback to frontline health workers "out of the box"?

- Key for scalability across countries/settings
  - Range of (often low-skill) providers from public clinics to CHWs to informal care; frequently no documentation, no medical record keeping, and no training data
- Mimics private use by providers/patients

Implementation choices:

- Test with Community Health Extension Workers (CHEWs, $\tilde{3}$ years of training) at EHA Clinics in Nigeria
- CHEWs use clinic EMR, but: compiled into 'SOAP' note that's submitted for one-off feedback
- Prompt engineering:
  - Set up scenario & give basic guidelines, e.g. to reduce excessive medical testing (McPeak et al. 2024)
  - Focus on concise, actionable advice

Introduction

(General) challenge: measure quality of care improvement of intervention at patient contact

- Assessing effects across health problems and patients
- "True" patient status generally unknown
- Health outcome effects almost always small

Introduction

(General) challenge: measure quality of care improvement of intervention at patient contact

- Assessing effects across health problems and patients
- "True" patient status generally unknown
- Health outcome effects almost always small

- Our solution: Within-patient comparison of unassisted vs. LLM-assisted care plans
  - CHEW "unassisted" and "assisted" SOAP, plus independent SOAP from high-skill provider, and universal medical testing
  - Objective measure of patient welfare: CHEW decisions benchmarked against medical test results
  - Subjective measure: blinded assessment of potential harm caused by errors, by Medical Officers *who saw the patient*

Experimental Design and Estimation

Experimental Design

1. CHEW sees patient, creates "unassisted" SOAP note
2. Receives LLM feedback, creates "assisted" SOAP note

Experimental Design

1. CHEW sees patient, creates "unassisted" SOAP note
2. Receives LLM feedback, creates "assisted" SOAP note

3. Medical Officer (MO) sees patient independently and creates MO SOAP note
4. Medical testing for eligible patients: malaria, urine dipstick for UTI, PCV for anemia
5. MO creates final SOAP note ("groundtruth")

Experimental Design

1. CHEW sees patient, creates "unassisted" SOAP note
2. Receives LLM feedback, creates "assisted" SOAP note

3. Medical Officer (MO) sees patient independently and creates MO SOAP note
4. Medical testing for eligible patients: malaria, urine dipstick for UTI, PCV for anemia
5. MO creates final SOAP note ("groundtruth")

6. MO conducts blinded, randomized evaluation of treatment errors in CHEW SOAP notes

Sample Characteristics

	After Jan 30 Mean (SD)	After Feb 25 Mean (SD)
Share Female	0.64 (0.48)	0.66 (0.47)
Share Age 0-4	0.06 (0.24)	0.07 (0.25)
Share Age 5-14	0.11 (0.31)	0.13 (0.33)
Share Age 15-45	0.72 (0.45)	0.71 (0.45)
Share community clinic	0.49 (0.50)	0.48 (0.50)
Share CHEW would refer to MO	0.77 (0.42)	0.77 (0.42)
Share MO rated seriously ill	0.33 (0.47)	0.24 (0.42)
Share 'unassisted' is note A	0.62 (0.49)	0.49 (0.50)
N (Patients)	660	491
N (CHEWs)	28	28
N (MOs)	8	6

- Full sample 660, correctly randomized sample 491 patients
- More complex cases than typical community clinic
- Wide range of cases beyond maternal and newborn

Estimation

Effect of LLM feedback b estimated with patient fixed effects:

$$Y_{ik} = a_i + bT_k + cB_{ik} + \epsilon_{ik}$$

- Outcome Y_{ik}
- Indicator for "assisted" SOAP note T_k
- Patient i (or patient×condition): a_i fixed effect
- SOAP note shown to MO in randomly assigned order $k = \{A, B\}$
- Indicator for note rated second, B_{ik}: control for order effects (equiv. to period FE)

Results

Provider (CHEW) Response to LLM Feedback

How are CHEWs using and perceiving the LLM feedback?

Provider (CHEW) Response to LLM Feedback

How are CHEWs using and perceiving the LLM feedback?

	Mean (SD)
Fraction of notes with medical test change	0.329 (0.47)
Fraction of notes with Rx change	0.539 (0.50)
Fraction of notes with diagnostic code change	0.409 (0.49)
N	660

- CHEWs frequently change their treatment plan
- From pulse surveys: Overall positive sentiment towards LLM feedback, less than 1% error reported

Effect on MO Assessment of Quality of Care

How do high-skill providers rate the improvement in care for the patient?

Effect on MO Assessment of Quality of Care

How do high-skill providers rate the improvement in care for the patient?

- After seeing the patient, MOs rate the (blinded) CHEW notes
- Frame: treatment errors that could cause harm to patients

Effect on MO Assessment of Quality of Care

How do high-skill providers rate the improvement in care for the patient?

- After seeing the patient, MOs rate the (blinded) CHEW notes
- Frame: treatment errors that could cause harm to patients
- Qualitative error/harm:
  - Was there any error in the treatment plan?*
  - Temporary harm present: discomfort, additional symptoms?
  - Permanent harm present: potential for disability or loss of life?
  - Indicator for SOAP note judged to cause less harm
- DALY loss based on anchored harm scale:
  - Severe harm: error judged to be in top 5% of DALY loss*
  - Lower harm: indicator for strictly lower recorded DALY loss*

* Indicates pre-specified outcome indicator.

Effect on MO Assessment of Quality of Care

	Qualitative error/harm				DALY loss	
	Any error?*	Temp harm	Perm harm	Less harm	Severe loss	Lower loss
Effect of assisted note	-0.009	-0.018	-0.009	0.028	-0.011*	0.036
	(0.016)	(0.017)	(0.005)	(0.033)	(0.006)	(0.026)
Order effect	0.027*	0.018	-0.009	-0.072**	-0.002	-0.019
	(0.016)	(0.017)	(0.005)	(0.033)	(0.006)	(0.026)
Avg. unassisted note	0.642***	0.524***	0.086***	0.262***	0.054***	0.140***
	(0.012)	(0.012)	(0.004)	(0.023)	(0.004)	(0.018)
R-squared	0.01	0.00	0.01	0.01	0.01	0.01
N. of cases	900	900	900	900	900	900

SE in parentheses, $* \ p < 0.10$, $** \ p < 0.05$, $*** \ p < 0.01$.

- Plenty of room for improvement
- Positive effects, but small and not significant
- In 14% of cases, "assisted" SOAP judged to cause greater harm than "unassisted" SOAP

Objective Indicators of Treatment Quality for Malaria, Anemia, UTI

- Sample-wide screening with three most-used tests
  - Malaria: all patients; urine analysis for UTI: female patients 7 yrs. and older; PCV for anemia: all adults

Objective Indicators of Treatment Quality for Malaria, Anemia, UTI

- Sample-wide screening with three most-used tests
  - Malaria: all patients; urine analysis for UTI: female patients 7 yrs. and older; PCV for anemia: all adults
- Use test results as groundtruth to determine (prevented) patient harm from treatment and testing decisions:
  - Misallocation of testing: CHEW ordered test for those with a negative test and no MO test; CHEW did not order a test for those with a positive test
  - Misallocation of treatment*: those with a positive test received no or wrong treatment; those with a negative test received (some) treatment

* Indicates pre-specified outcome indicator.

Effect on the Misallocation of Medical Testing

	Malaria		UTI		PCV	
	Over-testing	Under-testing	Over-testing	Under-testing	Over-testing	Under-testing
Effect of assisted note	-0.037***	0.002	0.028***	-0.003	0.045***	-0.002
	(0.011)	(0.002)	(0.010)	(0.003)	(0.012)	(0.002)
Avg. unassisted note	0.327***	0.003***	0.062***	0.037***	0.073***	0.100***
	(0.006)	(0.001)	(0.005)	(0.001)	(0.006)	(0.001)
Share abnormal test	0.026	0.026	0.042	0.042	0.109	0.109
Misallocation MO	0.224	0.010	0.031	0.034	0.036	0.100
R squared	0.019	0.002	0.020	0.003	0.032	0.002
Patients	575	575	354	354	449	449

Standard errors in parentheses

* $p < .1$, ** $p < .05$, *** $p < .01$

- LLM feedback reduces (seasonal?) overtesting for malaria
- Increase in UTI and PCV tests does not catch the relevant patients: more overtesting, no less undertesting

Effect on the Misallocation of Medical Treatment for Malaria, Anemia, UTI

	(1) Aggregate	(2) Overtreatment	(3) Undertreatment
Effect of assisted note	-0.004	-0.004	-0.001
	(0.004)	(0.003)	(0.001)
Avg. unassisted note	0.075***	0.028***	0.047***
	(0.002)	(0.002)	(0.001)
R squared	0.001	0.001	0.000
Patient-Condition	1,378	1,378	1,378

Standard errors in parentheses

* $p < .1$, ** $p < .05$, *** $p < .01$

- Common problems in this population not identified better or treated more reliably
- Also very few effects on treatment plan inconsistencies such as missed tests, dosage errors (not shown)

Direct Analysis of LLM Feedback (Preliminary)

Why is the LLM not able to improve quality of care?

Direct Analysis of LLM Feedback (Preliminary)

Why is the LLM not able to improve quality of care?

	Mean (SD)
Suggestions for Changes to CHEW SOAP	2.831 (1.430)
Suggestions Closing Gap with MO	1.381 (1.339)
Suggestions Widening Gap with MO	0.726 (0.934)
N	658

Direct Analysis of LLM Feedback (Preliminary)

Why is the LLM not able to improve quality of care?

	Mean (SD)
Suggestions for Changes to CHEW SOAP	2.831 (1.430)
Suggestions Closing Gap with MO	1.381 (1.339)
Suggestions Widening Gap with MO	0.726 (0.934)
N	658

- SO part of CHEW SOAP seems to lead the LLM down the wrong path (not shown)
- CHEWs don't take the good advice – for identified differences between CHEW and MO note:
  - Less than 10% addressed by a pertinent LLM suggestion
  - In only 3% of instances, the CHEW followed the suggestion

Conclusion

Conclusion

- (Prompt-engineered) LLM received positively by providers and has significant effects on treatment plans
- BUT likely no strong impact on patient welfare
- Even 'low-hanging fruit' inconsistencies for common conditions not addressed
- A lot of LLM feedback irrelevant, misguided, or disregarded
- Caveats:
  - Outside of malaria season
  - EHA Clinics CHEWs better trained than average
  - CHEWs know their treatment plans are not implemented

Appendix

CHEW LLM Pulse Feedback

Table: Pulse Responses on LLM Feedback

	Mean (SD)
Fraction of CHEWs reporting 'LLM improves documentation'	0.952 (0.21)
Fraction of CHEWs reporting 'LLM helps provide better care'	0.950 (0.22)
Fraction of CHEWs Reporting 'LLM made an error'	0.005 (0.07)
N	642

Direct Analysis of LLM Feedback

(1) Diff. between 'unassisted' CHEW and MO SOAPs	Mean (SD)
Total Number of Differences	3.723 (2.558)
Diff. with Pertinent LLM Suggestion	0.324 (0.674)
Diff. where CHEW followed LLM Suggestion	0.123 (0.422)
(2) All LLM Suggestions	
Suggestions for Changes to CHEW SOAP	2.831 (1.430)
Suggestions Closing Gap with MO	1.381 (1.339)
Suggestions Widening Gap with MO	0.726 (0.934)
N	658

Effects on Treatment Plan Inconsistencies

	(1) Treatment without testing	(2) Prescription error	(3) Any Inconsistency
Effect of assisted note	-0.003	0.001	0.002
	(0.003)	(0.004)	(0.004)
Avg. unassisted note	0.026***	0.070***	0.086***
	(0.001)	(0.002)	(0.002)
R squared	0.000	0.000	0.000
Patient-Condition	1,980	1,980	1,980

Standard errors in parentheses

* $p < .1$, ** $p < .05$, *** $p < .01$

Note: Treatment without testing is coded as 1 whenever the CHEW intended to treat a condition but did not order the necessary diagnostic test. Prescription error is coded as 1 whenever the CHEW intended to treat a condition but prescribed an incorrect dosage or drug choice (including unnecessary medications). The dependent variable for any inconsistency is coded as 1 whenever either treatment without testing or prescription error occurred.