ARTICLE

AI-Infused Discovery Environments

Information Retrieval Boon or Overpromised Hype?

Blake L. Galbreath, Erica England, Corey M. Johnson, and Jen Saulnier Lange

ABSTRACT

Although still in its infancy, artificial intelligence (AI) is rapidly making inroads into most facets of the library and education spheres. This paper outlines steps taken to examine Primo Research Assistant, an AI-infused discovery environment, for potential deployment at a large US public research university. The researchers aimed to evaluate the quality and relevance of the AI results in comparison to sources retrieved from the conventional search functionality, as well as the AI system's multi-paragraph overview reply to the search query. As a starting point, the authors collected 103 search strings from a Primo Zero Result Searches report to approximate a corpus of natural language search queries. For the same research area, it was discovered that there was only limited overlap between the titles returned by the AI tool versus the current discovery layer. The researchers did not find appreciable differences in the numbers of topic-relevant sources between the AI and non-AI search products (Yes = 46.3% vs. Yes = 45.6%, respectively). The overview summary is largely helpful in terms of learning more details about the recommended sources, but it also sometimes misrepresents connections between the sources and the research topic. Given the overall conclusion that the AI system did not constitute a clear advancement or decline in effective information retrieval, the authors will turn to usability testing to aid them in further implementation decisions.

INTRODUCTION

Generative artificial intelligence (AI) research tools span a wide range of functions and applications across disciplines that aid scholars throughout the research lifecycle. These tools assist in the early stages of research by formulating research questions, organizing ideas, and helping scholars quickly digest complex academic texts by summarizing research papers and extracting key concepts. In addition to supporting discovery, AI research tools are also used to facilitate language and grammar, enhance clarity and coherence, and adhere to disciplinary conventions in scholarly writing; these tools support the drafting, revising, and refining of academic writing, helping scholars produce more polished academic works. Furthermore, AI-enhanced citation and reference management tools simplify citation generation and integration, as well as the organization of sources while maintaining academic integrity. The scope of this research is focused on the AI-enhanced discovery tools that help researchers find relevant sources and synthesize existing academic literature more efficiently by reducing the time spent on existing conventional search practices.

One such AI-enhanced discovery tool is Ex Libris' Primo Research Assistant (PRA), which was released in beta in September 2024.[1] According to the vendor's documentation upon release,

About the Authors

Blake L. Galbreath (blake.galbreath@wsu.edu; corresponding author) is Assistant Head of Systems and Technical Operations, Washington State University. **Erica England** (erica.england@wsu.edu) is First-Year Experience Librarian, Washington State University. **Corey M. Johnson** (coreyj@wsu.edu) is Instruction and Assessment Librarian, Washington State University. **Jen Saulnier Lange** (jennifer.saulnier@wsu.edu) is Online Learning Librarian, Washington State University.

Submitted: 9 June 2025. Accepted for Publication: 12 September 2025. Published: 15 December 2025.

https://doi.org/10.5860/ital.v44i4.17465

> Primo Research Assistant is a generative AI-powered tool designed to streamline time-intensive tasks. It enables users to query academic content in natural language and utilizes the breadth of your library to pinpoint five articles that can aid in answering your question. The tool distills the most pertinent information from the descriptions/abstracts of each article to craft the response's overview that includes in-line references to the sources to provide transparency in how each source contributed to the response. Beneath the overview, these sources and additional sources are available for further exploration of the subject and for verifying the tool's responses.[2]

Specifically, PRA uses a retrieval augmentation generation (RAG) architecture that leverages the OpenAI GPT 3.5 large language model (LLM).[3] Details of the RAG process in PRA are summarized as follows:

1. The LLM converts the user search into ten variant strings, plus the original search query.[4] Example: User query "what is the effect of vitamin D deficiency?" is converted to (vitamin D deficiency effects) OR (impact of vitamin D deficiency) OR (consequences of low vitamin D) OR ((vitamin D deficiency) OR (hypovitaminosis D outcomes)) OR ((vitamin D deficiency) AND (health effects)) OR ("vitamin D deficiency effects") OR (vitamin D insufficiency health outcomes) OR (((vitamin D deficiency) OR (hypovitaminosis D) impact)) OR ((vitamin D deficiency) AND (physiological consequences)) OR (vitamin D) OR (what is the effect of vitamin D deficiency).
2. The converted query is searched against the CDI to find the top thirty matching records.[5]
3. The user query and the title and abstract of the top thirty search results are converted into numeric representations and then mapped together to identify the five that best fit the user query.[6]
4. The original user query, top five results, and instructions are sent to the LLM.[7]
5. The LLM creates an overview (answer) in the user's language with one to five inline references.[8]

AI tools utilizing a RAG model include Scite, Elicit, Consensus, and Scopus AI and are becoming increasingly popular within academia.[9]

Unlike purely generative models, RAG models both retrieve relevant documents from an external knowledge source *and* process information to generate human-friendly text.[10] This reliance on an external knowledge source helps ground the LLM-generated answer and thereby reduces the occurrence of hallucinations.[11] The term *hallucination* has many definitions today, from simply "false statements" to "bullshit" to a response to a "question (or prompt) with text that seems like a plausible answer, but is factually incorrect or irrelevant."[12] In this case, PRA retrieves documents from the Central Discovery Index, with the caveat that newspaper content and documents from certain providers are omitted.[13] The addition of such an AI-enhanced tool to the discovery layer would represent a significant change to the services the university has been offering library patrons.

The primary objective of this study is to evaluate the potential benefits of PRA beta for users conducting research at Washington State University, a large R1 land-grant university. To achieve this objective, the researchers address two central questions: First, they explore how the relevance of AI-generated PRA results compares to that of librarian-created search queries within the context of this research study. Second, they examine the usefulness of the overview statements produced by PRA.

LITERATURE REVIEW

Since the public launch of ChatGPT in November 2022, there has been immeasurable discussion regarding AI in education. Proponents argue that AI can streamline pedagogical tasks, such as test design, grading, lesson planning, and communication with parents.[14] Additionally, AI can personalize student learning by creating adaptive questions.[15] It also has the ability to create instantaneous translations for English as a Second Language (ESL) students, construct reading-level summaries, enhance writing support, and formulate accurate citations.[16] In short, proponents argue that AI will positively revolutionize the formal educational experience.

Critics of AI in education are equally numerous and vocal. One central critique of AI is that it creates academic integrity violations as students outsource their writing to it, since AI tools are capable of creating informative, coherent, and mostly high-quality writing.[17] Although a persuasive argument can be made that the use of AI is not intellectual theft from other scholars per se, it can still be classified as "unauthorized collaboration," resulting in a cheating infraction.[18]

Beyond concerns about AI in writing, there are other limitations. Paywall models risk deepening educational inequalities, and AI cannot understand value questions or emotions and does not know about very current events.[19] Its outputs are constrained by training data, limiting the context or intent of research and the ability to understand when it unwittingly perpetuates biases or discrimination.[20] Additionally, AI systems consume vast energy resources, contributing to climate change.[21] There are significant and pending legal questions about whether AI practices infringe on data privacy and violate intellectual property laws.[22] Finally, many fear a future overreliance on AI, which may weaken students' ability to perform deep critical thinking and close reading.

Despite extensive speculation about how AI will affect education, relatively little has been said specifically about how AI might impact the source acquisition component of the research enterprise. For example, Cotton states that AI could provide recommendations for resources but does not explain how it would work.[23] The most frequently cited concern in the literature is source hallucinations in AI results lists.[24]

Conversely, AI offers possible improvements to current discovery layer systems through superior interpretation of natural language searches. Patrons more easily express queries in natural language rather than through Boolean search queries. Furthermore, conventional information retrieval systems are not able to parse out natural language searches as well as AI systems, and natural language searching outperforms simple keyword searching, thus yielding results more closely aligned with user intent.

Additionally, AI systems can make stronger connections among articles and within individual articles as well. AI develops citation trails to create lists of relevant sources for researchers.[25] Some AI systems like Scite classify citations as supporting, contrasting, or merely mentioning the cited work.[26] Text-extraction tools, such as Research Rabbit, can extract specific parts of articles such as citations, datasets, summaries, abstracts, research findings, and methodologies while identifying common themes and conflicting viewpoints.[27]

One area of great promise for AI-infused literature searching concerns the personalization of the information retrieval process. AI systems can effectively use browsing history, user-prescribed preferences, and data from similar users to customize results and improve resource relevance.[28] AI is adept at converting discipline-specific jargon into more accessible language, which is an

important key to learning.[29] As Mitcham writes, "Everything taught at 5th grade science [class] today, has its roots in a journal article of the past."[30] Beyond sophistication levels within researchers' native languages, AI stands ready to efficiently translate across world languages and between text-based and audio/visual content.[31] Hyde et al. call language translation one of the seven key roles of AI, as information format conversion is one of the capabilities of AI, which can be accomplished at lightning-fast speeds.[32]

Although little has been published on PRA to date, what does exist suggests a positive review. One column review notes that the results returned by the PRA "View more results" functionality (a function that expands the user's search string by augmenting it with synonyms and Boolean OR structures) can return more relevant results than the same search string in Primo VE (PVE).[33] However, the study is mostly demonstrative in nature and very limited in size, citing only two examples of search comparisons. A second study invites staff and faculty to interact with PRA and provide feedback via a survey.[34] Users are instructed to ask PRA three to five questions on a subject in which they are knowledgeable to "assess the accuracy and quality of the answers provided."[35] From this survey, the majority of respondents find the PRA response "somewhat or highly accurate" and "somewhat or highly relevant."[36] Another recent study compares PRA against two other popular AI add-ons: Web of Science Research Assistant (WoSRA) and Scopus AI (SAI).[37] The review compares qualities such as the interpretability and reproducibility of searches, the Boolean search strategies employed, the ability to find relevant papers on a subject or correctly refuse to answer if no relevant papers exist, and the quality and scope of documents that are available to the RAG process.[38] With regard to PRA, the author finds that it provides appropriate answers and citations where the retrieval challenge is *Easy* and *Moderate*, but it struggles where the retrieval challenge is *High* and does not refuse to answer a question for which there are no relevant documents.[39] The study utilizes a small sample size of test questions (four questions), but more importantly, it runs each query three times (across the three products) and finds that "while RAG answers varied slightly, their overall performance or failure remained constant."[40]

This paper contributes to the literature by adding a large-scale comparative analysis of the relevancy of results returned by natural language queries (NLQ) within PRA versus those returned by expert Boolean searches within the discovery layer. To a lesser extent, this study also evaluates the usefulness of PRA overview statements. In the context of this study, *relevancy* is defined as the degree to which a retrieved item appropriately aligns with a user's query or information needs. The judges of relevancy in this study are instruction librarians, who are professionals in creating keyword Boolean searches but are not subject-matter experts.

METHODS

The team of researchers consisted of one systems librarian and three instruction librarians. The systems librarian was responsible for creating the corpus of search queries and recording multiple data points. The instruction librarians were responsible for developing expert Boolean searches and for evaluating the relevancy of the search results from both PRA and expert-created Boolean queries; they are later referred to as "raters" in the paper.

Research Questions

The overall objective of this study is to determine if *PRA is beneficial to users performing research in its current (beta) version* at a large state university in the Pacific Northwest. Specifically, the researchers were guided by the following questions:

- How does the relevance of AI-generated PRA results compare to the relevance of results from librarian-created search queries?
- Is the overview statement generated by the PRA useful?

Data Collection

To approximate a corpus of natural language search queries to test PRA against librarian-created search queries, the researchers created a Primo Analytics report within the Primo Zero Result Searches (PZRS) subject area from January 1 to October 10, 2024. The PZRS table reports daily on "the top 100 basic and advanced searches that had zero results."[41] Many of the search strings that are included in this report appear here, due to user error. Strings of this type might include egregious typos or nonexistent ISBN/ISSN identifiers, or searches on extremely narrow topics. Other search strings are included in this report simply because they are written in natural language syntax, and PVE does not handle such syntax well. In other words, this latter string-type does not necessarily imply that the queries are difficult (although some of them might be), but rather they have a syntax that is not compatible with PVE.

This original report consisted of 25,477 rows (including empty strings), which was then filtered to search strings containing a question mark. Six hundred and ninety-six search strings with question marks were exported and manually reviewed to remove those that did not appear to be natural language search queries. During the first pass, the researchers winnowed the list down to 138 search strings by removing exact duplicate entries, known items, and other unrelated items, such as website addresses and instances of a question mark used as a wildcard or diacritic.

These 138 search strings were then cleaned and normed. For example, double quotation marks were removed, extraneous characters were removed from the beginning of strings, and spelling errors were corrected. At this point, the search strings were referred to as NLQs. The researchers then ran the 138 NLQs through PRA and recorded the following data points for each string: the number of citations returned, the overview summary, and the actual citations returned, as shown in Table 1.

Table 1. Example of a natural language query and its associated summary and citations.

Natural Language Query	Number of Citations	Overview Summary	Citations
How are college athletes affected by policies and punishments related to marijuana use?	5	The policies and punishments related to marijuana use significantly impact college athletes, particularly in light of evolving regulations and societal attitudes towards cannabis. As the NCAA modifies its THC-testing policies and more states legalize recreational marijuana, college athletics administrators face challenges in adapting their drug-testing policies. This shift creates a complex environment for intercollegiate athletics, where administrators must balance compliance with legal standards and the health and well-being of student-athletes [1]. Research indicates that a notable minority of collegiate athletes report past-year marijuana use, despite the potential negative consequences on performance, eligibility, and scholarships. Factors influencing marijuana use among athletes include sensation-seeking behavior, negative mood, and perceptions of peer use, which are similar to those observed in nonathletes. However, athletes may be particularly motivated to use marijuana for enhancement purposes, suggesting that their risk profiles differ from nonathletes [3][4]. Moreover, the consequences of marijuana use can vary based on athlete characteristics, such as gender and competition season status. For instance, male athletes who continue using marijuana during their competitive season exhibit different motivations and risk profiles compared to those who abstain [4]. The	1. Consider new dynamics in student-athletes' marijuana use Neal, Timothy. 2022. Abstract: Between the NCAA modifying its THC-testing policies and more states legalizing the recreational use of marijuana, intercollegiate athletics administrators are struggling to make appropriate adjustments in their departmental drug-testing policies… 2. An Examination of Athlete Status as a Moderator of Multiple Alcohol and Marijuana Use Relationships in College Reynolds, Linda M. 2021. Abstract: Objective: To test if intercollegiate athlete status moderates various alcohol and marijuana use relationships in a large college student sample… 3. Risk of Marijuana Use in Male and Female College Student Athletes and Nonathletes Buckman, Jennifer F, Farris, Samantha G, Yusko, David A, White, Helene R, Pandina, Robert J. 2011. Abstract: A large minority of collegiate athletes report past-year marijuana use even though there is a significant possibility of experiencing negative athlete-specific consequences related to performance, eligibility, and scholarship… 4. Sport-Related and Psychosocial Factors Associated With Motives and Consequences Of Alcohol and Cannabis Use Among NCAA Athletes: A Systematic Review

Natural Language Query	Number of Citations	Overview Summary	Citations
		implications of these findings highlight the need for tailored approaches in addressing substance use among athletes, considering their unique circumstances and the pressures they face [2][3]. In summary, college athletes are affected by marijuana-related policies and punishments through the interplay of evolving regulations, personal motivations for use, and the potential consequences on their athletic careers. The literature suggests a need for further research to understand the nuances of these dynamics and to inform policy development in collegiate athletics.	Zale, Emily L, Mastroleo, Nadine R, Shaw, Michael T, Williams, Callon M. 2022. Abstract: To review differences in alcohol- and cannabis-related motives and consequences among National Collegiate Athletic Association (NCAA) athletes as a function of athlete characteristics (e.g. gender and competition season status)… 5. Synthetic Cannabinoid Use and Descriptive Norms among Collegiate Student-Athletes Milroy, Jeffrey J., Erausquin, Jennifer Toller, Egan, Kathleen L., Wyrick, David L.. 2016. Abstract: Synthetic cannabinoids have gained popularity over the past decade, especially among young adults, due to sharing similar psychoactive properties with Tetrahydrocannabinol (THC)…

Expert Searches Rule Set

The three instructional librarians then created expert-level Boolean searches from these NLQs, based on their experiences teaching the same subject matter in classrooms, limiting themselves to the following workflow:

1. Create expert searches from the patron-provided NLQs.
2. Use any words from the NLQ and/or the AI search (long string of synonyms with Boolean ORs) to create their expert search queries.
3. Use Boolean logic, symbol limiters (e.g., ""), and/or expanders (e.g., *) and parentheses for grouping terms, but no other scoping, material type, local availability, or post-search facets.
4. Create up to three searches, performing a second and third search as needed to get a search that provides the best quality results set.
5. The results set should have five or more relevant results on the first page or two and mostly be academic materials (books and journal articles).
6. Availability of the sources (electronic, print, interlibrary loan required) is not used as part of the assessment of quality (because PRA beta does not offer availability information).

During their analysis, the instruction librarians removed an additional thirty items from the corpus: duplicates, known items, those in non-English languages, those where PRA returned zero citations, and those where there were difficulties understanding the meaning of the NLQ.

Once the expert search queries were constructed, the researchers ran the remaining 108 queries in PVE. The instruction librarians allowed themselves a total of two possible revisions to the initial expert query attempt with the aim of producing the “best” set of high-quality results. Results of the final expert query were recorded, including the total number of results, as well as titles of the first ten search results returned (see Table 2).

Table 2. Example of an expert search query and its associated first ten Primo VE (PVE) search results.

Expert Search Query	Number of Results	First Ten PVE Search Results
(college athletes) AND (marijuana policy)	134	1. Socially Constructing Marijuana Policy and Target Populations in the News Media 2. College drinking and drug use 3. Consider new dynamics in student-athletes' marijuana use 4. An Examination of Athlete Status as a Moderator of Multiple Alcohol and Marijuana Use Relationships in College 5. Marijuana Goes Mainstream At South By Southwest (SXSW(R)) As Cannabis Industry Leaders, Athletes, Media And Politicians Gather To Discuss Business, Investing, Medicine, Legalization, Politics, Sex, Sports And Startups: SXSW(R) To Feature Leading Cannabis Companies Including 420Games, Canna Advisors, Foria, Gateway, Harborside, Headset, HelloMD, Leafly, Marijuana Policy Project, Marley Natural, New Frontier Data, Strainz and VapeXHale 6. Synthetic Cannabinoid Use and Descriptive Norms among Collegiate Student-Athletes 7. Teenagers and substance abuse 8. College Athletes and Alcohol and Other Drug Use. Infofacts/Resources

		9. The Business Buds: How the Booming Marijuana Industry is Potting Seeds in Professional Sports Enterprises. 10. Cannabis and Exercise Science: A Commentary on Existing Studies and Suggestions for Future Directions

Title Matching

The researchers compared the first ten results from PVE against the five results from PRA, using Excel Fuzzy Lookup, as shown in Table 3.

Table 3. Example of overlapping citations returned by natural language query and expert search query.

Search Queries	Overlapping Citations	Number of Overlapping Citations
How are college athletes affected by policies and punishments related to marijuana use? (college athletes) AND (marijuana policy)	1. Consider new dynamics in student-athletes' marijuana use 2. An Examination of Athlete Status as a Moderator of Multiple Alcohol and Marijuana Use Relationships in College 3. Synthetic Cannabinoid Use and Descriptive Norms among Collegiate Student-Athletes	3

Determining the Relevancy of Primo Research Assistant Citations

As a normative exercise to assess the relevancy of the PRA citations, the raters individually scored the PRA citations from the first five NLQs. They scored the citations as being relevant with values of *Yes*, *Maybe*, or *No* and then convened to discuss their cumulative results. From this normative discussion, the raters devised a rubric for scoring the remainder of the corpus (see Table 4).

Table 4. Relevancy scoring rubric.

Yes	Maybe	No
• Fully addresses main topic • Largely relevant to most facets of the topic • Recent based on subject • Insights/conclusions based on scientific study(ies) and/or systematic analysis	• Might have some useful information, but would have to dig deeper • Unsure of relevance based on knowledge of topic—need more information • Older publication date, but culturally significant/significant to topic	• Not relevant to main topic • Too focused on a very specific part of the topic or tangential topic • Outdated based on subject • Insights/conclusions based largely on author opinion and/or anecdotal experiences

The five NLQs that were used in the norming process were removed from the corpus and the raters scored the remaining cases, which consisted of 103 NLQs and 499 citations. (Five hundred and three citations were originally recorded, but four of them could not be retrieved when the raters performed their analysis.)

Determining the Relevancy of Primo VE Search Results

To make comparisons against the PRA scores, the raters also scored the relevancy of the first ten results returned from the PVE results. The researchers decided to score the first ten results, compared to only five from PRA, because this represents the default number of results patrons normally encounter on the first page of results.

Assessing Usefulness of Overview Summaries

As an avenue to learn more about these comparative results, the researchers examined more closely the five overview statements of the NLQs where all five of the PRA results were scored as *No*. The authors analyzed how the overview summaries might impact how patrons would view the relevance and potential usefulness of the sources (see Overview Summary Analysis in the Results section).

RESULTS

Overlapping Citations

In this section, the researchers report on the overlap of citations returned by PRA and records returned by Primo VE. In total, twenty-four of the 103 (23%) natural language queries (NLQ) returned at least one overlapping citation between PRA and PVE. As shown in Table 5, fifteen NLQs returned exactly one, six NLQs returned two, and three NLQs returned three overlapping citations, bringing the total number of overlapping citations to thirty-six.

Table 5. Overlapping citations between Primo Research Assistant and Primo VE.

Frequency of Overlapping Citations	Number of Natural Language Queries	Total Number of Overlapping Citations
1	15	15
2	6	12
3	3	9
Total	24	36

This means that of the 499 citations returned by all of the NLQs, thirty-six (7.21%) were successful in matching records returned by the expert search queries.

Relevancy Ratings

The following is a breakdown of the relevancy ratings for the citations returned by PRA and the records returned by PVE.

Table 6. Aggregate relevancy rating between Primo Research Assistant (PRA) and Primo VE (PVE).

	Yes	Maybe	No	Total
PRA Total	231	130	138	499
PRA Percentage	46.3%	26.0%	27.7%	100%
PVE Total	448	204	330	982
PVE Percentage	45.6%	20.8%	33.6%	100%

Overview Summary Analysis

The researchers examined more closely the five overview statements of the NLQs where all five of the PRA results were deemed irrelevant. The overview summary analysis comprised two basic questions, the answers to which are summarized in Table 7:

1. Does the overview summary try to answer the questions or simply summarize the sources?
2. Does the overview summary attempt to make the sources seem artificially relevant to the question?

Table 7. Overview summary analysis.

Natural Language Query	Raters' Commentary (Summarized)
Sports and journalism have had a close relationship for more than 150 years in America. Which has had the greater influence on the other?	All researchers agree the overview summary just summarizes the sources and inaccurately attempts to connect them to the research question. The sources and summary do not address the historical perspective mandated by the research query.
How are book covers and their components represented in the digital market?	Sources are summarized well, but most—including the last one mentioned—do not tie to the question well; the summary throws in the phrase "including book covers" to tie the unrelated source to the topic, even when it does not. The most irrelevant sources are not mentioned in the overview summary, illustrating at least some sense that the AI understands something about relevancy.
How does music genre influence brain development in children?	Sources largely make connection to the exposure that music has on brain development, but no source compares specific music types or genre. The overview summary says influence of genre remains unclear and further research needed.
Do Video Doorbells Really Prevent Crime?	There was some disagreement among the raters regarding the summary, based on whether or not they believed that the effect of CCTV on crime was a useful extrapolation for the effect of video doorbells on crime. The summary did not answer the question directly, but used language such as "video doorbells may help deter certain crimes."
How and why did increased industrialization and mass production lead to an increase in drug use in the nineteenth century? How does this relate to globalization?	Summary attempts to answer question, but source coverage does not reach back to the 1800s; AI sets up sources by tying them to the topic even though they are not relevant. Raters question whether this is about all drugs or just illicit drugs; results focus on increased supply, not demand.

In addition to examining the overview statements for the five queries with no relevant documents (1–5 in Table 8), the researchers also looked at the PVE results for these five cases to determine if the ratio of relevant to irrelevant sources would be similar. While four of the five have at least one relevant source (Y) in the PVE results, there is no overall appreciable difference between PRA and PVE in producing relevant results. In a second evaluation inquiry, the researchers took the three queries where all PVE sources were irrelevant (5–7 in Table 8) and compared them to the corresponding PRA results. While two of the three have at least one relevant source (Y) in the PRA results, there was no notable difference between PVE and PRA in producing relevant results.

Table 8. Comparing largely irrelevant results sets between Primo Research Assistant (PRA) and Primo VE (PVE).

Query	PRA Results (Y/M/N)	PVE Results (Y/M/N)
1. Sports and journalism	0/0/5	3/3/3
2. Book covers	0/0/5	1/0/9
3. Music genre	0/0/5	2/4/4
4. Video doorbells	0/0/5	2/2/5
5. Increased industrialization	0/0/5	0/0/10
6. Giving middle schoolers iPads	3/1/1	0/0/10
7. Controversy around tobacco	1/0/4	0/0/10

DISCUSSION

Natural Language and Boolean Queries

One of the great innovations in searching technology is the ability to use NLQ syntax. All of the NLQs in this study originated from a Primo Zero Results report, meaning that patrons attempted and failed to find *any* materials via these queries. However, 96% of the original 138 NLQs within PRA *and* 100% of the expertly crafted Boolean counterparts within VE produced at least one citation or record. Further, 79% (81 out of 103) of the NLQs produced at least one *relevant* citation—closely following the expert-level Boolean queries—which successfully produced at least one relevant record in 86 of the 103 cases. Therefore, if library users do not want to rely on the creation of traditional Boolean structures—which is clearly effective in its own right—this study's data show they could instead rely on the new generative-AI technology to fill their preference to use NLQ syntax.

PRA and PVE Overall Source Relevancy

At the outset of this study, the results returned by the expert search queries were considered the bar by which the PRA citations were being measured. In other words, the researchers assumed that a greater overlap between PRA citations and PVE records would indicate better PRA performance. However, the analysis found that only 7.21% of the 499 citations returned by all the NLQs were successful in matching records returned by the expert search queries. This rate of non-matching was partially attributed to the fact that PRA returns citations from only a subset of what PVE has at its disposal, but this non-matching rate was deemed unsuitable as a performance metric because its efficacy would have required predetermining the universe of relevant documents for each query, which was beyond the scope of this paper. As demonstrated by Tay, some queries can be categorized as *Easy*—having 100+ relevant documents.[42] In such cases, both PRA and PVE could have retrieved relevant documents, but they exhibited no overlap with one another.

The *Yes*, *Maybe*, *No* distribution of relevant results between PRA and PVE was largely similar, with slightly higher *Yes* and *Maybe* ratings in PRA and higher *No* ratings in PVE. Due to the minimal differences in ratings, the raters concluded there was not enough evidence to favor one search tool over another in terms of performance. However, a distinction that can be drawn between the two tools is that they tend to deliver different types of resources, because they rely on different source data. PRA heavily favors journal articles, because it relies solely on records from the Central Discovery Index, while PVE incorporates a wide variety of resource types, such as books,

newspapers, government documents, and reference entries. Because Washington State University's discovery layer prefers local items, books are much more frequent within the PVE results. Positive distinctions of PVE are its ability to filter by resource type, date, and local availability, all of which are not options in PRA beta. Researchers also noted that duplicates were less likely to occur in PRA, which is a common issue in PVE.

Over the course of this study, the researchers noted that when repeating the same search in PRA, the five results returned were nearly always completely—or almost completely—different than the other identical searches, even over very short intervals (minutes) of time. Similar observations of non-determinism within the PRA Boolean search strategy and top five results are cited in Tay 2025.[43] Because the Boolean search query created when the user's question is sent to the LLM can change from search to search, the researchers speculate that this behavior can contribute to the retrieval of new sets of citations each time the system is run.[44] Regardless of the cause of this PRA behavior, the inability to replicate results consistently complicates the evaluation of the tool.

PRA Overview Summary

The researchers found a number of shortcomings in the presentation of the overview summaries. The raters analyzed the five PRA summaries where all five citations were rated *No* and found that even when the search result sources are overwhelmingly irrelevant, the summary attempts to answer the query with the sources anyway. This finding is similar to other studies that have also demonstrated inconsistencies between the retrieved sources and generated text.[45] The researchers also found that PRA rarely cites all five sources in its overview, periodically excluding those sources deemed most irrelevant by the researchers. Furthermore, the summary often incorporated language from the initial question to tie irrelevant sources to the topic, which led to a misrepresentation of the citations' content, often to the point of elevating questionable connections. For example, to answer a query about how book covers are represented in the digital market, PRA returned an irrelevant citation regarding queer women and digital identity on social media platforms. Although this citation was not related to the topic, the summary parroted language from the query and presented the source as relevant to the topic via the phrase "including book covers."

Also, across multiple cases, the overview (and the recommended sources) did not address the historical perspective mandated by the research question. This idea is tempered by the fact that PRA does not include newspaper articles or records from certain vendors (e.g., JSTOR), both of which likely contributed to the reduced prevalence of historical documents.

Despite these limitations, the overview summaries exhibit positive qualities. One of the PRA summaries included language that the topic connection "remains unclear" and "further research is needed." This indicates some evidence that PRA has the ability to exhibit restraint and a broader perspective in attempting to address research questions with the five sources offered. Although this type of statement would help a researcher be critical of the included sources, it only appeared in one of the five evaluated summaries.

CONCLUSIONS AND NEXT STEPS

The researchers devised this study to better understand the implications of implementing generative-AI technology (PRA) in parallel with the current discovery layer (PVE). Different avenues were explored, each with its own limitations. At the time of this study (December 2024 to March 2025), PRA was in beta, and the researchers look forward to the forthcoming enhancements that will allow for such functionality as filtering by resource type and the ability to

incorporate local records into the PRA search results. To test user NLQs, the researchers harvested actual searches performed by patrons via Primo Analytics. It was decided to use zero-result searches to provide answers to queries that previously went unanswered—both through AI-assisted technology and through expertly constructed Boolean searches. However, it could be argued that zero-result queries are not "average" queries, nor those that would naturally tend to produce significant sets of relevant results. As PRA is implemented at more academic institutions, we will have a greater number of datasets containing successful and failed NLQs, as they would occur in their *authentic* environments, to help further research in this regard.

There are other limitations to this study as well. The testing was not anonymized. The instruction librarians who rated the relevancy of the PRA results were the same as those who rated the relevancy of the PVE results. Ideally, the rating of relevancy should be done by an independent group who did not already know if the records were retrieved by the traditional algorithm or by AI.

The approach of using a title overlap analysis may not have been an adequate indicator of PRA performance because the two systems draw from different databases. Although the researchers were interested in comparing the default behaviors of each technology (as configured at Washington State University), future studies could attempt to align the PVE search scope with the types of materials and sources present in PRA. However, such an alignment would be difficult given that PRA excludes certain resource types, including newspaper articles, newsletters, and text resources or content from some providers, such as APA, DataCite, Elsevier, JSTOR, and Conde Nast.[46]

Since the overall ratings and analysis of PVE and PRA results did not reveal significant discrepancies, the primary distinction between the two tools is the summary feature in PRA. When the citations that PRA provides are relevant, the summary can be a helpful tool in putting the sources into context. However, when the citations that PRA provides are either irrelevant or questionable in quality, the summary attempts to assert their relevance by aligning the content to the research question, effectively "shoehorning" them to fit their topics. Researchers who are willing to look more closely at the citations and evaluate them in regard to the question could find this beneficial in thinking about sources more critically. However, researchers who may consider the inclusion of said citations in PRA at face value as being relevant sources could struggle with quality issues in the types of sources they include in their research.

Overall, PRA's leverage of natural language processing reduces the need for users to learn complex Boolean search strategies. However, further investigation is warranted with respect to audience and specific use cases. In the final analysis, the researchers were able to determine that the distribution of relevant results was relatively similar between PVE and PRA—a good sign for a beta product—and removes a significant barrier from the decision to implement PRA as an optional technology at Washington State University. Patrons who previously received no results to their queries could now receive both results and an associated summary of those results with a relevancy rate that is comparable to the existing discovery layer.

One topic that this study wholly does not address is the potential effect such technology might have on patrons' long-term ability to create search queries and discern relevant information on their own. Within the qualitative portion of the King study, one participant commented that PRA "would encourage laziness!"[47] The degree to which library staff help patrons varies, but it probably never carries to the extent of providing a full explanation of source interconnectedness

to research questions. This sentiment echoes the *dilemma of the direct answer*, described as "a user's choice between convenience and diligence."[48] Our next steps are to conduct usability studies to understand if this technology meets users' needs and to work with instruction librarians to understand how best to inform Washington State University's users of the inherent risks and advantages of AI.

ENDNOTES

[1] "IGeLU 2024 Full Program," International Group of ex Libris Users, 2024, https://igelu.org/archive-of-presentations/2024-copenhagen/igelu-2024-full-program/.

[2] "Getting Started with Primo Research Assistant," Ex Libris, Part of Clarivate, 2024, 1, https://knowledge.exlibrisgroup.com/Primo/Product_Documentation/020Primo_VE/Primo_VE_(English)/015_Getting_Started_with_Primo_Research_Assistant.

[3] "Getting Started," 1.

[4] "IGeLU 2024," Ex Libris, Part of Clarivate, 2024, https://knowledge.exlibrisgroup.com/Cross-Product/Conferences_and_Seminars/IGeLU/005IGeLU_2024.

[5] "IGeLU 2024."

[6] "IGeLU 2024."

[7] "IGeLU 2024."

[8] "IGeLU 2024."

[9] Aster Zhao, "Trust in AI: Evaluating Scite, Elicit, Consensus, and Scopus AI for Generating Literature Reviews," *Research Bridge* (blog), March 20, 2024, https://library.hkust.edu.hk/sc/trust-ai-lit-rev/.

[10] Shailja Gupta, Rajesh Ranjan, and Surya Narayan Singh, "A Comprehensive Survey of Retrieval-Augmented Generation (RAG): Evolution, Current Landscape and Future Directions" (arXiv, 2024), 2, https://doi.org/10.48550/arxiv.2410.12837.

[11] Nandan Thakur, et al., "'Knowing When You Don't Know': A Multilingual Relevance Assessment Dataset for Robust Retrieval-Augmented Generation," (arXiv, 2023), https://doi.org/10.48550/arxiv.2312.11361.

[12] Michael Townsen Hicks, James Humphries, and Joe Slater, "ChatGPT Is Bullshit," *Ethics and Information Technology* 26, no. 2 (2024): 1–4, https://doi.org/10.1007/s10676-024-09775-5; Karin Verspoor, "LLMs Can Combat LLM Hallucinations," *Nature (London)* 630, no. 8017 (2024): 569, https://doi.org/10.1038/d41586-024-01641-0.

[13] "Getting Started," 2.

[14] Will Douglas Heaven, "The Education of ChatGPT," *MIT Technology Review* 126, no. 3 (2023): 46.

[15] Debby R. E. Cotton, Peter A. Cotton, and J. Reuben Shipway, "Chatting and Cheating: Ensuring Academic Integrity in the Era of ChatGPT," *Innovations in Education and Teaching International*

61, no. 2 (2024): 229–230, https://doi.org/10.1080/14703297.2023.2190148; Heaven, "The Education of ChatGPT," 46.

[16] Heaven, "The Education of ChatGPT," 45–46; Danny Kingsley, "Can Generative AI Facilitate the Research Process?," *C&RL News* 84, no. 9 (2023): 343, https://doi.org/10.5860/crln.84.9.342.

[17] Damian Okaibedi Eke, "ChatGPT and the Rise of Generative AI: Threat to Academic Integrity?," *Journal of Responsible Technology* 13 (2023): 2, https://doi.org/10.1016/j.jrt.2023.100060.

[18] Eke, "ChatGPT and the Rise of Generative AI," 2.

[19] Jurgen Rudolph, Samson Tan, and Shannon Tan, "ChatGPT: Bullshit Spewer or the End of Traditional Assessments in Higher Education?," *Journal of Applied Learning & Teaching* 6, no. 1 (2023): 345, https://doi.org/10.37074/jalt.2023.6.1.9.

[20] Leo S. Lo, "My New Favorite Research Partner Is an AI: What Roles can Librarians Play in the Future?," *C&RL News* 84, no. 6 (2023): 210, https://doi.org/10.5860/crln.84.6.209.

[21] Adetoun A. Oyelude, "Artificial Intelligence (AI) Tools for Academic Research," *Library Hi Tech News* 41, no. 8 (2024): 20, https://doi.org/10.1108/lhtn-08-2024-0131.

[22] Susan Gardner Archambault and Jose J. Ricon, "An Evaluation of Cutting-Edge AI Research Tools using the REACT Framework," *Computers in Libraries* 44, no. 8 (October 2024): 10, https://www.infotoday.com/cilmag/oct24/Archambault-Rincon--An-Evaluation-of-Cutting-Edge-AI-Research-Tools-Using-the-REACT-Framework.shtml; Oyelude, "Artificial Intelligence (AI) Tools," 20; Kingsley, "Can Generative AI Facilitate the Research Process?," 343.

[23] Cotton, "Chatting and Cheating," 3.

[24] Kingsley, "Can Generative AI Facilitate the Research Process?," 342.

[25] Archambault, "An Evaluation of Cutting-Edge AI Research Tools," 6.

[26] Archambault, "An Evaluation of Cutting-Edge AI Research Tools," 9.

[27] Archambault, "An Evaluation of Cutting-Edge AI Research Tools," 8.

[28] M. D. Ashikuzzaman, "Exploring AI-based Recommendation Systems in Libraries," *LIS Education Network* (blog), March 7, 2024, https://www.lisedunetwork.com/exploring-ai-based-recommendation-systems-in-libraries/.

[29] Adam Hyde, John Chodacki, and Paul Shannon, "An Initial Scholarly AI Taxonomy," *Upstream* (blog), April 11, 2023, https://doi.org/10.54900/6p6re-xyj61.

[30] Joshua Mitcham, "Unlocking Hidden Knowledge: Harnessing AI, Technology, and Design for Accessible Research," *Information Service & Use* 44, no. 1 (2024): 44, https://doi.org/10.3233/isu-230223.

[31] Kingsley, "Can Generative AI Facilitate the Research Process?," 343.

[32] Hyde, "An Initial Scholarly AI Taxonomy"; Mitcham, "Unlocking Hidden Knowledge," 47.

[33] Can Li and Hesper Wilson, "Primo Research Assistant: Potential for Enhancing Resource Discovery (A Six-Month Review)," *Internet Reference Services Quarterly* 29, no. 2 (2025): 2–4, https://doi.org/10.1080/10875301.2025.2480251.

[34] Charlene King, "Primo Research Assistant: User Testing and Feedback" (The Open University, 2025), 5, https://doi.org/10.21954/ou.ro.00103099.

[35] King, "Primo Research Assistant," 6.

[36] King, "Primo Research Assistant," 10–12.

[37] Aaron Tay, "Deep Dive into Three AI Academic Search Tools," *Katina Magazine*, May 20, 2025, https://doi.org/10.1146/katina-052025-2.

[38] Tay, "Deep Dive."

[39] Tay, "Deep Dive."

[40] Tay, "Deep Dive."

[41] "Primo Zero Result Searches," Ex Libris, Part of Clarivate, 2025, https://knowledge.exlibrisgroup.com/Primo/Product_Documentation/Primo/Analytics/040Primo_Analytics_Subject_Areas/Primo_Zero_Result_Searches.

[42] Tay, "Deep Dive."

[43] Tay, "Deep Dive."

[44] "Getting Started," 2.

[45] Thakur, "A Comprehensive," 4; Zhao, "Trust in AI"; Tay, "Deep Dive."

[46] "Getting Started," 2.

[47] King, "Primo Research Assistant," 17.

[48] Martin Potthast, Matthias Hagen, and Benno Stein, "The Dilemma of the Direct Answer," *SIGIR Forum* 54, no. 1 (2020): 1, https://doi.org/10.1145/3451964.3451978; Frauke Birkhoff, "Guest Post: Eight Hypotheses Why Librarians Don't Like Retrieval Augmented Generation (RAG)," *The Scholarly Kitchen* (blog), May 8, 2025, https://scholarlykitchen.sspnet.org/2025/05/08/guest-post-eight-hypotheses-why-librarians-dont-like-retrieval-augmented-generation-rag.