Evidence Based Library and Information Practice

Evidence Summary

Popular GenAI Chatbots Vary in Capabilities to Answer Academic Reference Questions

A Review of:

Whitfield, S., & Yang, S. Q. (2025). Evaluating AI language models for reference services: A comparative study of ChatGPT, Gemini, and Copilot. *Internet Reference Services Quarterly, 29*(2), 153-167. https://doi.org/10.1080/10875301.2025.2478861

Reviewed by:
Lisa Shen
Business Librarian & Director of Public Services
Newton Gresham Library
Sam Houston State University
Huntsville, Texas, United States of America
Email: lshen@shsu.edu

Received: 12 Nov. 2025 **Accepted:** 21 Jan. 2026

DOI: 10.18438/eblip30949

Abstract

Objective – To assess and compare the quality of responses to reference questions by popular generative artificial intelligent (GenAI) chatbots.

Design – Content analysis.

Setting – Web browser platforms of four GenAI chatbots.

Subjects – Responses from ChatGPT 3.5, Chat GPT 4.0, Gemini, and Copilot to a set of 28 chat reference questions submitted by Rider University Library patrons between July 1, 2023, and May 14, 2024.

Methods – Transcripts of 112 responses from the four chatbots were evaluated using a content analysis scheme adapted from a previous ChatGPT study (Yang & Mason, 2024). Both researchers

independently rated every response using a 10-point scale on four categories, accuracy, relevance, friendliness, and instructiveness, and analyzed the results using inferential statistics.

Main Results – Responses from Gemini received the highest total score (592 out of a possible 1,120 points), followed by ChatGPT 4.0 (542), ChatGPT 3.5 (502), and Copilot (433). However, every chatbot fluctuated greatly in their performance in accuracy and relevance. A single-factor ANOVA test also found statistically significant differences between the quality of the GenAI chatbots' responses in relevance and friendliness, with Gemini and Copilot performing the best in these categories respectively. There were no statistically significant differences between the chatbots' performances in accuracy or instructiveness, although Gemini held the highest mean score for instructiveness.

Conclusion – The researchers concluded that popular GenAI chatbots should be used to supplement, not replace, the work of reference and instruction librarians, and noted the potentials for training GenAI to address basic or local, library-specific FAQs for after-hour reference support. The authors also advised librarians to stay current with the rapid development of chatbots and other GenAI tools and to continue differentiating librarianship competencies from services offered by these programs.

Commentary

This article presented a unique perspective amidst the influx of research on GenAI applications in library and information studies by assessing the competencies of GenAI chatbots in an academic reference setting. While emphasizing the timeliness of this research topic, the authors also conceded that, given the swift development of GenAI technologies, their assessment of these chatbots' performances could quickly become outdated. Even so, the study's focus on comparing the capabilities of popular chatbots highlighted nuances not often addressed by critiques of GenAI use in reference work.

An assessment using the EBL Critical Appraisal Checklist (Glynn, 2006) yielded an overall validity score of 86% for this study, with sectional validities for population, data collection, study design, and results all exceeding the acceptable threshold of 75%. The process for selecting chatbots and testing reference questions was logical and clear, and the question set, the assessment categories, and the scoring scale were included in the publication with sufficient detail to enable replication. The final inter-coder reliability of over 90% between the two researchers was also commendable.

However, both researchers were librarians at Rider University with reference responsibilities, and their connections to the test reference questions and personal perceptions of the GenAI chatbots could cause potential subjectivity and bias when scoring chatbot responses. Also, the researchers stated that "no hallucinations" (p. 158) is required for high scores in all four criteria, but the concept is only mentioned in the assessment for accuracy. For instance, how would—or should—the presence of a hallucination impact the friendliness dimension of a response? Finally, the researchers did not provide a timeframe for generating the test chats. Even though chat conversations cannot be replicated due to the generative nature of AI chatbots, given the rapid progress of GenAI technology, it could still be informative to know the approximate dates when these interactions took place.

The authors addressed some of these concerns by recommending that future researchers blind the chatbot responses when scoring. They also noted the study's lack of generalizability due to its relatively small sample size and the shortcomings of standardizing the test process. By limiting each chat interaction to the initial reference question, the study could not fully assess the chatbots' capacity for iterations—a common challenge yet to be resolved by LIS researchers. Nonetheless, this article provided a sound foundation and methodology option for academic library professionals interested in understanding and differentiating the capabilities of different GenAI chatbots and captured a snapshot of these chatbots' performances for posterity.

References

Glynn, L. (2006). A critical appraisal tool for library and information research. *Library Hi Tech, 24*(3), 387–399. https://doi.org/10.1108/07378830610692154

Whitfield, S., & Yang, S. Q. (2025). Evaluating AI language models for reference services: A comparative study of ChatGPT, Gemini, and Copilot. *Internet Reference Services Quarterly, 29*(2), 153-167. https://doi.org/10.1080/10875301.2025.2478861

Yang, S. Q., & Mason, S. (2024). Beyond the algorithm: Understanding how ChatGPT handles complex library queries. *Internet Reference Services Quarterly, 28*(2), 97–151. https://doi.org/10.1080/10875301.2023.2291441