Pre-Analysis Plan

Improving AI-Assisted Medical Diagnosis and Triage by the General Public

Ihsan Ayyub Qazi, Ayesha Ali, Zafar Ayyub Qazi,
Abdullah Ghani, Muhammad Mahasin Irfan, Maryam Waseem

22nd July, 2026

1 Introduction

As large language models (LLMs) become widely accessible, a growing share of the public already turns to AI-powered chatbots for health-related information: surveys suggest that one in six American adults consults AI chatbots for health queries at least once a month [5, 6]. At the same time, diagnostic errors and misplaced triage decisions represent a persistent source of preventable patient harm globally [4]. This has prompted considerable interest in whether LLMs can serve as a reliable "front door" to the healthcare system for patients who lack immediate access to a clinician.

However, incidents involving misguided medical suggestions from LLMs to the general public such as fatal overdose [8] and erroneous diagnosis leading to life-threatening treatment delay [2] have been reported. A recent controlled study shows LLMs perform significantly worse for medical assistance in real-world settings compared to their performance on controlled benchmarks. Bean et al. [1] conducted a large preregistered study with 1,298 UK participants in which laypeople were assigned to receive assistance from one of three popular LLMs (GPT-4o, Llama 3, or Command R+) or to use a source of their choice when assessing ten standardized medical scenarios. Although the LLMs alone correctly identified relevant conditions in up to 94.9% of cases, participants using those same LLMs did so in fewer than 34.5% of cases, significantly *worse* than the group using their typical home resources (∼60%). Similarly, triage accuracy showed no significant difference between LLM users and controls, with an overall correct response rate of 43.0% across all groups. Qualitative analysis of interaction transcripts points towards two principal failure modes:

1. **Incomplete information provision.** Users typically provide LLMs with only a subset of their symptoms, omitting details critical to accurate triage (e.g., onset speed, severity, location). Because LLMs often respond to the information they were given rather than proactively eliciting what was missing, this frequently leads to incorrect or vague recommendations.

2. **Failure to interpret LLM output.** Even when LLMs mention correct conditions during conversations, users do not reliably incorporate those suggestions into their final responses, pointing to a breakdown in how recommendations are communicated and understood.

These findings motivate a direct intervention targeting both failure modes. We propose modifying the conversational behavior of LLM through a carefully designed system prompt that: (1) instructs the model to ask a structured set of clarifying questions before offering any medical suggestions (diagnosis and triage), thereby ensuring more complete symptom elicitation; and (2) requires the model to present its final output in a standardized, human-readable format (listing possible conditions, their likelihood, and a recommended course of action) to reduce the cognitive burden of interpretation.

This randomized controlled trial (RCT) will test whether participants with access to this *new* LLM interface alongside any other assistance methods outperform participants with access to a standard LLM and any assistance they would typically use at home (for example, internet search) on medically validated triage scenarios. We use a similar setup as the prior study for comparability. Participants are drawn from the student body and administrative staff of a university.

1.1 Research Question

Does access to a modified LLM interface, designed to elicit symptom information and deliver structured diagnostic output (alongside conventional resources), improve layperson diagnosis and triage accuracy relative to access to standard LLM interface (alongside the same conventional resources)?

1.2 Hypotheses

- **Primary Hypotheses:**
  - **H1:** Participants in the treatment group will identify medical conditions (diagnoses) relevant to a scenario at a higher rate than participants in the control group.
  - **H2:** Participants in the treatment group will correctly assess the urgency (triage) of medical scenarios at a higher rate than participants in the control group.
- **Secondary Hypotheses:**
  - **H3:** Participants in the treatment group will report different average self-reported confidence in their conclusions than the control group.
  - **H4:** Participants in the treatment group will spend less time per scenario relative to the control group.

2 Methods

2.1 Study Design

This is a two-arm, single-blind, randomized controlled trial:

- **Treatment Arm:** Participants can use any assistance methods they would typically employ at home in addition to a new LLM (GPT-4o) interface to complete medical scenarios. The new LLM interface uses a fixed system prompt that (a)

instructs the model to ask targeted clarifying questions before providing any diagnostic or triage suggestions, and (b) requires all final responses to follow a structured template listing: possible conditions, approximate likelihood of each, and a recommended triage with brief reasoning.

- **Control Arm:** Participants can use any assistance methods they would typically employ at home (e.g., Google or health portals) in addition to a standard LLM (GPT-4o) to complete medical scenarios. AI-overview in web searches will be disabled via an extension. They would not be allowed to access any LLMs other than the standard LLM interface.

Randomization: Participants will be randomly assigned to treatment or control in a 1:1 ratio.

Blinding: Participants will not be informed of which arm constitutes the "treatment" or what the study hypothesizes.

2.2 The new LLM Interface

The treatment interface wraps GPT-4o with a fixed system prompt.

2.2.1 Targeted-Questioning Protocol

Rather than responding to whatever partial information the user volunteers, the new interface is instructed to take a clinical history before offering any diagnostic suggestion. The model asks relevant questions and waits for the patient's response before proceeding, using plain language throughout. It is directed to ask only questions necessary for diagnosis and triage, and to accept uncertainty when the patient does not know an answer.

2.2.2 Structured Output

Once the model judges that sufficient history has been gathered, it provides a brief closing reassurance and immediately appends a structured JSON response containing: a differential diagnosis of up to three conditions each with a full condition name, likelihood rating (*likely / possible / unlikely*), key supporting reasoning, and a triage recommendation constrained to one of five levels matching the NHS entry-point scale. The JSON is parsed and rendered as a formatted, human-readable summary on the participant's screen, directly addressing the output interpretation failure mode.

2.3 Participants

2.3.1 Inclusion Criteria

- Student or employed administrative staff member at a university.
- Aged 18 years or older.
- Proficient in English (the language of the vignettes and interface).

2.3.2 Exclusion Criteria

- Individuals with any formal education or professional training in medicine, nursing, or any other healthcare profession.
- Participation in a prior pilot or testing phase of this study.

2.3.3 Incentive

- Participants will be financially compensated to participate in the study, which is expected to take 20 minutes per person.

2.4 Sample Size and Power Analysis

We power the study to detect a 15 percentage point difference between arms. Because H1 and H2 are co-primary hypotheses, we apply a Bonferroni correction, setting the per-test significance level to $\alpha = 0.025$. A two-sided two-proportion z-test with $\alpha = 0.025$ and 80% power requires 181 responses per arm.

Because each participant provides two responses, we adjust for within-participant clustering using an ICC of 0.20. This yields a design effect of $1 + (2 - 1) \times 0.20 = 1.20$, increasing the required sample to 218 responses per arm. This corresponds to 109 participants per arm. We therefore target recruitment of 110 participants per arm, for a total sample size of 220 participants.

2.5 Clinical Vignettes

We will use the same ten clinical vignettes developed and validated by Bean et al. [1]. These vignettes and their gold standard answers (triage and relevant conditions) were developed with input from five general practitioners and two intensive care physicians, covering a range of acuity levels:

- *Ambulance-level:* Pulmonary embolism, Pneumonia.
- *A&E (Accidents and Emergency)-level:* Subarachnoid haemorrhage, Renal colic.
- *Urgent primary care:* Anaemia, Ulcerative colitis.
- *Routine GP:* Tinnitus, Gallstones.
- *Self-care:* Allergic rhinitis, Common cold.

Scenario assignment. Each participant will be randomly assigned two of the ten scenarios to complete.

3 Outcome Measures

3.1 Primary Outcomes

- **Urgency.** The rating which participants give for the urgency of each scenario, measured on a five-point scale: Self-care, Routine GP, Urgent Primary Care, A&E, and Ambulance. Responses will be compared against the gold-standard answers to produce an accuracy measure.

- **Relevant Conditions.** Participants will be asked to name all medical conditions they considered relevant to their decision. A participant's response will be scored as correct for that scenario if at least one named condition matches the physician-generated gold-standard list.

3.2 Secondary Outcomes

- **Confidence.** Participants will report their confidence in their urgency assessment using a visual analog scale from 0 to 100.
- **Time Spent.** Time taken to complete a given scenario per person will be recorded.

3.3 Covariates

- **Age.** Self-reported age in years.
- **Gender.** Self-reported gender.
- **Education level.** Highest level of education completed.
- **English fluency.** Self-rated fluency in speaking, reading, and writing English on a five-point Likert scale, collected separately for each.
- **Internet use.** Self-reported frequency of internet use for news, travel information, and health information, collected separately for each.
- **LLM use.** Self-reported frequency of LLM use for drafting, information seeking, health information, coding, editing, and image generation, collected separately for each.

4 Statistical Analysis Plan

4.1 Primary Outcomes

H1 and H2 are co-primary hypotheses. A Bonferroni correction is applied across both tests, setting the significance threshold at $\alpha = 0.025$ for each.

- **H1: Condition identification accuracy:** Condition identification accuracy will be tested using a mixed-effects logistic regression as the primary confirmatory analysis, with condition identification accuracy as the dependent variable, treatment arm as the key predictor, participant ID as a random intercept to account for the two responses contributed by each participant, and age, gender, education, English fluency (speaking, reading, writing), internet use (news, travel, health), and LLM use (drafting, information seeking, health information, coding, editing, image generation) as covariates. The coefficient on treatment arm will be evaluated against alpha = 0.025 per the Bonferroni correction described above.
- **H2: Urgency assessment accuracy:** Identical procedure to H1, with urgency accuracy as the dependent variable.

4.2 Secondary Outcomes

- **H3: Confidence:** Confidence ratings will be compared between arms using a linear mixed-effects model, with confidence rating as the dependent variable, treatment arm as the key predictor, participant ID as a random intercept to account for each participant's two responses, and age, gender, education, English fluency (speaking, reading, writing), internet use (news, travel, health), and LLM use (drafting, information seeking, health information, coding, editing, image generation) as covariates. The treatment coefficient will be evaluated using a two-sided test, with Cohen's d reported as the effect size.

- **H4: Time spent:** Time spent will be compared between arms using a linear mixed-effects model, with time spent as the dependent variable, treatment arm as the key predictor, participant ID as a random intercept to account for each participant's two responses, and age, gender, education, English fluency (speaking, reading, writing), internet use (news, travel, health), and LLM use (drafting, information seeking, health information, coding, editing, image generation) as covariates. The treatment coefficient will be evaluated using a one-sided test, with Cohen's d reported as the effect size.

For **H3** and **H4**, Cohen's d will be calculated by the fixed-effect coefficient for treatment arm by the total standard deviation of the outcome, defined as the square root of the sum of the random-intercept variance and residual variance from the corresponding mixed-effects model. This approach [9] accounts for both between-participant and within-participant variability given the clustered data structure.

Secondary hypotheses (H3, H4) will be evaluated at $\alpha = 0.05$, uncorrected for multiple comparisons. Results for these hypotheses will be interpreted as supportive rather than confirmatory findings.

4.3 Statistical Software

All analyses will be conducted in Python using the `statsmodels`, `scipy`, `pymer4`, and `thefuzz` packages.

4.4 Exploratory Analyses

- **LLM-Alone Benchmark Experiment:** Separate from the RCT, we will evaluate the new LLM interface by prompting it directly with each of the ten scenarios, without any human participant. Each scenario will be submitted up to 20 times per model (up to 200 responses per model total), with condition identification and disposition accuracy scored using the same gold-standard lists as the RCT. Results will be reported as proportions with bootstrapped 95% CIs (1,000 resamples).

- **LLM Experience Moderation:** Within the treatment arm, we will examine whether participant-level accuracy differs between more frequent LLM users and less frequent LLM users.

- **Internet Usage:** Participants who report high internet-based health information seeking will show better performance in diagnosis and triage relative to those who report low internet-based health information seeking.

- **Digital Literacy:** Participants with higher digital literacy will show better performance in diagnosis and triage relative to those with lower digital literacy.
- **Scenario-level Analysis**: We will analyze whether diagnosis and triage accuracy varies across different scenarios.

5 Risks, Harms, and Ethical Considerations

5.1 Potential Risks and Harms

This study poses minimal risk to participants. The scenarios are fictional vignettes and participants are not asked to make real healthcare decisions. Participants in the control arm may experience mild frustration if they find the web-search task difficult; those in the treatment arm may experience confusion if the LLM's questions feel unfamiliar.

5.2 Safeguards

Participants will receive clear instructions emphasizing that the task concerns a hypothetical scenario and that no real person's health is at stake. Participants will be reminded at the outset that they may withdraw at any time without penalty. Study instructions will make clear that no personally identifiable medical information should be shared.

5.3 Informed Consent

Informed consent will be obtained electronically from all participants before enrollment.

5.4 Ethics Review

The study protocol is submitted for review and approval by the Institutional Review Board (IRB) at the Lahore University of Management Sciences.

6 Conclusion

This pre-analysis plan describes a randomized controlled trial designed to test whether access to a structured, question-driven LLM interface along with conventional web sources can close the performance gap in medical diagnosis and triage for laypeople relative to those given access to standard LLM interface along with conventional web sources. By directly targeting incomplete information provision and poor output interpretation, this study is positioned to provide actionable evidence on a targeted design intervention. The findings will have direct implications for developers and regulators considering the deployment of LLMs as patient-facing medical assistance tools.

References

[1] Bean, A.M., Payne, R.E., Parsons, G., Kirk, H.R., Ciro, J., Mosquera-Gómez, R., Hincapié M., S., Ekanayaka, A.S., Tarassenko, L., Rocher, L., & Mahdi, A. (2026).

Reliability of LLMs as medical assistants for the general public: a randomized preregistered study. *Nature Medicine*. https://doi.org/10.1038/s41591-025-04074-y

[2] Saenger, J. A., Hunger, J., Boss, A., & Richter, J. (2024). Delayed diagnosis of a transient ischemic attack caused by ChatGPT. *Wiener Klinische Wochenschrift*, 136(7), 236–238. https://doi.org/10.1007/s00508-024-02329-1

[3] Ramaswamy, A., Tyagi, A., Hugo, H., Jiang, J., Jayaraman, P., Jangda, M., Te, A. E., Kaplan, S. A., Lampert, J., Freeman, R., Gavin, N., Tewari, A. K., Sakhuja, A., Naved, B., Charney, A. W., Omar, M., Gorin, M. A., Klang, E., & Nadkarni, G. N. (2026). ChatGPT Health performance in a structured test of triage recommendations. *Nature Medicine*, 32(5), 1671–1675. https://doi.org/10.1038/s41591-026-04297-7

[4] National Academies of Sciences, Engineering, and Medicine. (2015). *Improving Diagnosis in Health Care*. Washington, DC: The National Academies Press. https://doi.org/10.17226/21794

[5] Presiado, M., Montero, A., Lopes, L., & Hamel, L. (2024, August 15). *KFF Health Misinformation Tracking Poll: Artificial Intelligence and Health Information*. KFF. https://www.kff.org/public-opinion/kff-health-misinformation-tracking-poll-artificial-intelligence-and-health-information/

[6] Shahsavar, Y., & Choudhury, A. (2023). User Intentions to Use ChatGPT for Self-Diagnosis and Health-Related Purposes: Cross-sectional Survey Study. *JMIR Human Factors*, 10, e47564. https://doi.org/10.2196/47564

[7] Bickley, L.S., & Szilagyi, P.G. (2017). *Bates' Guide to Physical Examination and History Taking* (12th ed.). Wolters Kluwer.

[8] Novak Jones, D. (2026, May 12). OpenAI faces lawsuit in California court claiming chatbot gave advice that led to fatal overdose. *Reuters*. https://www.reuters.com/legal/litigation/openai-faces-lawsuit-california-court-claiming-chatbot-gave-advice-that-led-2026-05-12/

[9] Feingold, A. (2009). Effect sizes for growth-modeling analysis for controlled clinical trials in the same metric as for classical analysis. *Psychological Methods*, 14(1), 43–53. https://doi.org/10.1037/a0014699