

Leveraging Artificial Intelligence Models to Streamline Systematic Reviews in Hand Surgery

Russell Day HS³, Carter Dunaway BS², Zachary Chen BS², Mithil Gudi BS², Charles S. Day M.D, MBA^{1,2}
Henry Ford Health System¹, Wayne State University School of Medicine², University of Michigan School of LSA³

Introduction & Objectives

- Systematic reviews are essential for evidence-based medicine, providing comprehensive and unbiased syntheses of research to inform clinical decision-making.
- Exponential growth of medical knowledge presents a significant challenge. Medical knowledge now doubles every 73 days, leaving systematic reviews outdated by the time they are published.
- The most time-intensive step of systematic reviews is the literature retrieval and initial database development phase, often consuming twice as much time as other steps. Manually sorting vast numbers of papers is labor-intensive, error-prone, and inconsistent.
- Newer AI models have demonstrated remarkable capabilities in natural language processing, understanding context, and performing complex analytical tasks offering promising solution to expediting the systematic review process.
- Hypothesis:** In a much more efficient manner, AI – large language models – are highly sensitive at over 90%.

Materials & Methods

Systematic Review Golden Standard: Human Reviewers

- 2 authors queried all of the articles included from a specific search criteria looking for comparative value studies in hand and wrist care across three databases.
- 121 articles were reviewed by the 2 authors, experienced in hand and wrist surgery research, who would accept or reject the paper separately following the PRISMA guidelines and internal inclusion/exclusion criteria.
 - Inclusion: Articles that compare two or more patient cohorts with a value calculation based on collected clinical data.
 - Exclusion: Non-english, model based data, systematic review/review articles, articles only looking at cost or quality individually, and protocols.

Categories for Evaluation

- Accept, Reject, or Maybe the paper based on the criteria for the paper
- Any “Maybe” responses were treated as “Accepted” due to its limited number in the overall dataset.

Evaluation

- The performance of each model was compared to the human-sorted reference group, assessing accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and time.

AI Models

- Four common commercially available AI models were chosen for this test
- Gemini 1.5 Pro, Llama-3.1-70B, Claude 3.5 Sonnet, and ChatGPT 4o

Consistency

- Each model was provided the title and abstract for each paper along with the same inclusion/exclusion criteria given to the authors who performed the manual sorting
- To ensure consistent answers, each model’s “temperature” (a measure of randomness) was set to 0 guaranteeing identical responses for repeated inputs

Figures

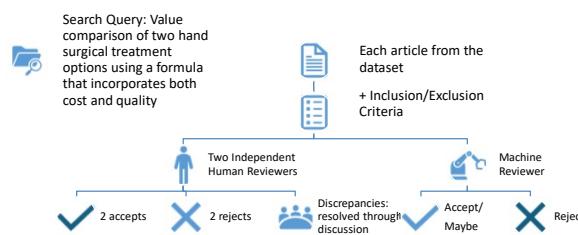


Figure 1. Workflow for systematic review screening using human and machine reviewers. Papers are first filtered based on inclusion/exclusion criteria. Human reviewers resolve discrepancies collaboratively, while machine reviewers make independent decisions on acceptance or rejection.

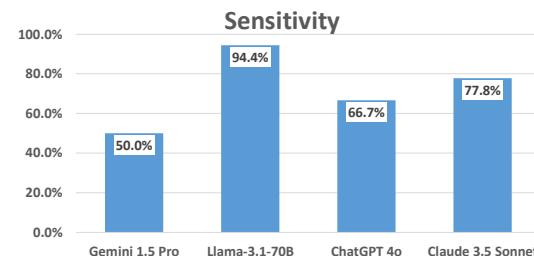


Figure 2. Sensitivity, proportion of true instances that are correctly predicted, for each models as compared to the golden standard, human reviewers.

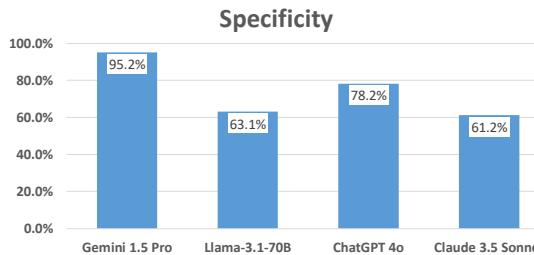


Figure 3. Specificity, proportion of false instances that are correctly predicted, for each models as compared to the golden standard, human reviewers.

Model	Sensitivity	Specificity	PPV	NPV	Accuracy	Rate (sec/article)	Time Savings versus Author Reviewers
Gemini 1.5 Pro	50.0%	95.2%	64.3%	91.6%	88.4%	8.5	5.6x
Llama-3.1-70B	94.4%	63.1%	30.9%	98.5%	67.8%	3.3	14.3x
ChatGPT 4o	66.7%	78.2%	35.3%	92.9%	75.2%	1.8	26.5x
Claude 3.5 Sonnet	77.8%	61.2%	25.9%	94.0%	63.6%	1.8	26.5x
Author Reviewers	-	-	-	-	-	47.2	-

Table 1. Table summarizing the performance of the LLM performance in comparison to human reviewers.

Results

- 121 papers were chosen from the query, human reviewers included 18 and excluded 103, serving as the golden standard for comparison
- On average, human reviewers took 47.2 seconds per article
- Gemini 1.5 Pro emerged as the model with the highest accuracy (88.4%), specificity (95.2%), and PPV (64.3%).
- Llama-3.1-70B demonstrated the highest sensitivity (94.4%) and NPV (98.5%).
- ChatGPT 4o and Claude 3.5 Sonnet were the fastest at 26.5x manual sorting
- Detailed performance metrics are summarized in Table 1

Conclusions

- All models significantly outperformed human reviewers in efficiency, with up to 26.5x faster processing.
- There was only one model, Llama-3.1-70B, surpassing the 90% sensitivity that was initially set as our hypothesis
- Despite their efficiency, AI models occasionally misclassified articles, highlighting the need for continued oversight by human reviewers.
- The small dataset limits generalizability to other areas of research.

Recommendations

- A hybrid approach is recommended using Llama-3.1-70B for the initial screening process due to its high sensitivity and then using human reviewers for the final screening

Future Research

- Optimizing performance using multiple AI models
- Further testing with other datasets in different fields

References

- Chiang W-L, Zheng L, Sheng Y, et al. Chatbot Arena: An open platform for evaluating llms by human preference. arXiv.org. March 7, 2024. Accessed November 30, 2024. <https://arxiv.org/abs/2403.04132>.
- Blanchard M. Closing the gap between medical knowledge and patient outcomes through new training infrastructure. Trans Am Clin Climatol Assoc. 2023;133:119-135.
- Borah R, Brown AW, Capers PL, Kaiser KA. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. BMJ Open. 2017;7(2):e012545. Published 2017 Feb 27. doi:10.1136/bmjopen-2016-012545
- Densen P. Challenges and opportunities facing medical education. Trans Am Clin Climatol Assoc. 2011;122:48-58.
- Popoff E, Besada M, Jansen JP, Cope S, Kanters S. Aligning text mining and machine learning algorithms with best practices for study selection in systematic literature reviews. Syst Rev. 2020;9(1):293. Published 2020 Dec 13. doi:10.1186/s13643-020-01520-5
- van Dijk SHB, Brusse-Keizer MGJ, Bucsán CC, van der Palen J, Doggen CJM, Lenferink A. Artificial intelligence in systematic reviews: promising when appropriately used. BMJ Open. 2023;13(7):e072254. Published 2023 Jul 7. doi:10.1136/bmjopen-2023-072254