Mitigating Automation Bias in Physician-LLM Diagnostic Reasoning Using Behavioral Nudges

Ihsan Ayyub Qazi, Ayesha Ali, Muhammad Asad Ullah Khawaja,
Muhammad Junaid Akhtar Ali Zafar Sheikh, Muhammad Hamad Alizai

Study Protocol and Statistical Analysis Plan

Document Date: December 26, 2025

1 Introduction

Automation bias presents substantial risks in clinical environments, particularly as artificial intelligence (AI) tools become more integrated into healthcare workflows. This cognitive phenomenon describes the tendency of individuals to favor suggestions from automated decision-making systems, even when those suggestions are incorrect. While Large Language Models (LLMs), such as ChatGPT-5.1, hold promise for reducing diagnostic errors by augmenting clinical reasoning, concerns arise about their potential to amplify cognitive biases. Of particular concern is the propensity of LLMs to *hallucinate*, generating plausible but false information, which, when combined with automation bias, can lead clinicians to accept erroneous recommendations without adequate scrutiny.

The emergence of automation bias in medical contexts reflects a complex interplay of environmental and psychological factors. Time constraints in high-volume clinical settings create pressure to accept AI-generated recommendations without adequate scrutiny. Financial incentives that prioritize efficiency over thoroughness may further discourage the critical evaluation necessary for sound clinical judgment. Cognitive fatigue during extended shifts diminishes physicians' capacity for sustained analytical thinking. These contextual pressures interact with psychological mechanisms including diffusion of responsibility, overconfidence in technological solutions, and cognitive offloading, collectively creating conditions where uncritical acceptance of AI-generated recommendations becomes more likely.

This randomized controlled trial evaluates the effectiveness of a behavioral nudge intervention designed to mitigate automation bias among medical doctors utilizing LLM-generated diagnostic recommendations. The primary objective is to determine whether this intervention improves diagnostic reasoning performance scores when physicians evaluate clinical vignettes that include deliberately flawed LLM recommendations. Secondary objectives include assessing whether physician experience level, gender, and prior LLM experience moderate the intervention's effectiveness.

Participants will be randomly assigned to control or treatment arms. Both groups will evaluate six clinical vignettes accompanied by LLM-generated recommendations, three of which contain deliberately introduced errors that are subtle yet clinically significant. LLM consultation will be discretionary, physicians will choose whether to view ChatGPT recommendations for each case.

The control arm will receive these recommendations without additional context. The treatment arm will receive a behavioral nudge intervention leveraging complementary cognitive mechanisms: (1) an anchoring cue displaying ChatGPT's average diagnostic reasoning accuracy on standard medical datasets before evaluation begins, establishing realistic performance expectations, and (2) a selective attention cue enabling deeper System 2 engagement through color-coded confidence signals when participants view ChatGPT recommendations. These confidence signals are generated through ensemble assessment: three independent state-of-the-art LLMs (Claude Sonnet 4.5, Gemini 2.5 Pro Thinking, and GPT-5.1) each provide confidence ratings for the recommendation, and the mean confidence determines the signal color. Multi-model aggregation mitigates the systematic overconfidence of single-model confidence reports, yielding more robust uncertainty estimates. Red signals indicate mean confidence below ChatGPT's baseline accuracy on medical benchmarks, prompting heightened critical scrutiny. Orange signals indicate above-baseline but less than 100% confidence, deliberately using orange rather than yellow to prevent complacency. Green signals indicate high ensemble confidence (= 100%), though standard AI safety warnings apply (e.g., "AI may make mistakes. Always double-check.").

Clinical vignettes will be randomly sequenced to prevent pattern detection. By comparing diagnostic accuracy scores between arms, we will quantify the degree of automation bias in the control group and assess whether the behavioral nudge intervention reduces uncritical acceptance of erroneous AI-generated recommendations in the treatment group. This design enables

us to isolate both the magnitude of automation bias and the causal impact of transparency-based interventions on physician decision-making.

All participants will complete a comprehensive AI-literacy training program covering LLM capabilities, limitations, prompt engineering techniques, and output evaluation strategies before beginning the trial. Responses will be evaluated by blinded reviewers using an expert-developed assessment rubric specifically designed to detect uncritical acceptance of erroneous information. This naturalistic approach will yield insights directly applicable to real clinical workflows, where increasing cognitive demands can progressively impact diagnostic decision quality, while providing evidence on scalable interventions to preserve clinical judgment in AI-augmented diagnostic settings.

1.1 Research Question

Does a behavioral nudge intervention combining baseline ChatGPT accuracy information and case-specific color-coded confidence signals reduce automation bias and improve diagnostic reasoning accuracy among physicians with discretionary access to LLM recommendations?

1.2 Hypotheses

- **Primary Hypothesis (H1):** Physicians receiving the behavioral nudge intervention (baseline accuracy information and case-specific color-coded confidence signals) will demonstrate significantly higher diagnostic reasoning scores when evaluating clinical vignettes with deliberately flawed LLM recommendations compared to physicians receiving the same LLM recommendations without the nudge.
- **Secondary Hypotheses:**
  - **H2:** Physicians with less clinical experience will exhibit greater baseline automation bias and show larger treatment effects from the nudge intervention compared to more experienced clinicians.
  - **H3:** The intervention will demonstrate greater effectiveness for red-signal (below-baseline) versus orange-signal (above-baseline) recommendations.

2 Methods

2.1 Study Design

This study will be a single-blind, randomized controlled trial with two parallel arms. Participants will be randomly assigned in a 1:1 ratio to the intervention arm or to the control arm. To control for variability in user prompting skills, participants will not interact directly with a live LLM chat interface. Instead, they will use a custom-built web platform that displays the clinical vignette and provides access to pre-generated LLM recommendations.

- **Control Arm:** Participants will evaluate six clinical vignettes (three with deliberately flawed diagnostic information and three with correct information). Through the study platform, they will be able to view "LLM Diagnostic Recommendations" generated by ChatGPT. These recommendations will be presented in a standard, neutral text format without additional context or nudge.

- **Intervention Arm:** Participants will evaluate the same six clinical vignettes using the same web platform, but with a behavioral nudge embedded directly within the LLM suggestion interface. Upon expanding the LLM recommendation panel for any vignette, participants will be presented with two synchronized cognitive cues:

1. **Anchoring Cue:** Displayed at the top of the panel, this visual cue presents ChatGPT's baseline diagnostic accuracy on standard medical datasets (e.g., Medbullets). This serves to explicitly anchor expectations to the model's realistic performance level, preventing both excessive skepticism and uncritical acceptance, before the participant reads the specific recommendation.
2. **Selective Attention Cue:** Below the baseline accuracy information, each case-specific LLM recommendation will be displayed alongside a color-coded signal reflecting the ensemble confidence for that particular recommendation. This score is derived from a consensus evaluation by three independent state-of-the-art LLMs (e.g., Claude 4.5 Sonnet, Gemini 2.5 Pro Thinking, and GPT-5.1) to mitigate single-model miscalibration:
   - **Red Signal:** Displayed when the ensemble's mean confidence falls below ChatGPT's established baseline accuracy, signaling a high-uncertainty case that requires deeper critical evaluation and indicating "*The LLM is less sure about this case than it usually is*".
   - **Orange Signal:** Displayed when the ensemble confidence exceeds the established baseline average but remains below 100%. The choice of orange (vs. yellow) is a deliberate nudge to prevent automation bias and maintain clinical scrutiny.
   - **Green Signal**: Displayed when the ensemble reaches a consensus of 100% confidence. Standard caution will apply (e.g., "AI may make mistakes. Always double-check.") to guard against over-reliance.

Randomization: Participants will be randomly assigned to either the intervention or control arm in a 1:1 ratio using a computer-generated random number sequence.

Blinding: Outcome assessors (those scoring the diagnostic reasoning performance) will be blinded to the participants' group assignments. Participants will be blinded to the study's specific hypothesis regarding automation bias but will be informed they are evaluating LLM diagnostic assistance tools.

Measuring change in automation bias: The primary measure of the change in automation bias will be the mean difference in diagnostic reasoning accuracy scores between the treatment and control arms. A positive difference favoring the treatment arm will indicate the reduction in automation bias attributable to the nudge.

2.1.1 Rationale for Ensemble Confidence Assessment

Our decision to utilize ensemble-based confidence signals rather than single-model self-assessment is grounded in extensive empirical evidence of LLM calibration failures. LLMs frequently exhibit overconfidence, expressing high certainty even when producing factually incorrect outputs [1]. Recent studies confirm these challenges persist in state-of-the-art models; standard prompting often yields poorly calibrated estimates that self-reflection mechanisms alone cannot rectify, particularly in high-stakes medical and scientific domains [2, 3].

To address this, we estimate confidence via an ensemble of three distinct LLMs, and use the mean score as a selective-attention cue. This shifts confidence estimation away from unreliable internal self-reports toward practical uncertainty quantification. By aggregating judgments from multiple models, we reduce sensitivity to any single model's miscalibration. When independently queried models disagree, the case may be genuinely ambiguous, warranting heightened human review. In the clinical context, this approach offers two critical advantages. First, the mean confidence score serves as a "soft cue" to modulate physician attention. Second, the architecture remains scalable; modern API frameworks allow for parallel querying with minimal latency overhead [4]. Ultimately, we treat the ensemble mean not as a definitive probability, but as a robust heuristic to safeguard against overstated confidence typical of single-model architectures.

2.2 Participants

2.2.1 Inclusion Criteria

- Full or Provisionally Registered Medical Practitioners with the Pakistan Medical and Dental Council (PMDC).
- Completed Bachelor of Medicine, Bachelor of Surgery (MBBS) Exam. The equivalent degree of MBBS in the US and Canada is the Doctor of Medicine (MD).
- Participants must have completed a structured training program on the use of ChatGPT-4o (or a comparable large language model), totaling at least 10 hours of instruction. The program must include hands-on practice related to the LLM's key aspects, specifically prompt engineering and content evaluation.

2.2.2 Exclusion Criteria

- Any other Registered Medical Practitioners (Full or Provisional) with PMDC (e.g., professionals with Bachelor of Dental Surgery or BDS).

2.3 Sample Size and Power Analysis

Target Sample Size: 50 participants (25 per arm).

Power Calculation: Based on pilot data and similar studies, we assume:

- An effect size (difference in mean diagnostic reasoning scores) of 8 percentage points between groups.
- A standard deviation of 16.2%.
- An intraclass correlation coefficient (ICC) between 0.05 and 0.15 (to account for clustering of cases within participants).
- An average of 4 completed cases per participant.

With these assumptions, a sample size of 50 provides $\geq 80\%$ power to detect a significant difference in the primary outcome, diagnostic reasoning score, between the two groups using a linear mixed-effects model ($\alpha = 0.05$, two-sided).

2.4 Clinical Vignettes & LLM Recommendations

Participants will evaluate six clinical vignettes specifically designed to measure automation bias. These vignettes will:

- Be based on modified real cases.
- Represent a range of diagnostic complexity across common medical specialties.
- Avoid overly simplistic or extremely rare conditions.
- Follow a standardized format, including:
  - Chief complaint and history of present illness.
  - Relevant past medical, social, and family history.
  - Physical examination findings.
  - Initial laboratory results.
- Include three clinical vignettes whose LLM recommendations contain deliberately introduced clinical reasoning errors and three with accurate recommendations, presented in randomized order. A panel of three licensed physicians will design these errors to be subtle yet clinically significant and detectable by trained clinicians.

- For every vignette, the LLM's recommendations will be pre-generated and fixed. This ensures all participants see identical recommendations for each vignette, the same error in flawed cases and the same correct reasoning in accurate cases, eliminating variance from real-time model variability.

Vignette Selection and Validation:

- A pool of 10 potential vignettes will be created and reviewed by an independent panel of three experienced medical doctors for:
  - Clinical accuracy and realism.
  - Appropriateness of difficulty level.
  - Clarity and completeness of information.
  - In flawed vignettes: subtlety and clinical significance of the introduced errors (errors should be detectable by competent physicians but not immediately obvious).
- Six vignettes will be selected based on the panel's ratings and feedback, aiming for a balanced representation of medical specialties and diagnostic challenges.
- For intervention arm participants, LLM output will be pre-generated and standardized for each vignette to ensure consistency in the presentation of correct or flawed information.

Vignette Completion:

- Each participant will complete all six vignettes in a single, proctored, approximately 75-minute session.
- The order of vignettes will be randomized for each participant to control for order effects and prevent pattern recognition.

2.5 Data Collection

2.5.1 Session Protocol

- Participants will complete the study under proctored conditions (in person or via secure video conferencing).
- Participants will complete a brief baseline questionnaire collecting demographic information (gender, specialty, years of experience), technology proficiency, and prior LLM experience.
- Participants will receive instructions specific to their assigned group (intervention or control).
- Participants will evaluate clinical vignettes one at a time.
- For each vignette, participants will provide:
  1. Their top three differential diagnoses.
  2. Supporting and opposing factors for each differential diagnosis (clinical findings, test results, etc.).
  3. Their most likely final diagnosis.
  4. Recommended next steps in the diagnostic workup (e.g., further tests, consultations).
- Participants will record all responses using a standardized electronic form.
- Time spent per vignette will be automatically tracked.
- For each vignette, a binary indicator will record whether the participant accessed the LLM recommendations (1 = viewed; 0 = not viewed).

3 Outcome Measures

3.1 Primary Outcome

Diagnostic Reasoning Accuracy Score: This will be a composite score (expressed as a percentage) based on a structured rubric that evaluates the following components for each clinical vignette:

1. **Quality of Differential Diagnoses:** (0-3 points)
    - 0: No reasonable differential diagnoses listed.
    - 1: One reasonable differential diagnosis listed.
    - 2: Two reasonable differential diagnoses listed.
    - 3: Three or more reasonable differential diagnoses listed, including the correct diagnosis.
2. **Supporting Findings:** (0-1 points per diagnosis)
    - 0: No relevant supporting findings listed.
    - 1: Some relevant supporting findings listed, but incomplete.
    - 2: Comprehensive list of relevant supporting findings.
3. **Opposing Findings:** (0-1 points per diagnosis)
    - 0: No relevant opposing findings listed.
    - 1: Some relevant opposing findings listed, but incomplete.
    - 2: Comprehensive list of relevant opposing findings.
4. **Top Choice Diagnosis Accuracy:** (0-18 points)
    - 0: Incorrect final diagnosis.
    - 9: Partially correct final diagnosis (e.g., correct category but wrong specific diagnosis).
    - 18: Correct final diagnosis.
5. **Appropriateness of Next Steps:** (0-1 points per step)
    - 0: Inappropriate or unnecessary next steps.
    - 0.5: Partially appropriate next steps (e.g., some necessary tests missing).
    - 1: Appropriate and comprehensive next steps.

The total score for each vignette will be the sum of these components.

3.2 Secondary Outcomes

- **Top Choice Diagnosis Accuracy**: We will use the *Final Diagnosis Accuracy* as the secondary outcome variable.

4 Data Analysis Plan

4.1 Descriptive Statistics

- Baseline characteristics will be summarized using means and standard deviations for continuous variables (e.g., years of experience) and frequencies and percentages for categorical variables (e.g., gender, specialty).
- Diagnostic reasoning scores and top choice diagnosis accuracy will be summarized using descriptive statistics by group
- LLM consultation rate: For each vignette, a binary indicator will be recorded (1 = participant viewed LLM recommendations; 0 = did not view). The measure is the mean proportion of vignettes (range 0-1) per participant in which recommendations were accessed.

4.2 Primary Outcome Analysis

- A linear mixed-effects model will be used to compare the mean diagnostic reasoning accuracy score between the intervention and control groups.
- **Model Structure:**
  - **Dependent Variable:** Diagnostic reasoning performance score (percentage).
  - **Fixed Effect:** Treatment group (Intervention vs. Control).
  - **Random Effects:**
    - Participant ID (to account for within-participant correlation across cases).
    - Vignette ID (to account for potential variability in case difficulty).
  - **Covariates:**
    - Years of practice post MBBS, gender, and prior experience with LLMs.
- **Primary Comparison:** The difference in mean diagnostic reasoning scores between the two groups, adjusted for random effects and covariates.
- **Significance Threshold:** A two-sided p-value ≤ 0.05 will be considered statistically significant.

4.3 Secondary Outcome Analyses

- **Top Choice Diagnosis Accuracy:**
  - A linear mixed-effects model will be used, with the same random effects structure as the primary analysis, and the dependent variable being the top choice diagnosis accuracy.

4.4 Subgroup Analyses

Subgroup analyses will be performed to explore whether the effect of the intervention differs among physicians based on physician experience level (e.g., years since MBBS/MD), gender, and prior LLM experience.

4.5 Sensitivity Analyses

The following sensitivity analyses will be conducted to assess the robustness of the findings:

- **Incomplete Cases:** The primary outcome analysis will be repeated including incomplete cases (if any) to evaluate the potential impact of missing data (e.g., using multiple imputation).
- **Scorer Agreement:** Inter-rater reliability for the diagnostic reasoning scores across at least two raters will be assessed using Cohen's kappa or Fleiss' kappa depending on the final number of raters.
- **Internal Consistency:** The internal consistency of the diagnostic reasoning score across the six vignettes will be assessed using Cronbach's alpha.
- **Distribution Analysis:** Assessment of normality assumptions through:
  - Visual inspection of distributions (e.g., Q-Q plots, histograms)
  - Formal tests of normality (e.g., Shapiro-Wilk test)

4.6 Exploratory Analyses

- **Differential Impact of Confidence Signals:** We will compare diagnostic scores for vignettes receiving Red Signal (flawed) versus Orange Signal versus Green Signal by group.

- **Chronometric Analysis of System 2 Engagement:** Using timestamps, we will compare time spent on Red Signal versus Orange Signal vignettes within the Treatment group. Longer durations on Red Signal vignettes would indicate deeper System 2 processing.
- **Physician-Case Specialty Match:** Exploratory analysis will test whether automation bias is lower when a vignette matches the physician's specialty, independent of AI accuracy or signal type.

4.7 Missing Data

- The primary analysis will be based on complete cases only.
- The pattern and extent of any missing data will be described.
- If substantial missing data occurs ($\geq 5\%$ for any key variable), multiple imputation will be considered, and results will be compared to complete case analysis.

4.8 Statistical Software

All statistical analyses will be performed using R (version 4.3.2 or later) with appropriate packages (e.g., 'lme4', 'ordinal', 'mice') or equivalent softwares in Python or Stata.

5 Risks, Harms & Ethical Considerations

5.1 Potential Risks and Harms

This study poses minimal risk to participants. Some participants may experience mild discomfort or frustration, particularly those in the control group who may find clinical case studies challenging without LLM assistance.

5.2 Safeguards for Addressing Risks or Harms to Subjects

To minimize potential discomfort, participants will receive comprehensive instructions and ongoing support throughout the study. Participants will be informed that individual performance is not being evaluated, and that data will be analyzed for aggregate trends only. All participants will be advised of their right to withdraw from the study at any time without penalty or explanation. Research staff will monitor for signs of distress and provide appropriate support as needed.

5.3 Informed Consent

Informed consent will be obtained from all participants before enrollment in the study.

5.4 Participant Compensation

Participants will receive a complimentary one-month subscription to ChatGPT Plus (or equivalent) after completing the study. Additionally, each participant will receive a certificate of participation to formally acknowledge their valuable time and effort in contributing to this research.

5.5 Ethics Review

This study will be submitted for review and approval by the Institutional Review Board (IRB) of the participating institution(s) before any participants are enrolled.

6 Amendments

Any deviations from this pre-analysis plan will be documented, justified, and reported in the final study report. Substantial changes to the protocol will be submitted for IRB review and approval prior to implementation.

7 Conclusion

This pre-analysis plan outlines a study to evaluate the impact of a behavioral nudge to mitigate automation bias in LLM-assisted diagnostic reasoning. By adhering to this plan, we aim to minimize bias, enhance transparency, and produce high-quality evidence that can inform the integration of LLMs into clinical practice. The findings will contribute to a better understanding of the potential benefits and limitations of using LLMs as diagnostic aids, ultimately aiming to improve patient care and reduce diagnostic errors.

References

[1] M. Xiong, Z. Hu, X. Lu, Y. Li, J. Fu, J. He, and B. Hooi, "Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs," in *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2306.13063.

[2] Z. Ji, T. Yu, Y. Xu, N. Lee, E. Ishii, and P. Fung, "Towards Mitigating Hallucination in Large Language Models via Self-Reflection," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 1827–1843, 2023.

[3] X. Zhang, B. Peng, Y. Tian, J. Zhou, L. Jin, L. Song, H. Mi, and H. M. Meng, "Self-Alignment for Factuality: Mitigating Hallucinations in LLMs via Self-Evaluation," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*, pp. 1946–1965, 2024. doi:10.18653/v1/2024.acl-long.107.

[4] M. Liu, A. Zeng, B. Wang, P. Zhang, J. Tang, and Y. Dong, "APAR: LLMs Can Do Auto-Parallel Auto-Regressive Decoding," arXiv:2401.06761, 2024.