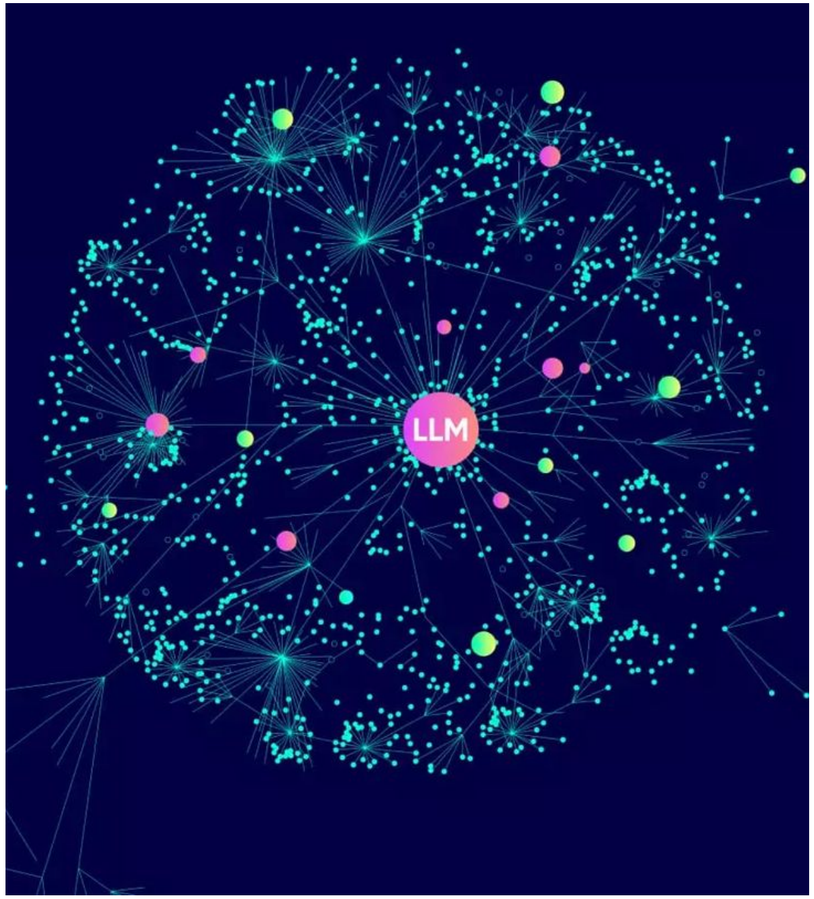
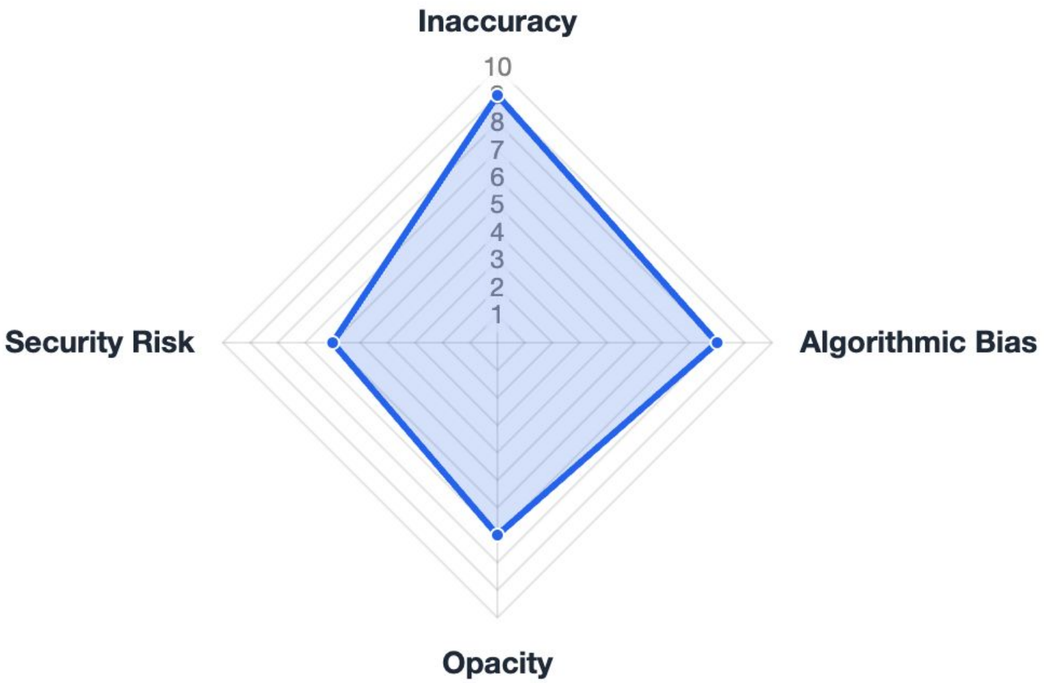
A FRAMEWORK FOR TRUSTWORTHY CLINICAL AI

Presented By Noah Schumacher



TRANSFORMATIVE POTENTIAL OF LLMS

- LLMs can enhance clinical decision support, automate administrative tasks, and improve patient communication.
- Demonstrated expert-level performance on medical exams (Med-PaLM).
- Potential to reduce clinician burnout and democratize access to knowledge.



THE RESEARCH PROBLEM & SIGNIFICANCE

The "Trust Gap"

- Significant barriers hinder safe and ethical adoption in routine practice.
- Fueled by unresolved technical and ethical challenges.
- This research aims to bridge the gap between abstract principles and practical tools.

CRITICAL BARRIERS: THE FOUR KEY RISKS

1. INACCURACY & HALLUCINATION

Models generate confident but false information, posing direct risks to patient safety.

2. ALGORITHMIC BIAS & INEQUITY

Models trained on biased data can perpetuate and amplify health disparities.

3. THE "BLACK BOX" PROBLEM

Lack of transparency erodes clinician trust and prevents critical evaluation of outputs.

4. DATA PRIVACY & SECURITY

Use of sensitive patient data requires strict adherence to regulations like HIPAA and GDPR.

PRIMARY RESEARCH QUESTION

How can a Human-in-the-Loop (HITL) dashboard, designed using a mixed-methods approach, enable healthcare professionals to effectively evaluate, monitor, and mitigate ethical and safety risks associated with clinical decision support LLMs in real-time?

PROJECT AIMS & OBJECTIVES

AIMS

1. Identify requirements for an LLM oversight tool.
2. Design and develop a high-fidelity HITL dashboard prototype.
3. Empirically evaluate the prototype's effectiveness and usability.

KEY OBJECTIVES

1. Analyze AI evaluation guidelines (TRIPOD-LLM, etc.).
2. Design specific dashboard modules (Bias, Explainability, Safety).
3. Develop an interactive prototype.
4. Create simulated clinical vignettes for testing.
5. Conduct a mixed-methods usability study with ≥10 clinicians.
6. Analyze data to evaluate risk detection and user trust.

INSIGHTS FROM KEY LITERATURE

THE "BLACK BOX" DILEMMA

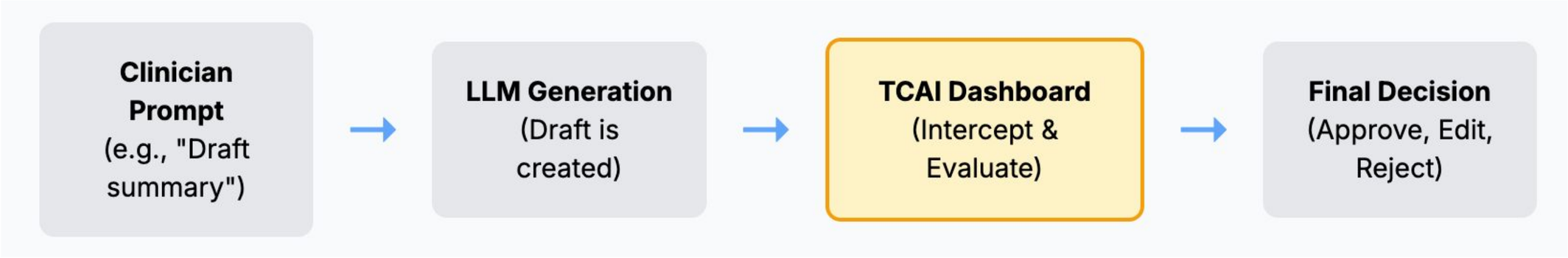
Clinicians cannot trust what they cannot understand. Transparency is essential. Retrieval-Augmented Generation (RAG) is a promising technical solution to ground outputs in verifiable source data, enhancing explainability.

THE EQUITY IMPERATIVE

Bias is a "wicked problem," not a simple bug. A one-time "debiasing" fix is insufficient. The only viable path is continuous, real-time monitoring and auditing of model performance within specific contexts.

PROPOSED SOLUTION: THE TCAI DASHBOARD

An interactive "control panel" that sits between the clinician and the LLM, transforming the user from a passive consumer to an active, informed evaluator.



DASHBOARD CORE FEATURES

Safety & Confidence Score

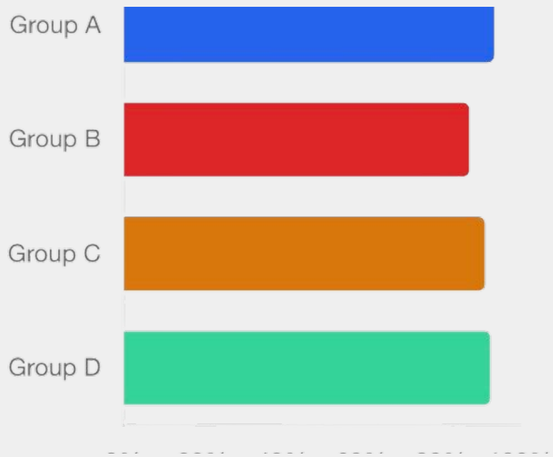
Flags hallucinations & low-confidence outputs using a traffic-light system.

Explainability & Grounding

Links every claim back to the source data in the patient's record for full transparency.

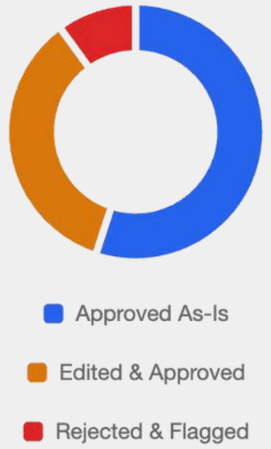
Bias & Equity Audit

Visualizes historical model performance across demographic groups.



Clinician Interaction Layer

Provides simple tools to Accept, Edit, or Reject & Flag the AI's output.



RESEARCH METHODOLOGY

PHASE 01

Requirements & Synthesis

Systematic analysis of AI evaluation frameworks (e.g., TRIPOD-LLM, CONSORT-AI) to synthesize a comprehensive set of functional and ethical requirements for the dashboard.

Key Deliverable:
Requirements Specification Document

PHASE 02

Artefact Development

Agile development of a high-fidelity, interactive prototype of the TCAI Dashboard, translating requirements into a functional user interface with core modules.

Key Deliverable:
Interactive Prototype v1.0

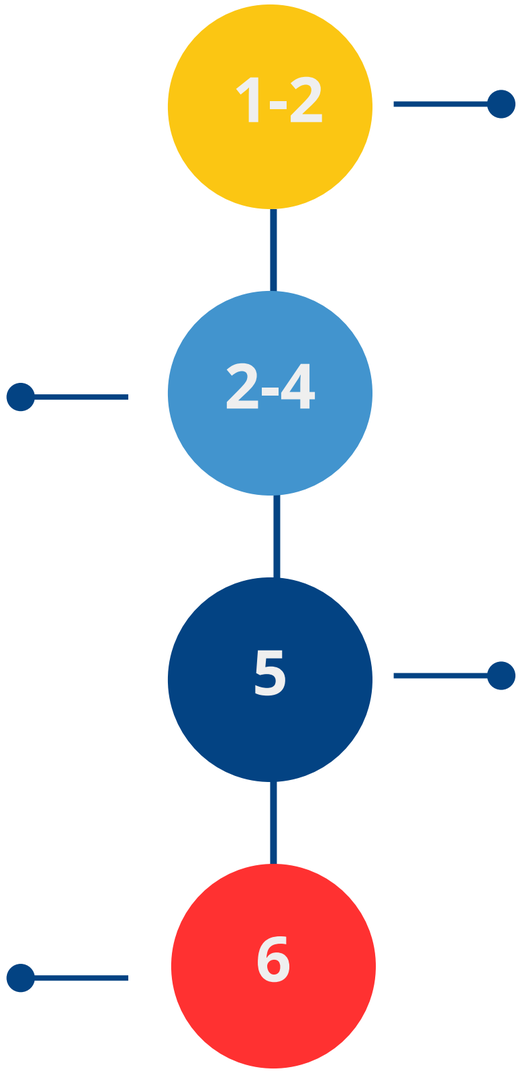
PHASE 03

Evaluation & Analysis

Mixed-methods usability study with clinicians using a think-aloud protocol and simulated cases to evaluate the prototype's effectiveness, usability, and impact on trust.

Key Deliverable:
Analyzed Data & Findings

PROJECT TIMELINE (6 MONTHS)



Foundation & Requirements

In-depth literature review and synthesis of evaluation frameworks. Milestone: Requirements Specification Document.

Development & Prototyping

Wireframing, high-fidelity prototype development, and creation of test vignettes. Milestone: Interactive Prototype v1.0.

Empirical Evaluation

Ethical approval, participant recruitment, and usability testing sessions. Milestone: Data Collection Complete.

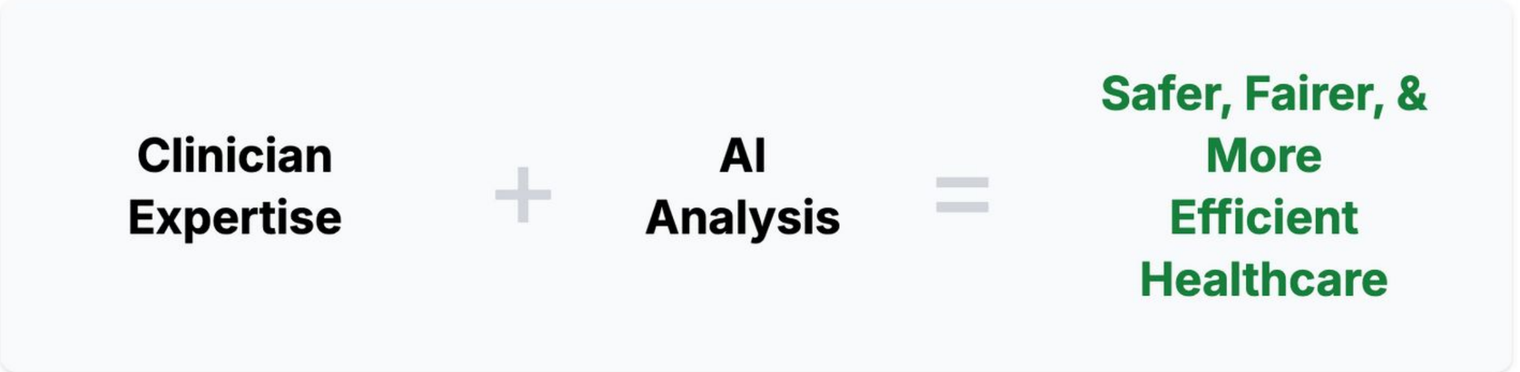
Analysis & Write-up

Data analysis, dissertation writing, and presentation preparation. Milestone: Final Submission.

ETHICAL CONSIDERATIONS

- **Participant Data Privacy:** No real Patient Health Information (PHI) will be used. All research data will be fully anonymized and stored securely.
- **Informed Consent:** Clinician participants will undergo a rigorous informed consent process, ensuring they understand the study and their right to withdraw.
- **Dual-Use Dilemma:** The research is explicitly framed around AI as an *assistive* tool to augment, not replace, human judgment and accountability.
- **Burden on Participants:** The study protocol is designed to be concise (~60-75 mins), and participants will be offered an honorarium for their time.

CONCLUSION: AUGMENT, DON'T AUTOMATE



This research aims to build the tools that make trust a tangible property of the clinical workflow, ensuring a human expert always remains at the heart of patient care.

REFERENCES

Bender, E.M., Gebru, T., McMillan-Major, A. and Shmitchell, S., 2021. On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT).

Khan, S., Sharma, P. and Kumar, N., 2023. Language models in global health. Nature Medicine.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I.D. and Gebru, T., 2022. Model cards for model reporting. Communications of the ACM, 64(12), pp.56-65.

Nori, H., King, N., Carignan, D. and Horvitz, E., 2023. Capabilities of GPT-4 in healthcare. MedRxiv. [Preprint]

Obermeyer, Z., Powers, B., Vogeli, C. and Mullainathan, S., 2019. Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), pp.447-453.

QUESTIONS