Human vs. AI: Performance of AI in Literature Review Screening

NYU Grossman School of Medicine

Amelia Shunk, MMCi[1], Natalie Henning, MPH[1], Gregory Laynor, PhD, MLS[2], Elizabeth R. Stevens, PhD, MPH[1]

1 NYU Grossman School of Medicine, New York, NY, 2 NYU Health Sciences Library, New York, NY

Department of Population Health
Section of Digitally Enabled Care
HiBRID Lab

INTRODUCTION

Literature reviews provide methods to identify and summarize existing research on a given topic. However, literature reviews are time-intensive, requiring labor and consensus from multiple research team members to complete the initial screening process.[1] Depending on the topic and search strategy, the screening phase could include thousands of titles. One such scoping review protocol seeking to identify research where researchers were the study subjects returns nearly **20,000** titles to be screened.[2] Artificial intelligence (AI) provides great promise to reduce laborious tasks such as screening and may be as, or more, effective than human reviewers.[3] This research aims to answer the question: **Can a general purpose AI tool perform sufficiently well in title and abstract screening compared to human reviewers to be used in a scoping review?**

METHODS

This study compared **human** and **AI (ChatGPT 4o)** performance in **title and abstract screening** for a scoping review. A total of **19,140 citations** were retrieved from **PubMed, CINAHL, and PsycInfo** in July 2024. A **10% random sample (1,912 citations)** was screened to identify studies in which **researchers were the study subjects**.

Human screening was conducted as follows:
- Each title and abstract was screened **in duplicate** by trained reviewers.
- Reviewers selected **"Yes," "No," or "Maybe"** based on predefined inclusion criteria.
- **Discrepancies** were resolved by **consensus** with a third reviewer.

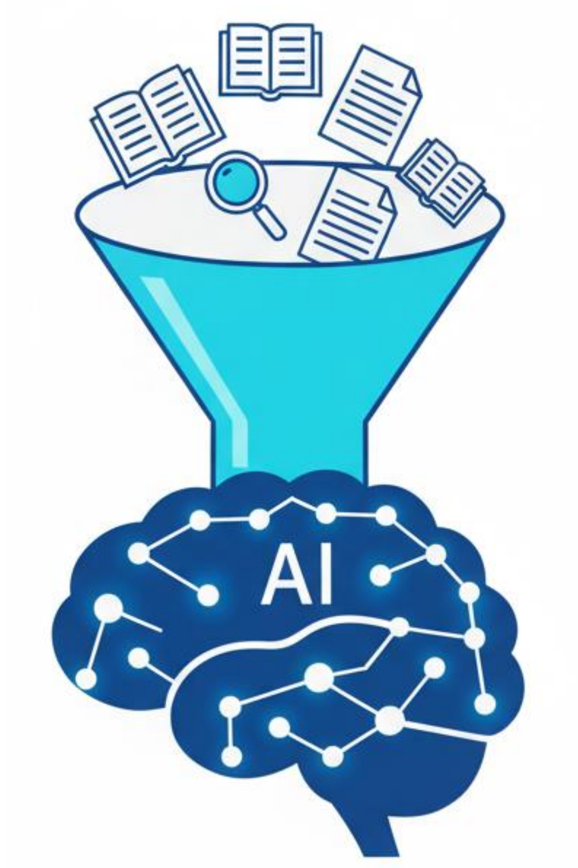
AI screening was conducted using **ChatGPT 4o**:
- A **screening prompt** was developed using the same inclusion criteria as the human reviewers.
- The prompt was **refined on a random subsample of 100 citations**, which were reviewed manually to assess performance.
- The full sample of **1,912 titles and abstracts** was then screened using **Chainforge**, producing outputs that mirrored the human review format.[4]

Following screening:
- A **manual review of discrepancies** was performed to identify any references that may have been incorrectly excluded by human reviewers.
- Articles included after **full-text review** were designated as the **"gold standard."**
- **Sensitivity and specificity** were calculated for both human and AI screening results relative to this gold standard dataset.

RESULTS

Gold Standard Dataset Creation:
- Among publications screened, there were **93 instances** where humans marked "No" but AI marked "Yes" or "Maybe."
- Upon **consensus reevaluation**, **42 of these 93** were reclassified as "Yes" or "Maybe."
- The final **gold standard dataset** included **523 publications** designated for full-text review.

Performance Metrics:
- **Human Screening:** Sensitivity **89.5%**, Specificity **77.1%**
- **AI Screening:** Sensitivity **76.0%**, Specificity **92.4%**



ChainForge

CONCLUSIONS

In the abstract and title screening phase of a scoping review, AI demonstrated **lower sensitivity** but **much higher specificity** compared to human reviewers.

Implications for Use:
- AI without specialized model training may serve as a **viable tool to reduce labor burden** during title and abstract screening when **comprehensiveness is not required** (e.g., scoping reviews).
- However, it is **not yet suitable** for **systematic reviews** or **medical decision-making**, where complete capture of evidence is essential.

Potential for Hybrid Approaches:
AI could **augment human reviewers** by identifying **inappropriately excluded or included studies**, improving both efficiency and rigor in the **downstream full-text review**.

Next Steps:
Future work will compare **AI and human performance in full-text screening** and evaluate **AI-assisted data extraction**.

REFERENCES

1. Nussbaumer-Streit B, Ellen M, Klerings I, et al. Resource use during systematic review production varies widely: a scoping review. *Journal of Clinical Epidemiology*. 2021/11/01/ 2021;139:287-296. doi:10.1016/j.jclinepi.2021.05.019
2. Laynor G, Stevens ER. Identifying meta-research with researchers as study subjects: Protocol for a scoping review. *PLoS One*. 2024;19(5):e0303905. doi:10.1371/journal.pone.0303905
3. Bajwa J, Munir U, Nori A, Williams B. Artificial intelligence in healthcare: transforming the practice of medicine. *Future Healthc J*. Jul 2021;8(2):e188-e194. doi:10.7861/fhj.2021-0095
4. Arawjo I, Swoopes C, Vaithilingam P, Wattenberg M, Glassman E. ChainForge: A Visual Toolkit for Prompt Engineering and LLM Hypothesis Testing. 2023:arXiv:2309.09128. doi:10.48550/arXiv.2309.09128