

How Well Does ChatGPT Handle Reference Inquiries? An Analysis Based on Question Types and Question Complexities

Katie Lai

To explore whether artificial intelligence can be used to enhance library services, this study used ChatGPT to answer reference questions. An assessment rubric was used to evaluate how well ChatGPT handled different question types and difficulty levels. Overall ChatGPT's performance was fair, but it did poorly in information accuracy. It scored the highest when handling facilities and equipment-related questions but the lowest when dealing with e-resources access problems. ChatGPT was weak in answering advanced research questions, complex inquiries, and known item searches relating to a specific local environment, but it could be adopted to enhance library communication with users.

Introduction

The launch of ChatGPT has created a wave of discussion on artificial intelligence (AI). While some are amazed by its ability to provide answers on wide-ranging topics in conversational style, others are skeptical about the accuracy and credibility of the information it provides. Aiming to enhance library reference services, this study used ChatGPT to answer reference questions emailed to the Marvin Duchow Music Library of McGill University between September 2022 and February 2023. An analysis using an assessment rubric was conducted to evaluate how well ChatGPT handled different types of questions and different difficulty levels based on the Reference Effort Assessment Data (READ) Scale. Statistical tests were employed to see whether there were statistically significant associations between variables. The goals are to explore whether ChatGPT could be used to enhance the quality and efficiency of the current music reference services, whether it can directly handle inquiries raised by users, and whether it can offer relevant information as a first step for librarians to handle complex research questions.

Literature Review

AI has become an integral part of everyone's life, from checking the weather forecast using digital voice assistants such as Alexa to navigating in the city using Tesla's full self-driving feature. In the context of libraries, the use of chatbots began as early as mid-2000s.¹ McNeal and Newyear gave an overview on the history of the use of chatbots in libraries and highlighted some of the early initiatives, such as *Stella*, developed by the Bibliothekssystem Universität Hamburg in 2004;

* Katie Lai is Associate Librarian and Liaison for Music at McGill University, email: katie.lai@mcgill.ca. © Katie Lai, Attribution-NonCommercial (<http://creativecommons.org/licenses/by-nc/4.0/>) CC BY-NC

Emma the Catbot, used by the Mentor Public Library in Ohio from 2009 to 2012; and *Pixel*, written in 2010 by the University of Nebraska-Lincoln Libraries.² These chatbots were designed to answer general library questions. In 2011, Tsinghua University Library created an AI talking robot Xiaotu to provide real-time virtual reference with capabilities to learn new knowledge from users through questions and answers.³ University of California Irvin Libraries also built the chatbot ANTswers in 2013 to handle simple and repetitive questions.⁴ In 2018, San Jose University started to develop its AI library chatbot, Kingbot, using Google's Dialogflow to answer basic circulation and introductory reference inquiries.⁵ California State University San Marcos also used Dialogflow to create a chatbot to help professors answer assignment and syllabus-related questions outside of class.⁶ Then, as chatbots were further developed, they began to focus not only on users' information needs but also their other needs. For instance, University of Technology Sydney's Lib-bot was designed to help undergraduates overcome research and library anxiety. With a potential to embed it into online learning management systems, it could proactively offer research advice before an upcoming assignment due date.⁷ In 2019, the California State University system introduced their chatbots to connect with students remotely and build rapport to help them get on track with their studies during the COVID outbreak.⁸ Thus, the implementation of chatbots has seen a gradual expansion in libraries and higher education.

In November 2022, a revolutionary chatbot called ChatGPT was launched. Developed by OpenAI and fine-tuned from the GPT-3.5 large language model (LLM), ChatGPT, which stands for Chat Generative Pre-Trained Transformer, is able to understand user inputs and interact in a human, conversational way.⁹ With a large corpus of data, it can produce relevant responses on a wide range of topics and handle language tasks such as translation and summarization. It was such a big hit to the world that within the first five days of its launch ChatGPT has attracted over one million users.¹⁰

Seeing these AI advancements, the author wanted to see if ChatGPT could be used to support library reference services. A search for literature about ChatGPT and library reference services in Google, Google Scholar, and the Library, Information Science, and Technology Abstracts database yielded little result. Those who attempted to evaluate ChatGPT's performance in answering academic research questions mostly came from blogs and websites. For instance, Davis asked ChatGPT about two scientific controversies. While one of them was answered correctly, the other replied with a fabrication of "scientific evidence."¹¹ Similarly, Kendrick asked ChatGPT to provide information on a research topic and its related citations. While the writing part was of a comparable quality of a Wikipedia article, ChatGPT "failed miserably" in the citations provided.¹² In the comments column of *Nature*, Van Dis et al. noted that ChatGPT's answers to questions that require in-depth subject knowledge were exceedingly general or often contained factual errors and misrepresentations.¹³ They advocate four priorities for research, including an author-contribution statement if AI technology is used, the non-recognition of LLMs as authors, more transparency in publishing policy and LLMs' underlying training sets, and investments into open-sourced independent non-profit AI technologies by universities and scientific-funding organizations in order to minimize possible biases produced by the underlying datasets and algorithms used by commercial enterprises.¹⁴ The concerns for these "datasets and algorithmic *black boxes*" were also echoed by Nayyer and Rodriguez, who flagged the potential danger of using them as a tool to violate academic library professional standards, patron respects, or ethical standards.¹⁵

Other than the above commentary-like articles, which are casual in nature, there is only one original study that more closely resembles scholarly research. In Chen's study, questions

were posed to both ChatGPT and traditional library chatbots, and their answers were compared.¹⁶ Though attempting to discuss the impact of AI in library reference services, Chen's study was however somewhat limited as only five questions were submitted to ChatGPT, and just one of the answers was compared with traditional library chatbots. Furthermore, the questions raised in Chen's research, as well as in the other articles mentioned above, were topics or questions that happened to come to the writers' minds. There was also no structured and systematic approach to analyze ChatGPT's performance. Seeing the void, the author therefore conducted this analysis based on question types and question complexities in order to understand ChatGPT's ability to handle inquiries received in an academic library setting.

Background

McGill University is a large research institution in Canada with a student body of around 39,000. Its Schulich School of Music offers undergraduate to doctoral programs in diverse subjects such as orchestral instruments, opera, jazz, and sound recording. The Marvin Duchow Music Library, one of the twelve branches of McGill Library, is charged to support the teaching, learning, research, and performance needs of the School. Its clientele, however, goes beyond current students and staff to include alumni and community members because of its large and unique collections of music materials.

The Music Library maintains an email to which both McGill- and non-McGill-affiliated users could send inquiries about its collections, services, research, and any music- or library-related matters. The questions received therefore cover all levels of studies and all disciplines of music, from performance to music technology. The email account is monitored by the Reference Team, which comprises music librarians and senior library assistants. Team members, in addition to a music degree, also possess a master of library science or are in pursuit of one. It is through this matrix of knowledge and expertise and a strong collaborative support system among team members that the Music Library ensures a high quality of reference services. The continuous pursuit of service excellence and efficiency thus motivates this research.

Methodology

An analysis using the questions received by the Music Library's designated email was conducted. The complexity of each question was rated using the READ Scale, and the answers provided by ChatGPT were evaluated using an assessment rubric. Fisher's Exact tests were used to determine whether there were statistically significant associations between the quality of ChatGPT's answers and the complexity and types of questions handled.

Since the intent of this study is to see whether ChatGPT could be incorporated to complement and/or enhance existing library services, McGill University's Research Ethics Board Office advised that an analysis conducted for program evaluation and quality improvement purposes as such does not need a research ethics approval.

Pool of Reference Questions

The 58 reference questions sent to the Music Library's designated email address between September 2022 and February 2023 were included in this study. Inquiries that took place verbally at the service desk were excluded, since there was no verbatim record of the actual reference interviews. To give a general picture of the nature of questions received, these inquiries were categorized into seven types (table 1).

TABLE 1
Seven Question Types (n=58)

Question Type	Examples	No. of Questions in the Study
Acquisitions	Purchase request	7
E-resources access problem	Remote access problem, failed to access e-resources	3
Facilities and equipment	Noise complaint, problem with computers	5
Known item search	Search for a specific title (either in the library or through ILL)	15
Other	Student jobs, donation	10
Patron records and policies	Extend due date, overdue fine, alumni access	8
Research	Find materials on a topic	10

To reflect the complexities of the questions involved, the inquiries were also ranked according to the READ Scale.¹⁷ It is a six-point scale, with 1 for questions that need the least amount of effort and no specialized knowledge, skills, or expertise, to 6 for questions that require the most effort and time and in-depth research.¹⁸ The READ Scale reflects “the effort, skills, knowledge, teaching moment, techniques and tools utilized by the librarian during a reference transaction,” and is used by over 400 libraries worldwide.¹⁹

Assessment Rubric

An analytic rubric was created to evaluate the quality of answers by ChatGPT (table 2). Three aspects, namely completeness, accuracy, and the provision of further assistance, were examined in order to produce meaningful insights on ChatGPT’s strengths and weaknesses and to avoid an overly general impression of its performance. The accuracy of all information provided was verified by the author and in certain cases also in consultation with members

TABLE 2
Assessment Rubric

Criteria	Quality of Response		
	1 Poor	2 Fair	3 Good
Completeness	Did not address any of the user’s question(s)	Only addressed some (part) of the user’s question(s)	Completely addressed all the question(s) raised
Accuracy	None of the information provided was correct	Provided both correct and incorrect information	All information provided was correct
Further assistance	Did not do any of the following: • Referred to other relevant sources/help when not able to fully answer question, or provided accurate additional information beyond initial inquiry; • Invited user to contact a librarian	Only did one of the following: • Referred to other relevant sources/help when not able to fully answer question, or provided accurate additional information beyond initial inquiry; • Invited user to contact a librarian	Did all of the following: • Referred to other relevant sources/help when not able to fully answer question, or provided accurate additional information beyond initial inquiry; • Invited user to contact a librarian

of the Reference Team to ensure that the answers were not “hallucination” or fabricated by AI. The relevancy of the information was assessed based on the context of the question and how the information would be used. Thus, if an answer was factually correct on its own but was unrelated to the essence of the inquiry or not deemed to be helpful to the user given the context, it was not considered as relevant or having provided further assistance.

ChatGPT

The author created a free account in February 2023. ChatGPT Feb 13 version with training data cut off in September 2021 was used to answer the reference questions in this study.²⁰

FIGURE 1
ChatGPT's General Suggestion on a Known Item Search



I'm on the hunt for a specific book:

"Searching for the Baroque spirit An exploratory study of teaching strategies for the successful performance of Baroque keyboard music on the modern piano" by Erika Bothma. I've found it in the "libraries worldwide" catalogue, and it seems to be an eBook but with no way to access it. I would rather not buy the book but I'm very interested in it for a presentation/paper this semester. Do you know if there's any way to get any access to this book?



It can be frustrating when you locate a book you need but can't seem to access it. Here are a few options to consider:

1. Interlibrary Loan: You can request the book through interlibrary loan (ILL) from your local library or university. ILL allows you to borrow materials from other libraries. Your library may be able to obtain a physical copy of the book or a digital copy if it is available.
2. Contact the Author or Publisher: You can try contacting the author or publisher of the book to see if they can provide you with a copy or direct you to a source where you can access it.
3. Open Access Repositories: Some academic institutions or organizations may have open access repositories that contain scholarly works, including dissertations and theses, which might include the book you're looking for. You can try searching open access repositories such as JSTOR, Google Scholar, or the Directory of Open Access Books.
4. Online Marketplace: You can try checking online marketplaces like Amazon or AbeBooks to see if the book is available for purchase or for rent at a lower cost.

I hope one of these options helps you to access the book you need for your presentation/paper!

Note that although ChatGPT Mar 14 version (GPT-4), which could handle advanced reasoning and complex instructions, was launched on March 14, 2023, there were no updates to the free accounts as GPT-4 was only available to paid subscribers at the time of writing this paper.

Process

Each question was copied from the email and pasted into ChatGPT. Sensitive, confidential, or personally identifiable data were removed or replaced with fictitious data prior to entering into the prompts.

Analyses

Qualitative Analysis

Below are selected examples of the “conversations” with ChatGPT.

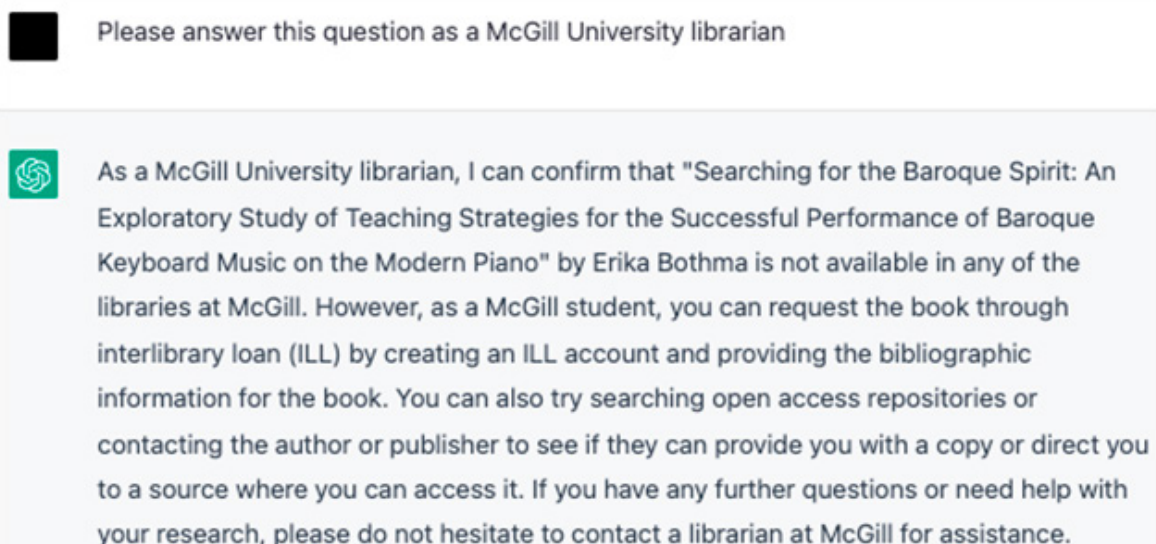
Example 1. Known Item Search

Since this was the first question entered in this free ChatGPT account, the exact text from the email was copied into the chat box. It was a known item search for a book that McGill did not own but was available in other libraries. The user wanted to see if they could access it without paying a fee or buying it. According to the READ Scale, it was a level 2 question.

In the first response, ChatGPT provided general suggestions of using interlibrary loan, contacting the author/publisher, checking open access repositories, and buying/renting it from online bookstores (figure 1).

While the options seemed sensible, they did not relate to McGill Library. So, the author revised the strategy and added “I’m a McGill University student” before retyping the question. This time, ChatGPT learned to tailor the answer to McGill Library and included a suggestion to search the McGill Library catalog. However, the response was still considered too general. Thus, the author once again tweaked the question and instructed ChatGPT to answer using a different role, as a McGill University librarian. This time, ChatGPT confirmed that McGill

FIGURE 2
ChatGPT’s Response as a McGill Librarian on a Known Item Search



Library did not own this book. It offered instructions on how to request an interlibrary loan (though not entirely correct), in addition to the options it had provided previously. It also invited the user to contact a McGill librarian should there be a need for further assistance (figure 2).

So, is the last version of the answer satisfactory? By telling ChatGPT to respond as a McGill librarian, it has learned to make reference to McGill Library. Though with some inaccuracies, the instruction to request the book through interlibrary loan was deemed useful. However, was it correct for ChatGPT to say that this title was not available at any McGill libraries? According to OpenAI, the free ChatGPT cannot search databases or access information outside of its static training data.²¹ So, it was likely that ChatGPT acknowledged the unavailability of the book by simply repeating what the user had inputted without checking the actual holdings in McGill's library catalog. In this sense, using the assessment rubric in table 2, this ChatGPT answer received a score of 3 for completeness (because it did fully address the user's question), 1 for accuracy and 2 for further assistance.

Example 2. Fact Finding Relating to Historical Research

The next question entered into ChatGPT was more complex. It was a research type question and fell under level 5 of the READ Scale as it required subject expertise, research skills, and consultations with multiple sources.

FIGURE 3
ChatGPT's Response to a Fact Finding Search for Historical Information



In the program note I wrote, I was challenged that the first performance of Beethoven's piano concerto no. 5 (Emperor) was not in Nov 1811, but some time earlier. Could you please check it for me? I know a book published in 2014 by Henle may have the info.



I'd be happy to help you fact-check the information about the premiere of Beethoven's Piano Concerto No. 5, also known as the "Emperor" concerto.

According to the standard reference work for Beethoven's works, "Thematic Catalogue of the Works of Ludwig van Beethoven" by Georg Kinsky and Hans Halm, the first performance of the concerto took place in Leipzig on November 28, 1811. This premiere featured the composer himself as the soloist.

However, it's worth noting that there were a few private performances of the concerto before this public premiere, which may be the source of confusion. Beethoven played the concerto for a small group of friends in Vienna in 1809, and he also played it in Prague in 1811, although this performance was also private.

As for the book you mentioned, the Henle edition of the concerto may provide more information about the early performances of the work, but it's unlikely to contradict the information about the Leipzig premiere. If you'd like, I can try to help you locate a copy of the book to confirm this.

Here the user was trying to verify the premiere date of Beethoven's Piano Concerto no. 5 from a 2014 publication by Henle because the date was believed to be different from what had been stated in most reference works (figure 3). In the first attempt, ChatGPT provided a date from the *Thematic Catalogue of the Works of Ludwig van Beethoven*, a major music reference work, but it also mentioned without citing the source(s) that there were a few private performances in Vienna in 1809 and in Prague in 1811 before the public premiere. With regard to the 2014 Henle publication, it is worth noting that ChatGPT did not recognize that the *Thematic Catalog* was indeed the 2014 Henle publication mentioned in the inquiry. Instead, it offered to search for a book by Henle to confirm the information quoted in the *Thematic Catalog*. Thus, ChatGPT failed entirely to correlate with this basic fact.

When being asked to provide the source relating to the private performances suggested, ChatGPT quoted a passage from the liner notes of a sound recording by the famous pianist Artur Schnabel produced by EMI Classics in 2002. This would have been an impressive discovery, if it had been true (figure 4).

Upon checking various sources and consulting with a Reference Team member, ChatGPT was correct that the first public performance date recorded in the latest 2014 edition of *Ludwig van Beethoven: Thematisch-bibliographisches Werkverzeichnis* (i.e., *Thematic Catalog of the Works of Ludwig van Beethoven*) was November 28, 1811, in Leipzig. However, the soloist was not Beethoven himself as claimed by ChatGPT, but Friedrich Schneider.²² As for the private performance prior to the public premiere, there was one in Vienna in the palace of Prince Lobkowitz, but it was in 1811,²³ not 1809 as ChatGPT stated. In fact, the author was not able to locate any documents that had a mention of the private performance in 1809 in Vienna or 1811 in Prague, nor the music album cited by ChatGPT.

FIGURE 4
ChatGPT's Response to the Request of Information Source



Please provide the source for "Beethoven played the concerto for a small group of friends in Vienna in 1809, and he also played it in Prague in 1811, although this performance was also private."



The information about Beethoven's private performances of his Piano Concerto No. 5 comes from the notes to the recording of the concerto by pianist Artur Schnabel, as included in the album "Beethoven: Piano Concerto No. 5 in E-Flat Major, Op. 73 'Emperor'," which was released by EMI Classics in 2002.

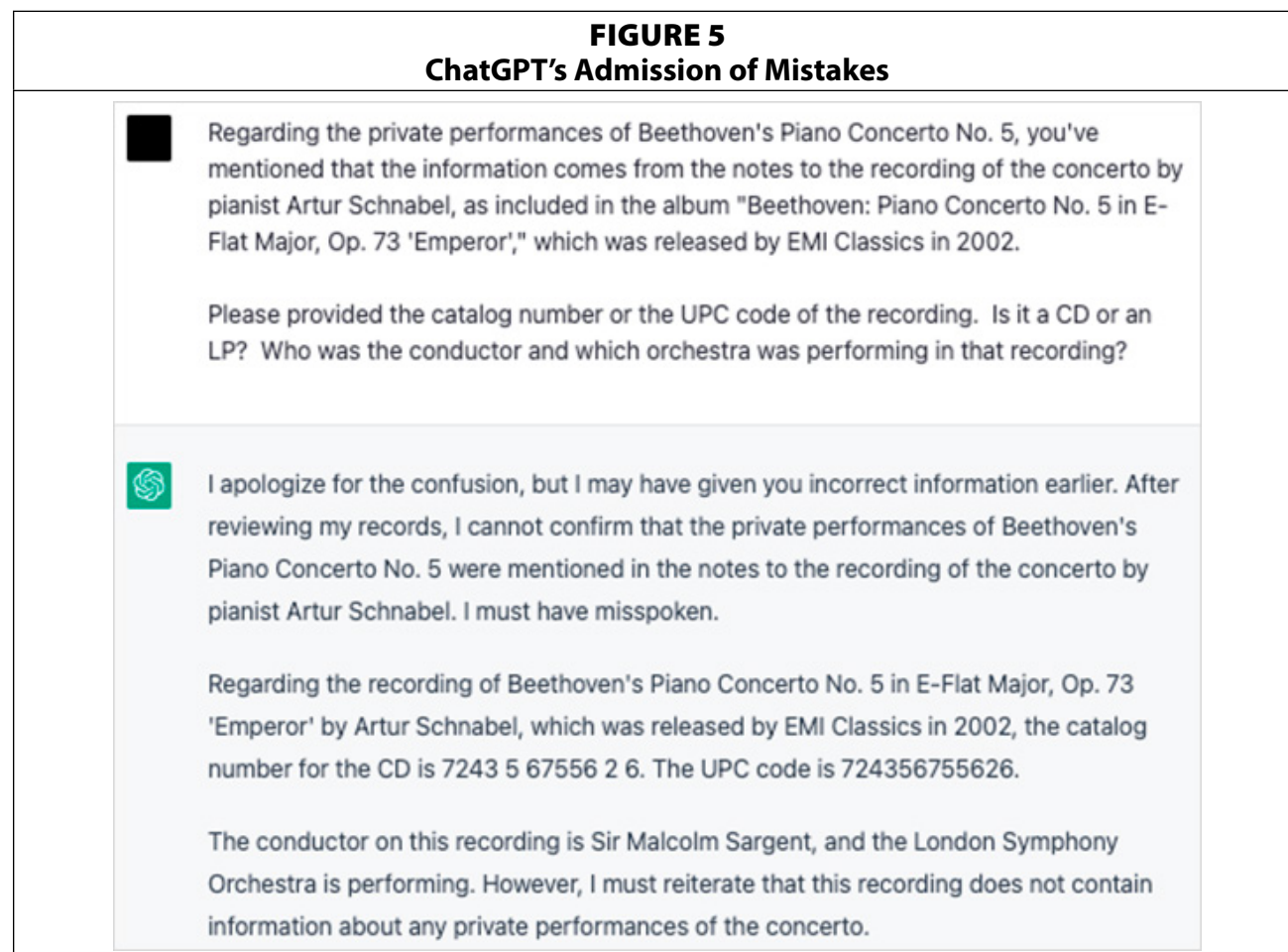
In these notes, Schnabel writes:

"Beethoven played his new concerto for a small group of friends in Vienna in 1809, but the first public performance was not until November 28, 1811, in Leipzig. He played the concerto again in Prague in 1811, but this performance was also private."

I hope this helps clarify the information for you!

When the author pushed for more details about the sound recording suggested, ChatGPT finally admitted that there was no mention of the private performances in the recording, and it has misspoken (figure 5).

FIGURE 5
ChatGPT's Admission of Mistakes



From this conversation, it is apparent that ChatGPT was quite confused, picked up bits and pieces of information from here and there, and mixed them together without adhering to the facts or what was written in the reference work. Hence, based on the wrong information provided and the lack of any proof for the alleged private performances, ChatGPT's answer to this question was far from satisfactory and was therefore rated with a 3 for completeness (because it did fully address the user's question), 1 for accuracy, and 1 for further assistance (since it did not suggest that the user contact a library for further assistance or provide additional accurate information beyond the initial inquiry).

Example 3. Identify a Musical Work on the Radio

This was also a research type question and fell under level 5 of the READ Scale due to the lack of specificity, the inclusion of potentially wrong information in the inquiry, and the possibility of false leads.

The user would like the Music Library to identify a sound recording of a Mozart quintet performed by the Menuhin Ensemble, heard by a friend on the Sirius XM radio. Here, ChatGPT again performed poorly by first making up information that the Menuhin Ensemble

was a student ensemble at McGill University. Then, ChatGPT suggested that the user contact McGill's Schulich School of Music for a copy of the recording (figure 6).

FIGURE 6
ChatGPT's Identification of a Musical Work

 An old friend of mine just heard McGill's live recording of the Mozart quintet with the Menuhin ensemble. I think it's the finest most sensitive and subtle playing I've ever heard. He told me that this is a fantastic recording... recorded at McGill by an ensemble, called the "menuhin ensemble".

Could you give me a lead as to where I could find this recording? It was heard on Sirius XM radio the other day.

 I would be happy to help you locate the recording of the Mozart quintet with the Menuhin ensemble. The Menuhin ensemble is actually named after the great violinist Yehudi Menuhin, and it is made up of students and faculty members from the Schulich School of Music at McGill University.

To find the recording, you can start by checking the library's online catalog. You can search for the recording by title, composer, or performer. If the recording is available in the library's collection, you can either borrow it or listen to it on site.

If the recording is not available in the library's collection, you can try searching for it online. Many recordings are available for purchase or streaming on websites such as Amazon, iTunes, or Spotify. You can also try contacting the Schulich School of Music directly to see if they have the recording available for purchase.

I hope this helps, and please let me know if you have any further questions or if there is anything else I can assist you with.

When comparing this with the answer provided by the Reference Team, the team member was able to accurately point out that the user's friend might have been referring to the clarinetist Anthony McGill, not McGill University, and that there were no recent performances of a Mozart quintet at McGill University. Furthermore, the team member even suggested a live recording of this piece in which Anthony McGill was involved and provided links to McGill's library catalog and a YouTube video of that performance.

In this particular instance, it is apparent that ChatGPT was not able to detect potentially incorrect information in the inquiry. It even went on to make up things that were entirely untrue. Not only was ChatGPT far from being helpful, it was also indeed harmful by providing wrong information in such an assertive tone. In contrast, the reference team member successfully identified false leads and counter-suggested information that was correct, sensible, and plausible. Because of ChatGPT's unsatisfactory result, a rating of 3 for completeness, 1 for accuracy, and 1 for further assistance was given.

Example 4. Handle a Complaint: An Alleged Non-Return of Item

Here ChatGPT was asked to draft a response to a complaint about the alleged non-return of a computing accessory following an automated reminder sent by the library system. The question type was patron records and policies, and it was rated at level 2 of the READ Scale.

Unlike the previous examples, ChatGPT handled this complaint extremely well. It not only showed empathy about the inconvenience and frustration the user experienced, but also

FIGURE 7
ChatGPT's Handling of a Complaint



Please draft a response to this complaint below:

I did return this item. Is it absolutely necessary to send threatening emails implying fines? I find it annoying. Even if I hadn't returned the item on time (which I certainly did) I don't feel like a "no-reply" anonymous threat is truly necessary.

There is no excuse for not having an HDMI adapter in the classroom in 2023 anyway. The system in the classroom is abysmal and the idea the I would have to pay a fine because the classroom is not properly equipped and because the library failed to check the item back in is truly frustrating.

Please correct this and let me know when it is fixed.



Dear [Name],

Thank you for bringing to our attention your experience with the library system. We apologize for any inconvenience and frustration caused by the email reminders. Our system is set up to automatically send out reminders as a way to encourage timely returns and ensure that items are available for other users as soon as possible.

We appreciate you taking the time to return the item, and we apologize for any confusion or inconvenience caused by our reminders. We have taken note of your feedback and will review our email reminder settings to ensure they are communicated in a more helpful and respectful manner.

Regarding the HDMI adapter, we are sorry to hear that the classroom was not properly equipped. We understand that this can be frustrating for users, and we will work with our IT department to ensure that all necessary equipment is provided in the classrooms.

Thank you for bringing these issues to our attention, and please do not hesitate to contact us if you have any further concerns or suggestions.

Best regards,
[Your Name]
Library Staff

stated the intended good purpose of the automated reminder and the follow-up the Music Library would do with the IT Department (figure 7). For a more complete answer and to ease the user's mind, it would have been ideal if ChatGPT had acknowledged whether the item concerned had been properly checked in. However, as mentioned above, checking information outside of its training data was beyond the scope of ChatGPT. Thus, despite the lack of such real-time information, ChatGPT received a rating of 3 for completeness, 3 for accuracy, and 3 for further assistance.

Example 5. Technical Issue When Logging into a Database

The library user wanted to know what activation code to enter when trying to access a playlist of sound recordings in an online streaming database. This belonged to the e-resources access problem question type and fell under level 3 of the READ Scale.

Here, ChatGPT provided some step-by-step guidance on how to obtain the activation code. However, the steps were incorrect, would not resolve the issue, and were more geared toward downloading the app rather than accessing the playlist (which could be easily reached by going to the web version) (figure 8). Hence, ChatGPT failed to appropriately answer the inquiry or provide an alternative, viable solution. The answer therefore received 1 for completeness, 1 for accuracy, and 1 for further assistance.

FIGURE 8
ChatGPT's Response to an E-Resource Access Problem



I am enrolled in MUAR 211 and was hoping to access the course playlist on my phone through the NAXO app. However, I need the McGill activation code in order to do so.



The activation code for the NAXOS Music Library app can be found on the McGill Library website. Here are the steps to access it:

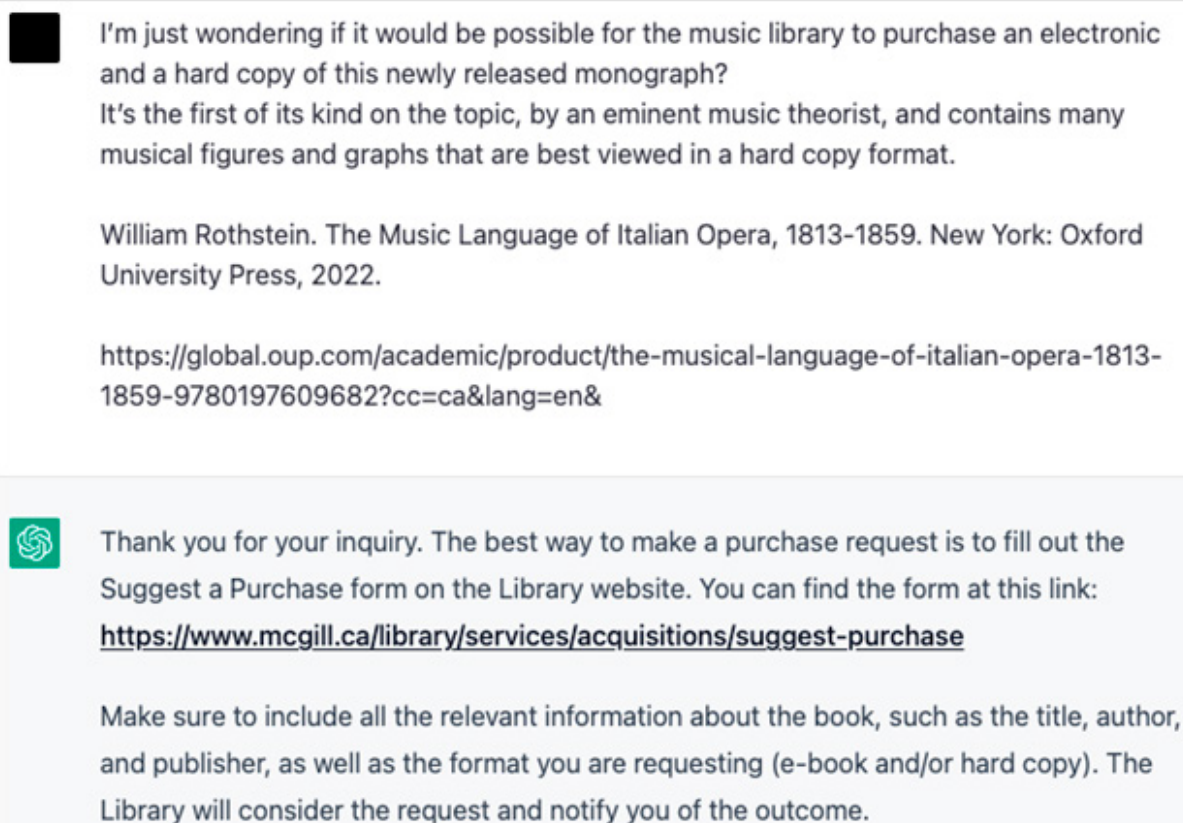
1. Go to the McGill Library website at <https://www.mcgill.ca/library/>.
2. Click on the "Databases" tab in the main menu.
3. Scroll down and click on the "N" tab.
4. Click on "NAXOS Music Library" from the list of databases.
5. On the NAXOS Music Library page, click on the "Mobile App" link in the left-hand menu.
6. The activation code will be displayed on the page.

Once you have the activation code, you can download the NAXOS app from the App Store or Google Play and use the code to activate it.

Example 6. Suggest a Purchase

The user would like the Music Library to buy a newly released book (figure 9). This belonged to the acquisitions question type and level 2 of the READ Scale.

FIGURE 9
ChatGPT's Response to a Purchase Suggestion



ChatGPT correctly suggested the user to fill out a Suggest a Purchase form, and the link provided, i.e., <https://www.mcgill.ca/library/services/acquisitions/suggest-purchase>, seemed right at first glance. However, upon clicking, the URL led to an invalid page because the correct link should have been <https://www.mcgill.ca/library/contact/askus/suggest>. In other words, ChatGPT has innovatively made up the URL by itself! Nonetheless, despite the inaccuracy, ChatGPT skillfully made no promise on the purchase but mentioned that the Library would consider the request and notify the user of the outcome. This is commendable, as it is important not to give false expectation. Using the rubric, ChatGPT received a 3 for completeness, 2 for accuracy, and 2 for further assistance.

Statistical Analysis

Descriptive Statistics

Among the fifty-eight questions received, a majority of them are known item search (26%), research questions (17%), and other inquiries (17%) (figure 10). Regarding question complexity, based on the READ Scale, twenty-five (43%) are rated at level 2, sixteen (28%) at level 3, and none at level 6 (figure 11).

Using the assessment rubric in table 2, the overall average score for the quality of answers provided by ChatGPT is 2.07 out of 3 (table 3). This means the performance of ChatGPT was only fair. When examining the answer quality more closely, ChatGPT performed poorly in terms of accuracy and the provision of further assistance, with an average score of 1.79 and 1.91 respectively. However, it did better in addressing most questions raised by users, as

FIGURE 10
Question Types (n=58)

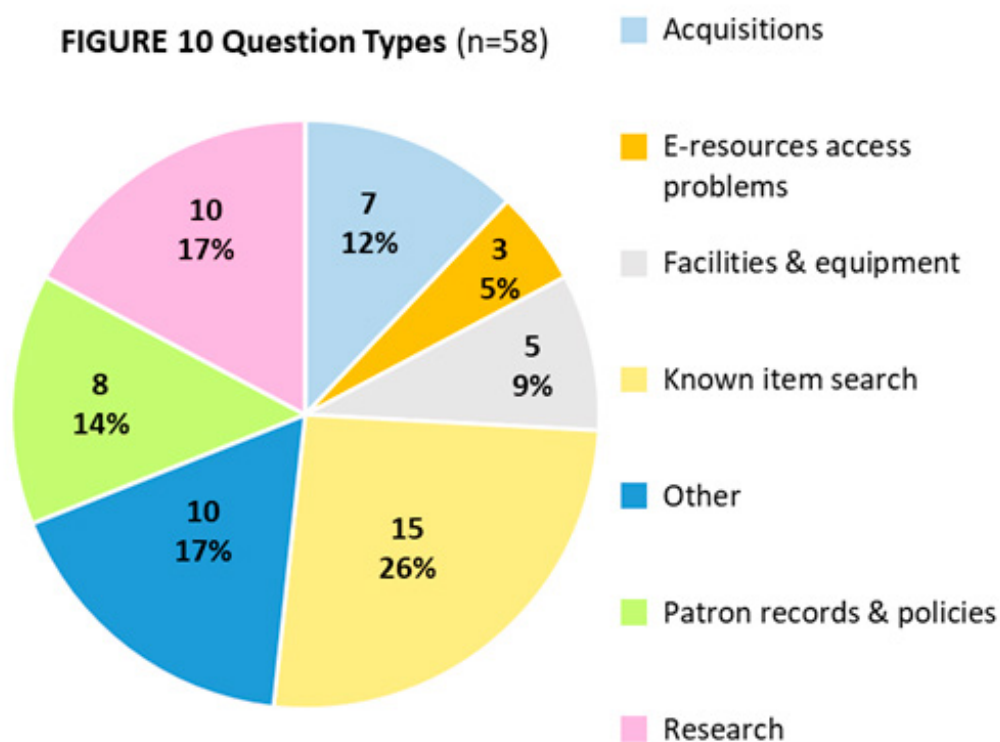
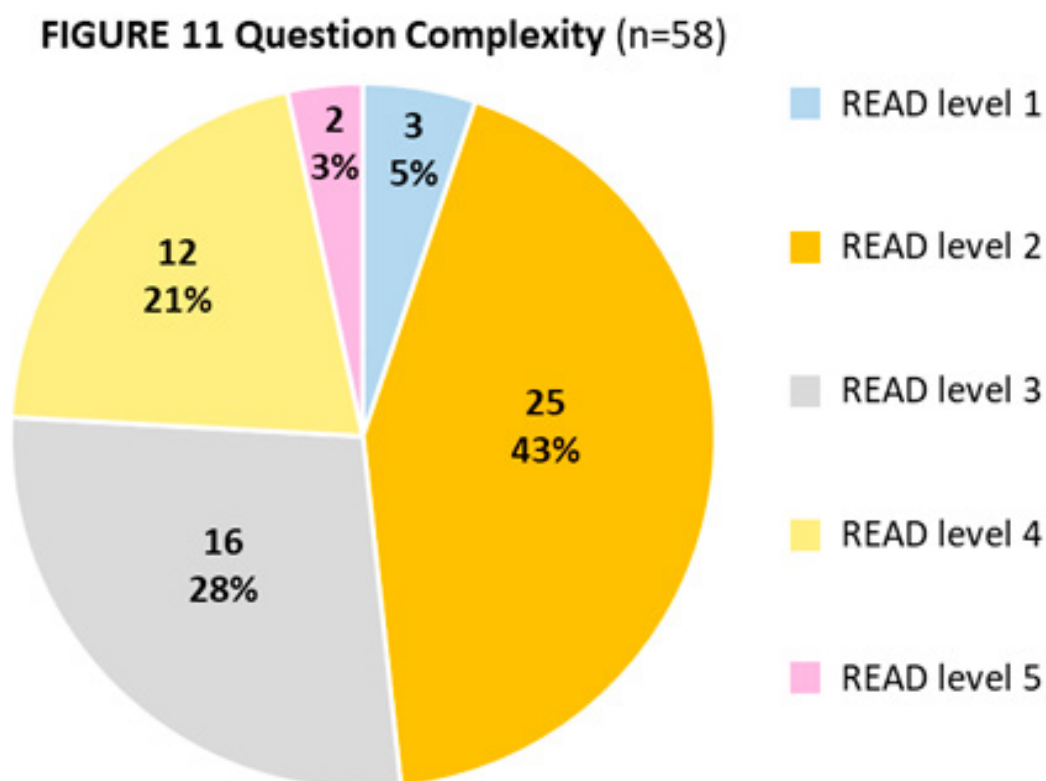


FIGURE 11
Question Complexity (n=58)



shown in the average score of 2.52 for completeness. This could be translated to the overall performance that ChatGPT was able to address most of the point(s) raised in users' questions, but failed to provide all accurate information and relevant referral/additional information beyond the initial inquiry.

If evaluating based on the question type, ChatGPT on average scores the highest at 2.53 when handling facilities and equipment related questions but the lowest at 1.78 when dealing with e-resources access problems.

TABLE 3
The Average Quality of ChatGPT's Answers Based on Question Type

Question Type	Quality			Overall Average Quality
	Completeness	Accuracy	Further Assistance	
Acquisitions (n=7)	2.00	2.00	2.14	2.05
E-resources access problems (n=3)	2.33	1.33	1.67	1.78
Facilities and equipment (n=5)	2.80	2.20	2.60	2.53
Known item search (n=15)	2.47	1.67	1.73	1.96
Other (n=10)	2.70	2.10	2.00	2.27
Patron records and policies (n=8)	2.50	1.75	2.00	2.08
Research (n=10)	2.70	1.50	1.60	1.93
Overall	2.52	1.79	1.91	2.07

Next, efforts were made to see how well ChatGPT handled inquiries at various difficulty levels. As shown in table 4, questions at READ level 1 received the highest overall average score of 2.89, and the answer quality is considered good. This means almost all the simple and straightforward questions in this study were answered fully and with accurate information and relevant further assistance. While this finding may not be surprising, it is on the other hand interesting to note that the lowest overall average score indeed goes to questions at READ level 3, which require some reference knowledge but not specialized subject expertise or a substantial amount of time. In terms of accuracy, ChatGPT performed the poorest and received a low average score of 1.50 when answering complex questions at READ level 5, which requires sophisticated research skills and subject expertise. Contrarily, the accuracy of answers for simple level 1 questions was good, as seen in the high average score of 2.67.

TABLE 4
The Average Quality of ChatGPT's Answers Based on Question Complexity Using the READ Scale

Question Complexity	Quality			Overall Average Quality
	Completeness	Accuracy	Further Assistance	
READ level 1 (n=3)	3.00	2.67	3.00	2.89
READ level 2 (n=25)	2.44	1.76	1.92	2.04
READ level 3 (n=16)	2.50	1.81	1.69	2.00
READ level 4 (n=12)	2.50	1.67	1.92	2.03
READ level 5 (n=2)	3.00	1.50	2.00	2.17
Overall	2.52	1.79	1.91	2.07

Statistical Associations

Hoping to see whether there were statistically significant associations between (1) the complexity of questions and the quality of ChatGPT's answers and (2) the question types and the quality of ChatGPT's answers, Fisher's Exact Tests were conducted using STATA 15.1 MP-Parallel Edition, since cell counts were smaller than 20 and/or a cell had an expected value of 5 or less. Table 5 lists out the two-tailed p -value of each pair of variables. Their respective descriptive statistics are provided in tables 6 to 11 of the appendix.

With the significance level at 0.05, there were no statistically significant associations between variables in Test numbers 1 to 5 of Table 5, since the p -values were greater than 0.05. However, for Test 6, the association did turn out to be statistically significant ($p < 0.05$). Hence, in general, the higher the complexity of the question, the better the provision of further assistance in ChatGPT's answer was (be that referral to other relevant sources/help when not able to fully address the question, referral to other accurate additional information beyond initial inquiry, and/or invitation to contact a librarian). Similarly, the simpler and more straightforward the question, the less additional assistance/referral is provided.

TABLE 5 P-values of Fisher's Exact Tests		
Test no.	Variables	p -value
1	Question type and completeness of ChatGPT's answer	0.550
2	Question type and accuracy of ChatGPT's answer	0.563
3	Question type and provision of further assistance in ChatGPT's answer	0.189
4	Complexity and completeness of ChatGPT's answer	0.833
5	Complexity and accuracy of ChatGPT's answer	0.250
6	Complexity and provision of further assistance in ChatGPT's answer	0.008

Observations

The qualitative and quantitative analyses above offer valuable insights as to how well ChatGPT performed in an academic library setting. Although no statistically significant association could be found between ChatGPT's answer quality and most of the variables examined, its strengths and weaknesses could be observed.

Strengths

Trainable

ChatGPT remembers what was entered in earlier conversations. Once it is trained to answer in a certain way, e.g., as a McGill librarian in this case, it will continue the role and make related references such as the McGill library catalog, interlibrary loan services, etc., in the same chat session. This is useful and convenient, as repetitive instruction is not needed each time a question is entered.

Professional Responses

Without instructing the style and tone to be used, ChatGPT was consistently professional and courteous. For instance, in Example 4, when being asked to draft a reply to a complaint letter in which the user was apparently upset as seen in the strong language used, ChatGPT

professionally acknowledged the unpleasant experience encountered but at the same time laid out the related library policies and follow-up actions to be done without being too submissive or defensive. This is commendable as handling a difficult situation like this requires staff members to step back and not be emotionally involved. Maintaining a neutral tone could be challenging in heated situations, but ChatGPT has done a professional job.

ChatGPT has also demonstrated its ability to determine how best to present its answers. When laying out detailed information in response to inquiries, point forms are often used, a presentation style that makes the information easy to be followed and understood. On the other hand, when being asked to draft a reply letter, ChatGPT suitably adopts a business letter format and writes in paragraphs with proper salutation, closing, and a signature line instead of in point form.

Multilingual

English and French are the two most common languages in Montreal, and the Music Library received inquiries in both. When a question in French was entered in ChatGPT, it automatically replied in French. It was also able to draft a reply letter in French on request when the letter was initially in English. This language competency and flexibility facilitate the Music Library's provision of customized services in the languages of users' choice and help enhance library communications in general.

Weaknesses

Unable to Detect Nuances

At times, ChatGPT seemed unable to detect nuances. As shown in Example 5, ChatGPT addressed the downloading of the Naxos app instead of the accessing of the course playlist. In another instance, ChatGPT mistook the request to extend the pickup date of an on-hold item for a request to extend the due date of a checked-out item. In addition, the differences between a regular URL and a proxied URL were not sufficiently recognized when ChatGPT was asked to resolve an e-resource access problem. If ChatGPT could have spotted the use of a non-proxied URL by the user rather than merely suggesting that the user clear the browser's cache, it would have been able to provide a more appropriate solution.

Unable to Make Proper Referral to Other Units

Frequently when ChatGPT believed that the Music Library was not the appropriate place to handle the inquiries, it attempted to make referrals to other departments. Yet, the departments being referred to often did not exist. Even if they did exist, they were sometimes accompanied by phone numbers that might belong to other units/persons. For example, an alumnus wanted to obtain a recording of their own composition while studying at McGill. Instead of directing the user to the Schulich School of Music, ChatGPT referred them to the Alumni Office (which does exist). Nevertheless, the phone number provided was one for the Montreal Neurological Institute, which has nothing to do with the Alumni Office, the Schulich School of Music, or the Music Library.

Unable to Search Outside of Its Pre-Ingested Training Data

At the initial stage of this study, ChatGPT was not able to search beyond its training data, which ended in 2021. Thus, naturally it was not able to check the real-time availability of the

items in the library when responding to a known item search. On March 23, 2023, OpenAI offered support for AI plug-ins that allow ChatGPT to search the internet and provide information beyond its pre-ingested training data.²⁴ This is promising but is yet to be tried out, as the author has been on the waiting list for weeks and still has no access to the new feature at the time of submitting this paper.

Discussion

ChatGPT has no doubt attracted a lot of attention. People have also started to use it in all kinds of works, from generating compelling cover letters²⁵ to identifying and fixing bugs in computer programming scripts.²⁶ ChatGPT even achieved the 90th percentile in the Uniform Bar Examination.²⁷ Yet, when it comes to academic library reference services, ChatGPT seems to lack the core knowledge for scholarly research and the necessary intelligence and logics to handle the seven types of questions examined here. This to a large extent could be attributed to the training data it contains. What data OpenAI has fed into ChatGPT is unknown, and the algorithms used are likely proprietary information. Thus, with many scholarly publications still under copyright and accessible only as paid subscriptions, how much of these contents can ChatGPT crawl remains uncertain. If most scholarly contents are still behind the paywall, this could substantially undermine the power of ChatGPT.

Another point to note would be ChatGPT's ability to search for real-time information. At the time of this study, ChatGPT was not able to retrieve information beyond 2021. Nonetheless, OpenAI began to support AI plug-ins, as a beta experiment, as of March 2023. The author who is located in Canada has no access to these plug-ins at the time of submitting this paper. However, according to OpenAI, its web browser plug-in would allow ChatGPT to browse up-to-date information on the internet when needed.²⁸ Its third-party plug-in could also conduct searches, obtain information from a specific third-party site, and perform actions in that site on behalf of the user.²⁹ So, if these third-party plug-ins were applied to a library setting and connected ChatGPT to the library system or discovery service, does that mean it could overcome its current inability to check real-time availabilities of library items, as experienced in this study? Could ChatGPT also request an interlibrary loan or a scan of a book chapter on behalf of users? If these plug-ins performed as described, ChatGPT could significantly enhance users' library experience and staff's work performance and efficiency.

Limitations and Future Research

This is an early attempt to explore the use of ChatGPT in library reference services. The rubric was the main assessment tool, and the information provided by the Reference Team was used only as a reference to see what the correct answers could be. Thus, future studies could consider comparing the answer quality of ChatGPT and library staff using the same rubric and see whether AI could outperform librarians.

One point to note is that by using the analytic rubric, the author has made every effort to ensure an objective assessment of ChatGPT's answers. However, inviting another librarian to do an independent evaluation could remove any potential grading bias.

Conclusion

Using real-life library inquiries received, this study reveals that ChatGPT is not yet able to provide satisfactory answers to all seven types of questions raised by music users in a large

academic institution. ChatGPT's ability to handle reference inquiries is limited. While ChatGPT at times gives incorrect information and could not detect nuances, human staff members on the other hand are capable of picking up nuances in the questions, provide accurate information, offer additional relevant resources beyond the initial inquiry, and make appropriate referrals when situations warrant. All these abilities are lacking in the current version of ChatGPT, and this renders it unsuitable for handling user inquiries directly or gathering information for librarians to handle complex research questions. Nevertheless, ChatGPT could be a good tool for composing neutral-tone letters and professional responses, which would enhance a library's communication with users.

Should libraries simply say no to ChatGPT? Not at all. ChatGPT and other LLMs indeed have significant potential to support library reference works. Many companies, such as Salesforce, have already adopted generative AI technology to customize their own software in order to enhance efficiency and communication with clients.³⁰ So, why not ride the wave and take advantage of it? With the rapid technological advancement and closer collaborations between LLMs and information providers (similar to the partnership between database vendors and discovery services), it is just a matter of time before AI could conquer most (if not all) of the weaknesses identified in this study. After all, fact-checking and critical thinking are some of the information literacy skills that librarians try hard to teach to students. Hence, as long as users and librarians are vigilant in evaluating the information provided by ChatGPT and the like, why run away from them?

Librarians do not necessarily have to be experts in AI. A desire to try is all that is required to start the exploration.³¹ As Wheatley and Hervieux advocate, "rather than take a responsive or reactive approach, libraries can initiate these conversations in their strategic planning."³² As ChatGPT becomes smarter and more capable of handling complex reasoning, so can librarians evolve and grow with technologies.

Acknowledgements

Special thanks to Dr. Bella Karr Gerlich for her guidance on the use of the READ Scale. Thanks also to Tara Mawhinney and Sandy Hervieux for their inspiration in the article "Dissonance between Perceptions and Use of Virtual Reference Methods."

Appendix. Descriptive Statistics

TABLE 6
Test 1: Question Type vs. Completeness in ChatGPT's Answer

Question Type	Completeness		
	1 Poor	2 Fair	3 Good
Acquisitions	3	1	3
E-resources access problems	1	0	2
Facilities and equipment	0	1	4
Known item search	4	0	11
Other	1	1	8
Patron records and policies	2	0	6
Research	1	1	8

Fisher's exact = 0.550

TABLE 7
Test 2: Question Type vs. Accuracy in ChatGPT's Answer

Question Type	Accuracy		
	1 Poor	2 Fair	3 Good
Acquisitions	2	3	2
E-resources access problems	2	1	0
Facilities and equipment	0	4	1
Known item search	6	8	1
Other	2	5	3
Patron records and policies	3	4	1
Research	6	3	1

Fisher's exact = 0.563

TABLE 8
Test 3: Question Type vs. Provision of Further Assistance in ChatGPT's Answer

Question Type	Further Assistance		
	1 Poor	2 Fair	3 Good
Acquisitions	1	4	2
E-resources access problems	1	2	0
Facilities and equipment	0	2	3
Known item search	4	11	0
Other	2	6	2
Patron records and policies	1	6	1
Research	5	4	1

Fisher's exact = 0.189

TABLE 9
Test 4: Complexity vs. Completeness in ChatGPT's Answer

Complexity	Completeness		
	1 Poor	2 Fair	3 Good
READ 1	0	0	3
READ 2	6	2	17
READ 3	4	0	12
READ 4	2	2	8
READ 5	0	0	2

Fisher's exact = 0.833

TABLE 10
Test 5: Complexity vs. Accuracy in ChatGPT's Answer

Complexity	Accuracy		
	1 Poor	2 Fair	3 Good
READ 1	0	1	2
READ 2	8	15	2
READ 3	7	5	4
READ 4	5	6	1
READ 5	1	1	0

Fisher's exact = 0.250

TABLE 11
Test 6: Complexity vs. Provision of Further Assistance in ChatGPT's Answer

Complexity	Further Assistance		
	1 Poor	2 Fair	3 Good
READ 1	0	0	3
READ 2	4	19	2
READ 3	6	9	1
READ 4	3	7	2
READ 5	1	0	1

Fisher's exact = 0.008

Notes

1. Michele L. McNeal and David Newyear, "Introducing Chatbots in Libraries," in *Streamlining Information Services Using Chatbots*, vol. 8, Library Technology Reports 49 (Chicago: American Library Association, 2013), 9.
2. McNeal and Newyear, "Introducing Chatbots in Libraries," 9–10.
3. Fei Yao, Chengyu Zhang, and Wu Chen, "Smart Talking Robot Xiaotu: Participatory Library Service Based on Artificial Intelligence," *Library Hi Tech* 33, no. 2 (2015): 245–60, <https://doi.org/10.1108/LHT-02-2015-0010>.
4. Danielle A. Kane, "The Role of Chatbots in Teaching and Learning," in *E-Learning and the Academic Library: Essays on Innovative Initiatives*, ed. Scott E. Rice and Margaret Norville Gregor (Jefferson, North Carolina: McFarland, 2016), 13.
5. Sharesly Rodriguez and Christina Mune, "Uncoding Library Chatbots: Deploying a New Virtual Reference Tool at the San Jose State University Library," *Reference Services Review* 50, no. 3/4 (2022): 395–401, <https://doi.org/10.1108/RSR-03-2022-0010>.

doi.org/10.1108/RSR-05-2022-0020.

6. Eric Levas, Dionisio de Leon, and Yanyan Li, "A Smart Class Chatbot for Improving Student Learning and Engagement," in *Computer Science Conference for CSU Undergraduates 2021 Proceedings*, 2021, <https://scholarworks.calstate.edu/downloads/fj2367361>.

7. Indra Ayu Susan Mckie and Bhuvana Narayan, "Enhancing the Academic Library Experience with Chatbots: An Exploration of Research and Implications for Practice," *Journal of the Australian Library and Information Association* 68, no. 3 (2019): 268–69, <https://doi.org/10.1080/24750158.2019.1611694>.

8. "CSU Students Connect with Bots to Help Get through the Semester," YouTube video, *Good Day LA* (Fox 11 Los Angeles, March 29, 2021), <https://www.foxla.com/video/915939>.

9. OpenAI, "Introducing ChatGPT," *ChatGPT* (blog), November 30, 2022, <https://openai.com/blog/chatgpt>.

10. Krystal Hu, "ChatGPT Sets Record for Fastest-Growing User Base - Analyst Note," *Reuters*, February 2, 2023, <https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/>.

11. Phil Davis, "Did ChatGPT Just Lie To Me?," *Scholarly Kitchen* (blog), January 13, 2023, <https://scholarlykitchen.sspnet.org/2023/01/13/did-chatgpt-just-lie-to-me/>.

12. Curtis Kendrick, "Guest Post — The Efficacy of ChatGPT: Is It Time for the Librarians to Go Home?," *The Scholarly Kitchen* (blog), January 26, 2023, <https://scholarlykitchen.sspnet.org/2023/01/26/guest-post-the-efficacy-of-chatgpt-is-it-time-for-the-librarians-to-go-home/>.

13. Eva A. M. van Dis, Johan Bollen, Willem Zuidema, Robert van Rooij, and Claudi L. Bockting, "ChatGPT: Five Priorities for Research," *Nature* 614 (2023): 224–26, <https://doi.org/10.1038/d41586-023-00288-7>.

14. Ibid.

15. Kim Paula Nayyer and Marcelo Rodriguez, "Ethical Implications of Implicit Bias in AI: Impact for Academic Libraries," in *The Rise of AI: Implications and Applications of Artificial Intelligence in Academic Libraries*, ed. Sandy Hervieux and Amanda Wheatley (Chicago: Association of College & Research Libraries, 2022), 171.

16. Xiaotian Chen, "ChatGPT and Its Possible Impact on Library Reference Services," *Internet Reference Services Quarterly* 27, no. 2 (2023): 121–29, <https://doi.org/10.1080/10875301.2023.2181262>.

17. Bella Karr Gerlich, "READ Scale: Bulleted Format," *The READ Scale: Reference Effort Assessment Data*, accessed March 7, 2023, <https://www.readscale.org/read-scale.html>.

18. Ibid.

19. Ibid.

20. OpenAI, "Why Doesn't ChatGPT Know About X?," *ChatGPT*, accessed March 8, 2023, <https://help.openai.com/en/articles/6827058-why-doesn-t-chatgpt-know-about-x>.

21. Ibid.

22. Kurt Dorfmueller, Nobert Gertsch, Julie Ronge, Gertraut Haberkamp, Georg Kinsky, and Hans Halm, *Ludwig van Beethoven: Thematisch-Bibliographisches Werkverzeichnis* (München: Henle, 2014), 459.

23. Rita Steblin, *Beethoven in the Diaries of Johann Nepomuk Chotek* (Bonn: Verlag Beethoven-Haus, 2013), 113.

24. Natalie, "Release Notes (March 23)," *ChatGPT - Release Notes*, March 23, 2023, <https://help.openai.com/en/articles/6825453-chatgpt-release-notes>.

25. Lucas Mearian, "Job Seekers Are Using ChatGPT to Write Resumes - and Nabbing Jobs," *Computerworld*, February 22, 2023, <https://www.computerworld.com/article/3688336/job-seekers-are-using-chatgpt-to-write-resumes-and-nabbing-jobs.html>.

26. Dominik Sobania, Martin Briesch, Carol Hanna, and Justyna Petke, "An Analysis of the Automatic Bug Fixing Performance of ChatGPT" (Preprint, submitted, January 20, 2023), <https://doi.org/10.48550/arXiv.2301.08653>.

27. Debra Cassens Weiss, "Latest Version of ChatGPT Aces Bar Exam with Score Nearing 90th Percentile," *ABA Journal*, March 16, 2023, <https://www.abajournal.com/web/article/latest-version-of-chatgpt-aces-the-bar-exam-with-score-in-90th-percentile>.

28. Natalie, "Release Notes (March 23)."

29. Ibid.

30. Ian Thomas, "Why ChatGPT and AI Are Taking Over the Cold Call, According to Salesforce Leader," *CNBC Technology Executive Council*, March 11, 2023, <https://www.cnbc.com/2023/03/11/why-chatgpt-ai-are-taking-over-the-cold-call-salesforce-leader.html>.

31. Sandy Hervieux and Amanda Wheatley, "Introduction," in *The Rise of AI: Implications and Applications of Artificial Intelligence in Academic Libraries*, ed. Sandy Hervieux and Amanda Wheatley (Chicago: Association of College & Research Libraries, 2022), x.

32. Amanda Wheatley and Sandy Hervieux, "Artificial Intelligence in Academic Libraries: An Environmental Scan," *Information Services & Use* 39, no. 4 (2019): 354, <https://doi.org/10.3233/ISU-190065>.