AI Generation Of Evidence Summaries: A Descriptive Study Of The Comparability With Human Annotations

Michelle Colder Carras, Riaz Qureshi, Faisal Aldayel, Mayank Date, Dahlia AlJuboori, Johannes Thrul

Submitted to: JMIR Medical Education
on: December 05, 2024

Table of Contents

Original Manuscript..... 4

AI Generation Of Evidence Summaries: A Descriptive Study Of The Comparability With Human Annotations

Michelle Colder Carras[1] PhD; Riaz Qureshi[2] PhD; Faisal Aldayel[3] MD, MPH; Mayank Date[3] BDS, MPH; Dahlia AlJuboori[3] MBChB, MHS; Johannes Thrul[3] PhD

[1]Department of International Health Johns Hopkins Bloomberg School of Public Health Baltimore US
[2]Department of Ophthalmology, School of Medicine University of Colorado Anschutz Medical Campus Aurora US
[3]Department of Mental Health Johns Hopkins Bloomberg School of Public Health Baltimore US

Corresponding Author:
Michelle Colder Carras PhD
Department of International Health
Johns Hopkins Bloomberg School of Public Health
615 N. Wolfe St.
Baltimore
US

Abstract

Background: Annotated bibliographies comprise summaries of relevant literature, and training, experience, and time is required to create useful annotations. However, summaries generated by artificial intelligence (AI) can contain serious errors.

Objective: We compared the quality of human- and AI-generated annotations directly to determine strengths and weaknesses of both.

Methods: We compared five criteria (word count, readability, capture of main points, presence of errors, broader contextualization/quality) between human- and native ChatGPT-produced annotations for 15 academic papers using descriptive statistics and non-parametric testing.

Results: Humans produced shorter annotations (90.20 vs 111.47 words, Z= 2.82, P=.01) with better readability than AI (15.25 vs. 8.03, Z= -2.28, P=.02), although readability was low for all annotations. There was no difference in the capture of main points (X2 = 6.12, P=.18)
or presence of errors (X2 = 5.27, P=.16). AI-produced annotations provided better contextualization than human annotations (X2 = 11.28, P <.001).

Conclusions: AI-produced summaries of academic literature are comparable to human annotations. Annotations generated by AI and verified by humans should reduce the time needed to produce summaries on a given subject.

(JMIR Preprints 05/12/2024:69707)
DOI: https://doi.org/10.2196/preprints.69707

Original Manuscript

AI Generation of Evidence Summaries: A DESCRIPTIVE STUDY OF the comparability with human annotations

Michelle Colder Carras, Ph.D.[1,2*], Riaz Qureshi, Ph.D.[3], Faisal Aldayel, M.D., M.P.H.[1], Mayank Date, B.D.S., M.P.H[1], Dahlia Aljuboori, M.B.Ch.S., M.H.S.[1], Johannes Thrul, Ph.D.[1,4,5]

[1] Department of Mental Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, USA

[2] Department of International Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, USA

[3] Department of Ophthalmology, School of Medicine, University of Colorado Anschutz Medical Campus

[4] Sidney Kimmel Comprehensive Cancer Center at Johns Hopkins, Baltimore, MD, USA

[5] Centre for Alcohol Policy Research, La Trobe University, Melbourne, Australia

*** Correspondence:**

Michelle Colder Carras

Department of International Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, USA

mcarras@jhu.edu

Abstract

Background: Annotated bibliographies comprise summaries of relevant literature, and training, experience, and time is required to create useful annotations. However, summaries generated by artificial intelligence (AI) can contain serious errors.
Objective: We compared the quality of human- and AI-generated annotations directly to determine strengths and weaknesses of both.
Methods: We compared five criteria (word count, readability, capture of main points, presence of errors, broader contextualization/quality) between human- and native ChatGPT-produced annotations for 15 academic papers using descriptive statistics and non-parametric testing.
Results: Humans produced shorter annotations (90.20 vs 111.47 words, Z= 2.82, *P*=.01) with better readability than AI (15.25 vs. 8.03, Z= -2.28, *P*=.02), although readability was low for all annotations. There was no difference in the capture of main points (X2 = 6.12, *P*=.18)
or presence of errors (X2 = 5.27, *P*=.16). AI-produced annotations provided better contextualization than human annotations (X2 = 11.28, *P* <.001).

Conclusions: AI-produced summaries of academic literature are comparable to human annotations. Annotations generated by AI and verified by humans should reduce the time needed to produce summaries on a given subject.

Keywords: Large Language Model, ChatGPT, Artificial Intelligence, Evidence Synthesis, Annotated Bibliography, Information Management

INTRODUCTION

The fields of medicine and public health are dynamic and constantly evolving, with a continuous influx of new research articles and evidence. Staying up to date with the latest literature poses a significant challenge for researchers and practitioners alike (1), and it is difficult to familiarize oneself with the literature when diving into a new subject area (2). Evidence summaries such as meticulously curated reviews and guidelines are lengthy and often complex, and their number has exponentially increased over the past decade (3,4). In the past, individuals (oftentimes students) may have resorted to sources such as blogs or Wikipedia to understand the basics of a topic, ideally before diving into more reputable peer-reviewed publications. However, this approach has been superseded by use of AI large language models (LLM) such as ChatGPT, which provide summaries based on accessible scientific literature but often have questionable accuracy and frank errors (5,6).

As a research tool, annotated bibliographies may be a valuable remedy for efficient and insightful literature exploration. These condensed summaries highlight the main points, findings, and scientific validity of the respective articles, while also providing context by relating them to existing knowledge within the field (7).

Recent studies on the use of LLMs show mixed evidence for these as evidence synthesis methodologies. While newer models may provide good summaries, it is unclear whether they have improved in their ability to effectively summarize and analyze scientific materials.(5,6,8,9) AI Chatbots and LLM may have limitations in capturing nuances and answering specific requests accurately, in addition to their known challenges of hallucinating information and citations.(5,6,9) These limitations may stem from an inability to discern the main points of articles (for example, distinguishing factual statements in findings from background material) and may result in output that diverges from the intended context.(10) On the other hand, developing the skill to write succinct and

insightful annotations requires time and practice, as well as a solid background knowledge of the subject matter.(11) Keeping abreast of recent advancements and understanding them in the context of existing knowledge can be a daunting task, even for experienced professionals in the field.

Given the varying opinions about the utility of LLMs in generating summaries and citing them in context, our objective was to explore the quality and comparability of annotated bibliographies produced by authors of varying professions and levels of training with outputs generated by a widely-available large language model (ChatGPT). By analyzing the strengths and weaknesses of both human-authored and AI-generated annotated bibliographies, we aim to shed light on the potential benefits and limitations of utilizing AI in the context of literature exploration and synthesis in medicine and public health.

Methods

We created annotated bibliographies comprising 15 papers, then compared characteristics and quality of these bibliographies between the human annotators and chatbot-developed annotations. We hand-curated a selection of 15 papers in areas related to the biomedical sciences by first selecting categories from Web of Science, then choosing from among the most highly-cited open-access empirical studies in chosen categories (see Supplemental Information).

Our human annotators consisted of dentist who recent obtained a master's degree in public health and a faculty-level public health faculty. Papers were added to the evidence synthesis platform PICO Portal (12) and annotators were provided with instructions about what an annotation should comprise and how to create annotations in the platform.

To explore differences between chatbot versions and application, two pre-specified versions of ChatGPT (3.5 and 4) were used in March of 2023 to provide AI-generated summaries. Each of the 15 articles was exported into a text file, and a web-based splitter utility (13) was used to break up text into chunks to comply with ChatGPT's restrictions on text input length. After each text chunk was fed to the conversation, two researchers used a naturalistic approach to prompt generation to generate summaries for each article. As initial attempts to create summaries with chatbots often produced very long summaries, instructions were modified to include general word count recommendations. If the chatbot replied that it was unable to create a summary (three cases), additional prompts or the "Regenerate response" button were used until a summary was created. Full text of prompts and chatbot conversations can be found in the project OSF website (14). Differences in resulting prompts and summary quality led us to add additional sets of LLM-based annotations post-hoc using PICO Portal (12) and ScholarAI (an academically-oriented and trained plug-in for ChatGPT-4) (15). We did not formally evaluate these annotations for comparison, but describe their characteristics and provide the full annotations on OSF.

AI annotations were scrubbed for formatting anomalies (see Supplemental Information and OSF), blinded, and all annotations randomly ordered using random.org.

We calculated word count and Flesch Reading Ease score for each annotation using Microsoft Word and compared mean values between AI-generated and human annotations with two-sample Wilcoxon Rank Sum (Mann-Whitney). To assess quality, an epidemiologist specializing in evidence synthesis methods evaluated three characteristics: main points (whether the annotation captured the main points of the article rated on a scale of one to seven), presence of errors on a scale of one to four ranging from no errors to multiple errors, and context (quality of evidence and/or contextualization

with field, assessed as yes or no). A guess was also made as to whether the annotation was created by a human or AI, although this was not formally evaluated because the assessor had prior knowledge that the four annotations for each paper were evenly split between human and AI. Summary measures were created for each annotator and aggregated across annotator type, then provided in tables and narrative synthesis. Scoring results, annotations, and expert notes can be found on our OSF website.

Results

Annotation characteristics are described in Table 1. Whereas the AI produced summaries within seconds, the average times for each human annotator were 14 and 15 minutes (see OSF). Readability was low overall with the Flesch Reading Ease being below 30 for both sets of annotators. Although there was a significant difference between human and AI annotators, with human annotators receiving higher scores/better readability, these scores indicate the text from all summaries was very difficult to read and best understood by university/college graduates.(16) We noted that AI often had difficulty keeping to length restrictions when asked to respond in a certain number of words or sentences and produced annotations on average 20 words longer than humans. We did not find any differences between human and AI annotations in terms of capturing the main points of the articles (P = .18) or having errors in the text (P = .16). However, we did find a few cases of significant errors in both human and AI annotations (such as referring to the wrong author for the annotated paper, describing the wrong study design, or drawing an incorrect inference; see Supplemental Information). Interestingly, we found a difference in the proportion of annotations that provided context for the findings within the field: Of the 30 AI annotations, 21 (70%) provided some contextualization whereas only 8 (27%) of the human annotations did the same (P <.001).

Table 1. Annotation characteristics (n=60 annotations)

Indicator	**AI (n = 30)**		**Human (n = 30)**		**Test statistic (*P*-value)**
Readability, Mean (SD)	8.03	(7.95)	15.25	(12.45)	Z = -2.28 (P=.02)
Word count, Mean (SD)	111.47	(17.47)	90.20	(36.82)	Z = 2.82 (P=.01)
Main points, n (%)					X^2 = 6.12 (P=.18)
1 – Included no main points	0	(0.00)	0	(0.00)	
2	0	(0.00)	0	(0.00)	
3	2	(6.67)	1	(3.33)	
4 – Included a few minor points	3	(10.00)	4	(13.33)	
5	1	(3.33)	7	(23.33)	
6	10	(33.33)	9	(30.00)	
7 – Included all main points	14	(46.67)	9	(30.00)	
Errors, n (%)					X^2 = 5.27 (P=.16)
No errors	20	(66.67)	25	(83.33)	
Minor error	4	(13.33)	3	(10.00)	
Major error	0	(0.00)	1	(3.33)	
Multiple errors	6	(20.00)	1	(3.33)	
Context, n (%)					X^2 = 11.28 (P <.001)
Yes	21	(70.00)	8	(26.67)	
No	9	(30.00)	22	(73.33)	

Note: Each annotator produced annotations for 15 articles, resulting in 60 total annotations. Main

points evaluated whether the annotation captured the main points of the article; Errors evaluated presence of errors in the summary, and Context evaluated the quality of evidence and/or contextualization with field.

Analysis of reviewer's guess as to the origin of the annotation compared to the true origin showed low accuracy. Of the initial 60 masked annotations, the assessor correctly identified 18/30 (60%) AI-generated and 18/30 (60%) human generated annotations, suggesting that the annotations were more similar than different.

Discussion

Our goal for this exploratory study was to directly compare AI and human-generated evidence summaries to determine whether the AI-produced annotations would be so inappropriate that further development would not be fruitful. Despite being longer (and violating requested word counts), the annotations produced by ChatGPT-3.5 and ChatGPT-4 often needed additional prompting to produce annotations that were consistent and compliant with the desired output. We also found some artefacts in the AI annotations that triggered identifying it as an automated summary and counted as a mistake. Despite those rare errors, overall, we found the AI produced annotations with similar number of errors and better contextualization to those produced by humans and in much shorter time.

Although we did not formally evaluate our post-hoc annotations using PICO Portal and ScholarAI, we felt both of these performed better than base ChatGPT-3.5 and ChatGPT-4, although the issues of annotation length and reading difficulty were still apparent. We expected these approaches to perform better because the models were trained on scientific literature, which is more specialized than the base ChatGPT portal on OpenAI's website. There are many apps and platforms that build upon existing large language model frameworks (e.g., Claude 3, Gemini, Bard, Co-Pilot, etc.) using more specific training sets to help with specialized tasks. Some, such as ScholarAI, have been trained on scientific articles and have access to PubMed and bibliographic databases – a feature that base ChatGPT does not have – which gives them a greater ability to provide useful contextualization and even suggest potentially related references.
This was supported by our finding that AI annotations more often provided some contextualization than did the human annotations. While the first AI prompts did not include a request to contextualize the findings within the broader literature, it may be that the AI could infer what was "expected" given the request to make a summary for an annotated bibliography. A difference in the effective contextualization likely arose because LLMs are predictive models and may be able to produce an overall statement and extrapolation/contextualization about the literature on a topic based on whatever is available on in its training set – a task that is very difficult for humans unless they are familiar with the literature already.

A limitation of using LLMs is the varying nature of output and the reliance on the prompts that were used. Although LLMs are functionally large predictive models with trillions of parameters, given the same prompt, a model will return different results each time a prompt is run. This is a limitation of using these models to summarize literature, because different annotations will be produced for the same article. While this variability is also true of multiple human annotators, and both humans and AI can make mistakes, humans may be better able to adapt prompts or instructions to different settings and situations in a way that LLMs cannot. The potential risk is that a carefully written prompt that produces an accurate summary of a paper in one LLM may not produce a reliable or valid summary (1) of a paper that has a different structure (e.g., social science versus medical literature) or (2) in an LLM built upon a different framework. Another limitation is the potentially

low readability, however in our study, this may have come from the fact that we were summarizing scientific articles and did not ask the AI to write for a better readability score.

As our intention was to explore the comparability of AI and human annotations, our sample was small. For our generation of annotations, assessments, and comparisons, we followed a pre-defined protocol as closely as possible. We generated further annotations using other models than we had originally specified, however we did not conduct any formal analyses with these as they were created post-hoc.

Contrary to other studies that look only at AI summaries or answers to questions, (5,6,9,17) our findings suggest that LLMs produce summaries of academic articles that are comparable to human annotators in errors and contextualization. While the annotations were not perfect and sometimes needed further prompting to produce better summaries, summaries were created 15 times faster than humans and did not require the experience and training of human researchers. We believe the utility is clear, and that human-assisted AI annotation – whereby the AI generates the summary with accuracy verified by a human – will be the most efficient way to create annotated bibliographies that can be used to provide a rapid overview of a breadth of literature for a wide range of users.

ACKNOWLEDGMENTS

Authors [blinded] contributed to the study concept and design; authors [blinded] contributed to the analysis and interpretation of data; authors [blinded] obtained funding, and authors [blinded] contributed to study supervision. All authors had full access to all data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. This research was supported by a grant from Aramco Services Company, a subsidiary of Saudi Aramco. The funders had no role in conducting the study or reporting the results.

CONFLICT OF INTEREST

Author Riaz Qureshi declares consulting with PICO Portal from 2020-2024. The other authors declare no competing interests.

ABBREVIATION

AI: Artificial Intelligence

ChatGPT: Chat Generative Pre-Trained Transformer.

OSF: Open Science Framework

PICOS: population, intervention, comparator, outcome, and study design

References

Note: Summarized article citations can be found in Supplemental Information.

1. Bastian H, Glasziou P, Chalmers I. Seventy-Five Trials and Eleven Systematic Reviews a Day: How Will We Ever Keep Up? PLOS Medicine [Internet]. 2010 Sep 21 [cited 2024 Apr 12];7(9):e1000326. Available from: https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1000326

2. Kraker P, Kittel C, Enkhbayar A. Open Knowledge Maps: Creating a Visual Interface to the World's Scientific Knowledge Based on Natural Language Processing. 0277 Zeitschrift für Bibliothekskultur [Internet]. 2016 Nov 11 [cited 2024 Apr 12];4(2):98–103. Available from: https://zenodo.org/records/4705327

3. McKenzie JE, Brennan SE. Overviews of systematic reviews: great promise, greater challenge. Systematic Reviews [Internet]. 2017 Sep 8 [cited 2024 May 7];6(1):185. Available from: https://doi.org/10.1186/s13643-017-0582-8

4. Lunny C, Reid EK, Neelakant T, Chen A, Zhang JH, Shinger G, et al. A new taxonomy was developed for overlap across "overviews of systematic reviews": A meta-research study of research waste. Research Synthesis Methods [Internet]. 2022 [cited 2024 May 7];13(3):315–29. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/jrsm.1542

5. Chelli M, Descamps J, Lavoué V, Trojani C, Azar M, Deckert M, et al. Hallucination Rates and Reference Accuracy of ChatGPT and Bard for Systematic Reviews: Comparative Analysis. Journal of Medical Internet Research [Internet]. 2024 May 22 [cited 2024 May 26];26(1):e53164. Available from: https://www.jmir.org/2024/1/e53164

6. Gravel J, D'Amours-Gravel M, Osmanlliu E. Learning to Fake It: Limited Responses and Fabricated References Provided by ChatGPT for Medical Questions. Mayo Clinic Proceedings: Digital Health [Internet]. 2023 Sep 1 [cited 2024 May 26];1(3):226–34. Available from: https://www.mcpdigitalhealth.org/article/S2949-7612(23)00036-6/fulltext

7. Basham SL, Radcliff VP, Bryson SL. How to Write an Annotated Bibliography. Journal of Criminal Justice Education [Internet]. 2023 Apr 3 [cited 2024 May 7];34(2):289–97. Available from: https://doi.org/10.1080/10511253.2022.2131859

8. Qureshi R, Shaughnessy D, Gill KAR, Robinson KA, Li T, Agai E. Are ChatGPT and large language models "the answer" to bringing us closer to systematic review automation? Systematic Reviews [Internet]. 2023 Apr 29 [cited 2024 Apr 12];12(1):72. Available from: https://doi.org/10.1186/s13643-023-02243-z

9. Day T. A Preliminary Investigation of Fake Peer-Reviewed Citations and References Generated by ChatGPT. The Professional Geographer [Internet]. 2023 Nov 2 [cited 2024 May 26];75(6):1024–7. Available from: https://doi.org/10.1080/00330124.2023.2190373

10. Tang L, Sun Z, Idnay B, Nestor JG, Soroush A, Elias PA, et al. Evaluating large language models on medical evidence summarization. npj Digit Med [Internet]. 2023 Aug 24 [cited 2024 May 7];6(1):1–8. Available from: https://www.nature.com/articles/s41746-023-00896-7

11. Flaspohler MR, Rux EM, Flaspohler JA. The Annotated Bibliography and Citation Behavior: Enhancing Student Scholarship in an Undergraduate Biology Course. LSE [Internet]. 2007 Dec [cited 2023 Jun 8];6(4):350–60. Available from: https://www.lifescied.org/doi/full/10.1187/cbe.07-04-0022

12. PICO Portal. PICO Portal: Systematic review management platform [Internet]. [cited 2024 Dec 4]. Available from: https://picoportal.org/

13. ChatGPT Splitter. Web-based splitter utility [Internet]. [cited 2024 Dec 4]. Available from: https://chatgptsplitter.com/

14. OSF. Study data and materials [Internet]. [cited 2024 Dec 4]. Available from: https://osf.io/cmt7r/?view_only=24440ba39cb04fbcb4f297f8217a465e

15. ScholarAI. An academically-oriented and trained plug-in for ChatGPT-4 [Internet]. [cited 2024 Dec 4]. Available from: https://scholarai.io/

16. Flesch–Kincaid readability tests. In: Wikipedia [Internet]. 2024 [cited 2024 May 7]. Available from: https://en.wikipedia.org/w/index.php?title=Flesch%E2%80%93Kincaid_readability_tests&oldid=1214583693

17. Alkaissi H, McFarlane SI, Alkaissi H, McFarlane SI. Artificial Hallucinations in ChatGPT: Implications in Scientific Writing. Cureus [Internet]. 2023 Feb 19 [cited 2024 May 26];15(2). Available from: https://www.cureus.com/articles/138667-artificial-hallucinations-in-chatgpt-implications-in-scientific-writing

Acknowledgments

Generative AI (ChatGPT-3.5) contributed to writing the background section of this article, which was revised extensively by the authors.

Competing interests

Author MCC [blinded] conducts consulting related to video games and well-being and has held leadership positions in nonprofit organizations related to the promotion of healthy video gaming. The All other authors declare no other conflicts of interest.

Funding

This research was supported by a grant from Aramco Services Company, a subsidiary of Saudi Aramco. The funders had no role in conducting the review or reporting the results.

Data availability

The datasets generated for this study and other documents can be found in the project's OSF website (anonymous link for review: https://osf.io/cmt7r/?view_only=24440ba39cb04fbcb4f297f8217a465e)