more direct title

Evaluating LLM-Driven Book Discovery: A Comparative Study of a RAG-Based Chat Interface vs. Traditional Search in Digital Libraries

May 5th, 2025

NICOLE LEÓN*, Pennsylvania State University, USA
MATT MURTAGH WHITE*, Trinity College Dublin, Ireland
YUNKAI XU*, Pennsylvania State University, USA

Keyword-based search remains the default interaction model for public digital libraries, yet it often limits some users in discovering relevant literature. We present a conversational book discovery interface powered by a Retrieval-Augmented Generation (RAG) Large Language Model (LLM), designed as an alternative to traditional search on platforms like Project Gutenberg. The system supports natural language queries, visualizes literature relationships through interactive graphs, and summarizes book metadata using bibliographic data. We conducted a within-subjects study in which participants performed book retrieval tasks using both our prototype and the original website. Quantitative and qualitative results indicate notable improvements in retrieval accuracy, user satisfaction, and support for exploratory search. We also discuss challenges related to LLM-generated content and propose design directions for integrating LLMs into digital library interfaces. Our work demonstrates how LLMs can extend the interaction capabilities of open-access tools toward more natural, engaging, and discoverable digital repository experiences.

so, note them SD or % or raw #

CCS Concepts: • **Information systems** → *Search interfaces*; • **Human-centered computing** → **Empirical studies in interaction design**.

Additional Key Words and Phrases: Human-centered Design, Searching System, Interaction Design

ACM Reference Format:

Nicole León, Matt Murtagh White, and Yunkai Xu. 2025. Evaluating LLM-Driven Book Discovery: A Comparative Study of a RAG-Based Chat Interface vs. Traditional Search in Digital Libraries [0.5em] *May 5th, 2025*. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 22 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

The rise of large language models (LLMs) has fundamentally redefined the design of information retrieval (IR) systems, offering new opportunities for adaptive, conversational interaction [17, 32]. Conversational agents powered by these models increasingly supplement or replace traditional search interfaces, especially in domains where users face cognitive, linguistic, or conceptual barriers. Project Gutenberg, one of the most expansive and longstanding digital literature

*All authors contributed equally to this research.

Authors' Contact Information: Nicole León, Pennsylvania State University, University Park, Pennsylvania, USA, nicoleleon@psu.edu; Matt Murtagh White, Trinity College Dublin, Dublin, Dublin, Ireland, mmurtagh@tcd.ie; Yunkai Xu, Pennsylvania State University, University Park, Pennsylvania, USA, yqx5322@psu.edu.

Manuscript submitted to ACM

a ? — share w/ Noah to cody to Bin

repositories [22], exemplifies both the promise and the challenge of open-access cultural archives. While it provides free access to tens of thousands of literature, users must navigate an interface optimized for keyword lookup and categorical browsing modes that are often misaligned with exploratory or curiosity-driven engagement [6, 20].

Exploratory search, defined as an open-ended process of learning, investigating, and discovery [15], is especially salient in digital library contexts. Prior research highlights that users in these settings often confront unfamiliar domains and struggle with vague goals, uncertain pathways, or evolving information needs [2, 29]. Despite significant advances in visual interfaces and metadata design [3, 16], mainstream digital libraries still remain largely plain, static, and text-driven. They offer limited scaffolding for sense making, reflection, or iterative exploration [15, 25, 26]. This creates an opportunity to examine whether conversational agents, particularly those enhanced by LLMs, can offer a more responsive, user-aligned alternative for accessing cultural and historical texts.

Recent studies have begun to explore this space. Custom-built LLM agents such as Walert et al. [21] and MoodleBot et al. [18] illustrate how dialogic interaction can support domain-specific information seeking and learning. These systems offer planning, memory, and contextual reasoning capabilities previously unavailable in IR systems [33]. However, most empirical studies have focused on well-structured academic or educational environments, and less is known about how these agents function in high-context, public cultural repositories like Project Gutenberg [18, 27]. Therefore, while the potential for exploratory engagement is often discussed, few studies have empirically examined how conversational interfaces influence actual search strategies compared to traditional digital library baselines.

Another critical dimension is trust. Users interacting with AI agents in complex information tasks must calibrate their trust based not only on correctness but also on the transparency, tone, and initiative of the system [1, 23, 28]. Studies have shown that interface modality alone can shift trust perceptions [30], and that ChatGPT-like models often outperform search engines in credibility judgments during some health information tasks [31]. As conversational systems begin mediating access to cultural heritage and historical literature, it is crucial to understand how design choices influence trust in both the content and the agent itself.

In this work, we reimagine public-domain book discovery through a GPT-powered conversational interface that engages users in multi-turn, adaptive dialogue and result visualization. We focus on two **research questions**:

- **RQ1:** whether conversational interfaces promote more exploratory search behavior compared to traditional digital library interfaces;
- **RQ2:** how trust and verification behaviors are shaped for static interfaces vis-a-vis knowledge graph enhanced language model agents.

By embedding LLM capabilities within a publicly accessible cultural archive, we contribute new empirical insights at the intersection of digital libraries, exploratory search and human-AI interaction.

2 Related works

2.1 Introducing *Project Gutenberg*

Project Gutenberg [22] is one of the earliest and most extensive digital repositories of public domain literature, providing open access to over 60,000 texts spanning diverse genres, languages, and historical periods. The primary mode of interaction is through its website, which offers keyword-based search, categorical browsing, and metadata filtering to help users locate specific titles of interest. Users can read books online or download them in plain text or e-book formats at no cost. Due to its availability and rich coverage, Project Gutenberg has become a widely used resource in academic

research. For example, prior work has leveraged the Project Gutenberg collection as a resource for computational analysis [9, 12, 24].

Building on this foundation, our work reimagines access to the Project Gutenberg corpus through a conversational search interface. By embedding these texts into an interactive dialogue system, we aim to lower cognitive entry barriers and foster more intuitive, exploratory engagement with cultural heritage materials.

2.2 Exploratory Search in Digital Libraries

Exploratory search refers to search behaviors where users aim to learn, investigate, or discover, rather than simply locate a known item [15]. This mode of search is particularly relevant in digital libraries, where users often engage with complex or unfamiliar domains. Existing studies have shown that digital libraries struggle to support such exploratory behaviors due to limited scaffolding for goal refinement, iterative discovery, and contextual serendipity.

Prior research has laid important theoretical and design foundations for supporting this mode of search. Marchionini [15] emphasized that exploratory systems must facilitate browsing, reflection, and sense-making across multiple stages of the search process. Shen et al. [29] further identified key usability challenges in digital library interfaces, including limited contextual cues and inflexible navigation pathways. In response, Bernard et al. [2] proposed a user-centered interface model that aligns system affordances with the cognitive strategies of exploratory users, particularly in the context of scientific research data.

Recent studies have expanded this discussion by focusing on user motivations, search tactics, and the importance of design features in shaping exploratory engagement. For instance, Boon et al. [3] examined complex search behaviors within public digital libraries, while McCay-Peet et al. [16] proposed empirical metrics to differentiate exploratory from fact-finding behaviors. Collectively, these works underscore the need for systems that not only expose diverse content but also dynamically support users in clarifying goals, navigating uncertainty, and sustaining engagement.

However, despite growing interest in exploratory search, existing digital library interfaces remain predominantly static and text-based, offering little interactive engagement for users navigating vast, open-ended archives. Few studies have explored how conversational interactions might serve as scaffolds for such search behavior. Our work addresses this gap by evaluating whether a GPT-powered chatbot can enhance exploratory book discovery in the Project Gutenberg corpus through adaptive, multi-turn dialogue.

2.3 Conversational Interfaces for Information Retrieval

Conversational interfaces are increasingly positioned as alternatives to traditional search systems, offering natural language interaction and adaptive support [17]. Yet, most real-world deployments remain constrained to simple tasks, and their value in supporting complex or exploratory information-seeking remains unclear.

Prior research has documented persistent mismatches between user expectations and the capabilities of conversational agents. Luger and Sellen [14] found that users often felt frustrated with assistants like Siri and Alexa due to their limited ability to handle ambiguous or nuanced queries. Papenmeier et al. [20] showed that users tend to overestimate agent intelligence, leading to unmet expectations and reduced trust. Schmitt et al. [27] demonstrated that user-centered design significantly improves trust, enjoyment, and performance in educational retrieval tasks, suggesting that default conversational agent designs often fall short. Building on this, Deng et al. [7] argued for a shift toward human-centered proactive conversational agents that prioritize intelligence, adaptivity, and civility, emphasizing that proactive behaviors must be socially appropriate and ethically aligned to avoid being perceived as intrusive.

Recent advancements in LLMs have catalyzed a paradigm shift in the design of information retrieval systems [17, 32]. Zhang et al. [33] argue that LLM-powered agents can overcome long-standing limitations in personalization, interactivity, and contextual reasoning by integrating human-like capabilities such as memory and planning. This shift has renewed interest in conversational interfaces as a medium for supporting more dynamic, user-aligned retrieval experiences [32]. Demonstrating this shift in practice, Cherumanal et al. [21] developed Walert, a customized LLM-based chatbot designed to support academic information-seeking scenarios. Their deployment highlights how best practices in conversational IR can be operationalized to meet domain-specific user needs, bridging the gap between research prototypes and real-world applications. Neumann et al. [18] introduced MoodleBot, an LLM-driven educational chatbot integrated into a learning management system to support self-regulated learning and help-seeking behavior. Their evaluation showed strong user acceptance and instructional benefits, reinforcing the potential of conversational agents to facilitate personalized support in structured academic contexts.

Despite these developments, few studies have examined how conversational agents perform in domain-specific, high-context environments like digital libraries. Our work addresses this gap by evaluating whether a GPT-based chatbot can support exploratory book discovery in the Project Gutenberg collection, and how it compares to traditional library interfaces.

2.4 Trust in Conversational AI

Trust is a foundational concern in the design of conversational agents, especially as LLM-powered systems are increasingly deployed for complex information-seeking tasks across domains. Prior work has identified a range of factors influencing trust, including agent appearance, communication style, and perceived interaction performance [23]. This understanding has been extended to high-stakes contexts such as health information retrieval, where ChatGPT has been shown to elicit higher trust than traditional search engines like Google [31]. Moreover, interface modality plays a critical role: even when the underlying LLM output is held constant, trust perceptions vary significantly depending on whether the interaction is text-based, speech-based, or embodied [30]. Beyond interface features, trust is also shaped by the system initiative, how actively the agent directs the interaction [1], and the degree of transparency and control users feel during the information-seeking process [28]. Building on this literature, our work explores trust in the context of cultural and literary discovery, where LLM-powered conversational and visual interfaces may promote greater user confidence, credibility, and exploratory engagement in open-access digital libraries.

3 Athena Design

3.1 Design Overview

The final prototype of Athena (Figure 1) encapsulates its core design principles through a workflow consisting of four stages: query, recommendation, visualization, and interaction. Users begin by entering prompts in the chat window, such as genres, moods, or author names. Athena then generates personalized book recommendations based on Project Gutenberg metadata. These books are immediately rendered as nodes and edges in a dynamic graph, revealing semantic or authorial relationships among them. Users can interact with the graph to select one or more books, which serve as anchors for follow-up dialogue. These follow-ups support reflective, comparative, or thematic inquiries contextualized by the selected items. Throughout the process, users can view metadata, generate summaries, and download books into persistent collections. By maintaining a synchronized conversational and visual state, Athena allows users to

fluidly transition between expression, exploration, and organization. The section below details each component of this interaction flow.

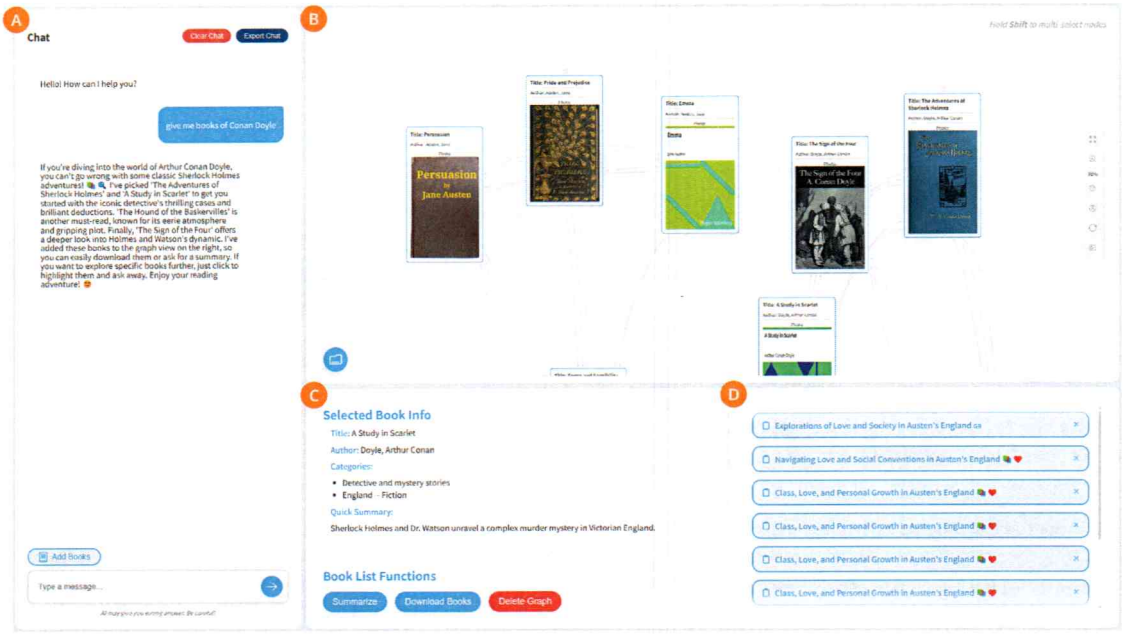

Fig. 1. Screenshot of Athena's user interface, which supports conversational book exploration through a multimodal system. (**A**) The chatbot panel enables natural language queries and returns rich, context-aware responses. (**B**) The knowledge graph visualizes books and their interconnections, allowing users to explore relationships. (**C**) The book information panel displays metadata and provides actionable functions such as summarization or download. (**D**) The dynamic tag panel stores user-generated thematic interpretations, supporting reflection and comparison.

3.2 Key Functions

3.2.1 Dialogue-Driven Exploration and Visual Integration. The conversational panel (Figure 2 A) in Athena serves as more than a query input interface; it anchors the entire exploration process. When users initiate book exploration by selecting the "Add Books" function and posing open-ended prompts (e.g., author names, genres, or moods), Athena responds with curated book recommendations grounded in literary metadata in Project Gutenberg, as seen in Figure 2. Here are some exploratory prompts:

- **Thematic discovery**: Queries about authors, topics, genres, or literary movements (e.g., "books about isolation", "books of Conan Doyle").
- **Historical or authorial curiosity**: Prompts about authors' backgrounds or publication contexts (e.g., "What's special about 19th-century detective fiction?").
- **Affective exploration**: Questions rooted in personal emotions or reading moods (e.g., "something calming to read").

Besides, these books retrieved by Athena are not only rendered in the conversation panel but also simultaneously instantiated as visual nodes in the graph (Figure 2 C) . The system further visualizes semantic relationships between newly

added books and those already in the graph, supporting continuity and thematic coherence. This visual augmentation enables users to externalize the results of conversational interactions, transforming ephemeral dialogue into the persistent, manipulable structure.

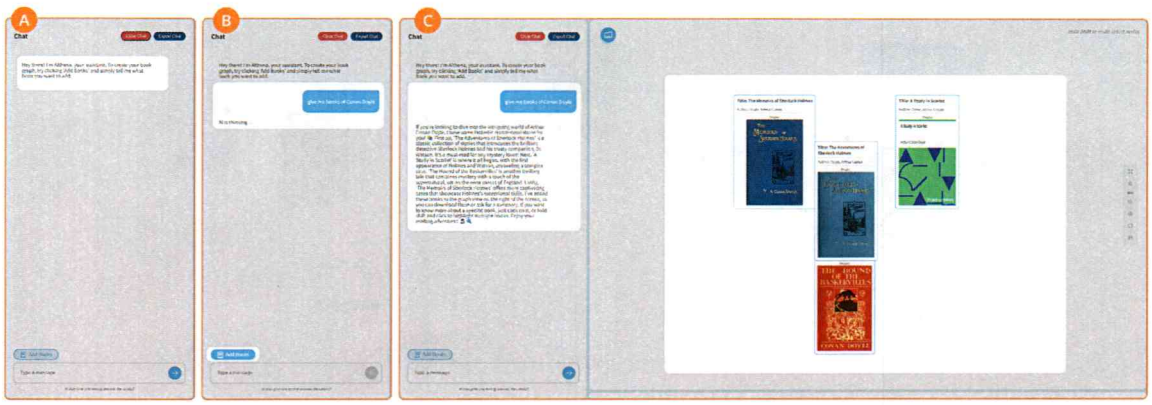

Fig. 2. Athena's dialogue-based book search process. (**A**) Athena's welcome message (**B**) The user initiates a conversation by requesting books, triggering the system to retrieve relevant titles and Athena begins processing the query, displaying a thinking state. (**C**) Athena generates a rich natural language response with curated book suggestions, which are then visualized on the right as a book graph.

3.2.2 Conversational Follow-up Grounded in Visual Selections. At the center of the interface is a force-directed graph view that serves as the primary visual navigation space. Each node represents a book, labeled with title, author and the cover. Edges denote meaningful relationships, such as shared authorship or user-tagged thematic similarity, which are aslo created by LLM when Athena is adding books to the graph. This layout allows users to see both the local context of selected books and broader patterns across the corpus. Users can freely pan and zoom the graph, select or multi-select nodes, and observe how visual connections correspond to semantic or authorial groupings. The graph affords exploratory interaction, enabling a shift from query-based discovery to visual sensemaking.

Beyond a visual representation, users may also engage in targeted dialogue by selecting one or more books from the graph and using them as conversational anchors. Once a selection is made, follow-up questions posed in the conversational panel are interpreted in the context of the selected items, as seen in Figure 3. This design supports multiple classes of interaction, such as:

- **Reflective engagement**: Asking for thematic interpretations, historical context, or genre placement of selected books.
- **Comparative inquiry**: Requesting differences or commonalities among a selected group of texts.
- **Exploratory extension**: Seeking related works that expand upon the themes or styles of the selected items.

In a word, the dialogue flow is tightly integrated with the graph's state: system-generated book recommendations from these conversations are always visualized, maintaining a unified interaction history across modalities.

3.2.3 Contextual Book Information. When a user selects a book node from the graph, a contextual detail panel appears at the bottom of the interface, displaying metadata such as title, author, categories, and a very brief summary, as seen in

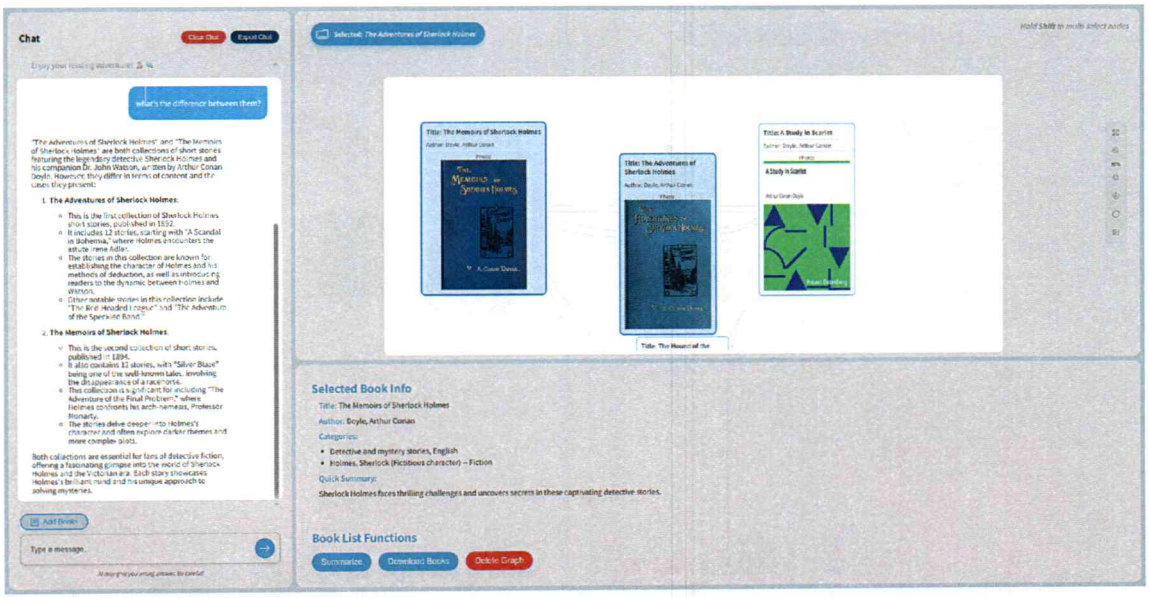

Fig. 3. Node-based querying workflow in Athena. Users can select one or more books as contextual anchors for conversation. In this example, the user asks for a comparison between two Sherlock Holmes collections, and the system generates a structured, paragraph-level comparison in the chat panel.

Figure 4. This panel supports deeper engagement by encouraging users to pause, interpret, and reflect on individual titles during exploration.

3.2.4 Summarize. The "Summarize" button (Figure 5) allows users to generate a concise, AI-authored summary based on the selected nodes. Rather than merely aggregating existing synopses, the system produces higher-order syntheses, highlighting thematic connections, stylistic contrasts, or narrative trends. For example, when selecting books within the Gothic genre, the generated summary might compare recurring motifs like isolation or madness across authors and time periods.

3.2.5 Download Books. The "Download" button links users to the full-text versions of selected books via Project Gutenberg. Clicking this button opens a new browser tab with the plain-text source, enabling offline reading or further annotation.

4 Implementation

Athena's backend is implemented using Flask, a lightweight Python web framework that manages HTTP requests and orchestrates interactions between frontend components, retrieval endpoints, and the language model. The system is designed around a modular architecture that integrates the Gutendex API—an open-access metadata service for Project Gutenberg—and OpenAI's GPT-4o for natural language understanding and generation. A key challenge for LLMs in information retrieval tasks such as this is hallucination, whereby an LLM will give probable answers that appear correct at face value but are actually incorrect [11]. An example in this context could be a suggestion of a book that is not in the public domain or not available via Project Gutenberg. To counter this, the system components described above

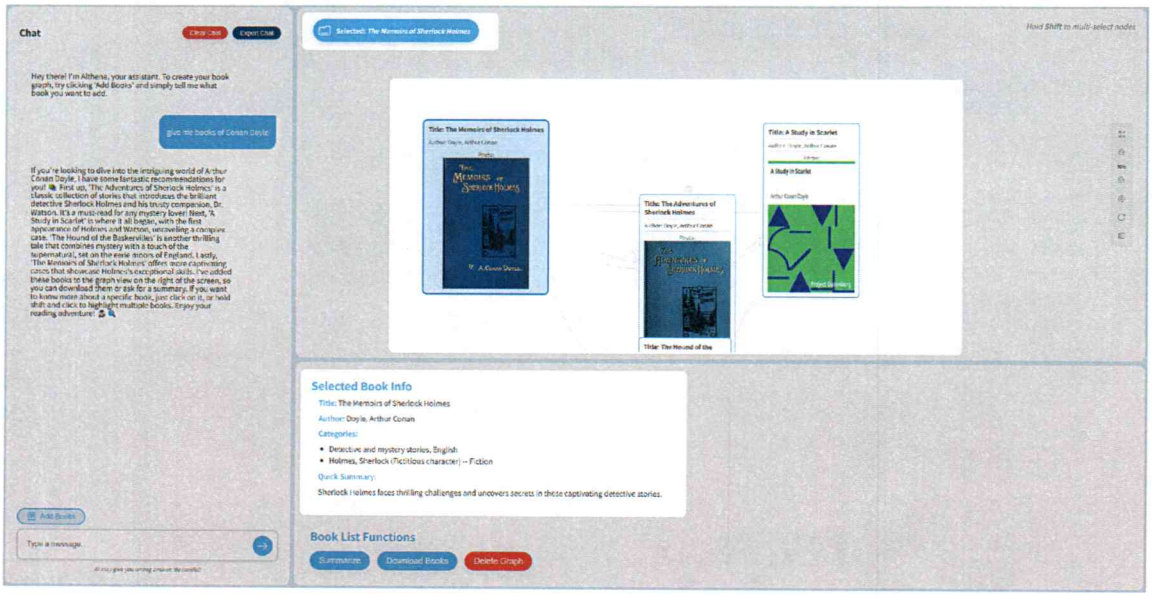

Fig. 4. When a user selects a book node (e.g., *The Memoirs of Sherlock Holmes*), Athena displays its metadata (title, author, categories, and a very brief summary) and summary in the lower panel. A top bar shows the selected book's title, or the relationship if a line is selected.

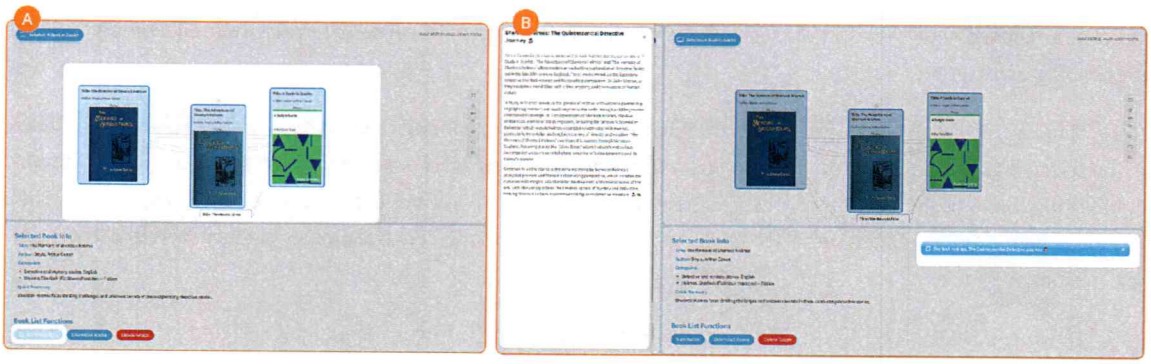

Fig. 5. Workflow of AI-assisted book summarization in Athena. (**A**) Users can click the `Summarize` button to generate a natural language summary based on the currently selected book node. (**B**) The generated summary is added to a persistent summary list, and clicking on a summary reveals its full content in an expandable panel.

work together to create Retrieval-Augmented Generation (RAG) [13], a system that helps ground the language model in the task at hand using retrieved information. This practice has been shown to reduce hallucination [13].

4.1 System Architecture and Workflow

The backend follows a RESTful API architecture, with distinct endpoints for processing user queries, managing graph generation, answering questions about selected books, and summarizing groups of texts. As illustrated in Figure 6, the

system is structured into three layers. The Presentation Tier is implemented in Vue and handles the user interface, including chat interaction, graph rendering, and metadata display. The Logic Tier, implemented in Flask and served through Nginx and Gunicorn, manages user inputs and coordinates responses. The Data/Service Tier integrates GPT-4o via OpenAI's API and performs retrieval via Gutendex.

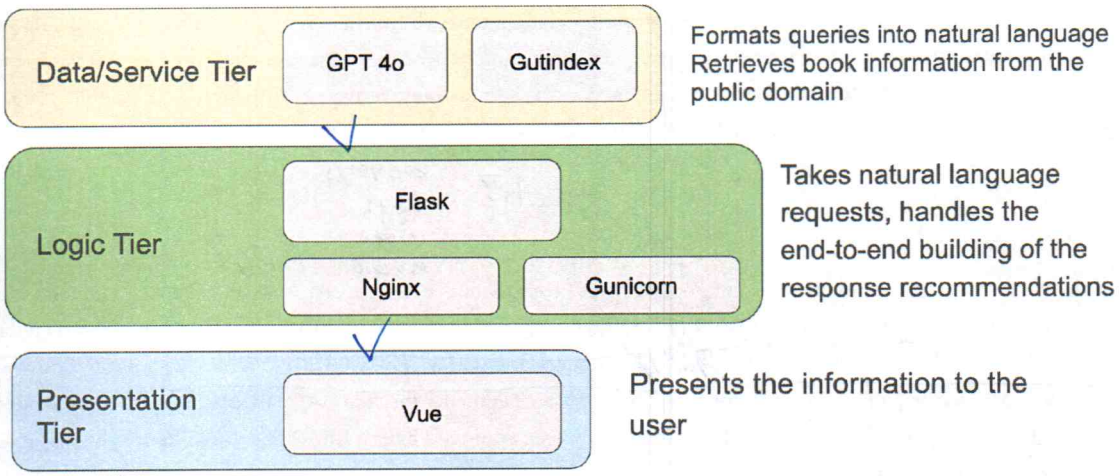

Fig. 6. System architecture of the AI-powered book recommendation engine. (**A**) The **Presentation Tier**, built with Vue, provides the user interface and displays generated recommendations. (**B**) The **Logic Tier** uses Flask, Nginx, and Gunicorn to manage natural language requests and orchestrate response generation. (**C**) The **Data/Service Tier** leverages GPT-4o for language processing and Gutindex for retrieving public-domain book data.

4.2 Retrieval-Augmented Recommendation Generation

When users submit a natural language prompt, such as "Show me books about dystopian societies", the backend routes this request through a LangChain pipeline to GPT-4o, which interprets the intent and reformulates the query into a structured call to the Gutendex API. Gutendex returns a collection of book metadata including titles, authors, subjects, short descriptions, and download links. This metadata is then filtered and deduplicated to ensure content relevance and diversity.

The selected metadata is injected back into a GPT-4o prompt that explicitly asks the model to generate a user-facing response grounded in the retrieved data. The LLM is instructed to justify each recommended book using the available metadata to reduce hallucination. Depending on the user's prompt, a variable number of results are selected and woven into a structured, paragraph-level reply that describes the books and explains why they were chosen in context. The returned results are also formatted into a consistent JSON structure for visual rendering, with each book instantiated as a node in a force-directed graph. Edges between nodes represent meaningful relationships such as shared authorship or LLM-inferred thematic similarity.

4.3 Interactive Book Exploration and Summarization

Athena supports follow-up queries anchored on selected books from the graph interface. When users select one or more books and ask additional questions in natural language, the backend retrieves the full metadata for the selected

titles and injects it into a grounding prompt to GPT-4o. This technique ensures that the model responds specifically to the books in question, allowing users to explore comparative, reflective, or thematic questions with targeted focus. For example, a prompt such as "How do these two Sherlock Holmes books differ in tone?" is answered with a response that references the known metadata, ensuring that the comparison remains faithful to the source information.

A similar RAG process underlies Athena's summarization functionality. When users invoke the "Summarize" action for a selection of books, the system collects the relevant metadata for each title and injects this into a tailored GPT-4o prompt that requests a concise, higher-order summary. The prompt encourages the model to move beyond a listing of features and instead synthesize common themes, contrasts, and interpretive patterns across the books. These AI-generated summaries are then rendered in the frontend's summary panel and stored for later review.

5 Methodology

5.1 Participants

Participants were recruited from the student population at Pennsylvania State University (PSU). A total of 8 undergraduate and graduate students took part in the study. Recruitment targeted individuals with a range of experience with digital libraries and LLMs; we recorded this data in order to understand patterns of use with familiarity. Participation was voluntary, and all participants provided informed consent before beginning the study. All participants were compensated financially for their participation in the study. This study was approved by the IRB Office at the Pennsylvania State University, study number 00026694.

Table 1. Participant Backgrounds: Reading Habits, Digital Library Use, and LLM Familiarity

ID	Reading Frequency	Digital Library Familiarity	LLM Familiarity	Typical Sources Used
1	Very frequently	Slightly familiar	Moderately familiar	Libgen
2	Rarely	Moderately familiar	Familiar	ACM DL, Google Scholar, IEEE Xplore
3	Rarely	Very familiar	Very familiar	–
4	Occasionally	Slightly familiar	Moderately familiar	–
5	Occasionally	Very familiar	Very familiar	ACM DL, IEEE Xplore
6	Somewhat frequently	Moderately familiar	Moderately familiar	–
7	Very frequently	Moderately familiar	Moderately familiar	Internet Archive
8	Somewhat frequently	Very familiar	Very familiar	ACM DL, IEEE Xplore, ACL Anthology, arXiv

5.2 Materials

The study compared two digital interfaces for book discovery. The first interface, Athena, was a conversational agent built on a Retrieval-Augmented Generation (RAG) architecture, ~~utilizing~~ GPT-4o for natural language generation. Athena supported multimodal exploration, including a dynamic book graph visualization and AI-generated metadata summaries. The second interface was the standard static Project Gutenberg website, which relied on traditional keyword-based search functionalities.

Additional materials included a set of eight structured information-seeking tasks, post-task surveys administered after each task, a final comparative survey, and the System Usability Scale (SUS), which participants completed separately for

each interface. Participants engaged in a think-aloud protocol during task completion. Screen recordings and audio data were collected for subsequent behavioral coding and analysis.

5.3 Experimental Design

The study employed a within-subjects experimental design, where each participant used both interfaces during the session. The order in which participants interacted with Athena and the static interface was counterbalanced to mitigate order effects and learning biases.

Each participant completed four assigned tasks with one interface, followed by four different but matched tasks with the other interface. Tasks were crafted to cover a spectrum of search activities, ranging from known-item lookups to open-ended exploratory queries, with particular attention to the cognitive demands and verification behaviors elicited by each system.

The independent variables in the study were the interface type (Athena vs. Static) and task type (lookup vs. exploratory). Dependent variables included objective behavioral measures such as query counts, time-on-task, and navigation depth, as well as subjective evaluations of perceived ease, trust, expressiveness, and exploration.

5.4 Tasks

Participants performed eight tasks in total, evenly divided between lookup-oriented and exploratory search scenarios. Lookup tasks involved structured retrieval goals, such as finding a specific title or author, while exploratory tasks required participants to discover books based on thematic, genre, or affective prompts.

Tasks were explicitly designed to elicit differences in search behavior between interfaces. Lookup tasks served as a baseline for measuring efficiency and accuracy, whereas exploratory tasks assessed the richness, breadth, and depth of users' information-seeking behavior.

Detailed task prompts and descriptions are provided in appendix A.

5.5 Procedure

After providing informed consent, participants received a brief onboarding session introducing both systems. Participants were allowed to complete an optional practice task to familiarize themselves with the interface mechanics.

Participants then completed four tasks using the first assigned interface. After each task, they filled out a short survey capturing perceptions of task ease, confidence in their answers, expressiveness afforded by the system, and whether they discovered anything unexpected. Throughout the task completion, participants were instructed to verbalize their thought processes following a think-aloud protocol [4, 8]. Their interactions were recorded for subsequent analysis. Subsequently, participants completed four additional tasks using the second interface under identical data collection conditions.

At the end of the study, participants completed a comprehensive post-evaluation survey. This survey captured comparative ratings between the two systems on trust, usability, perceived exploration support, and overall preference. Participants also completed two separate instances of the System Usability Scale (SUS) [5], one for each interface. Finally, participants were invited to provide open-ended reflections to offer deeper insights into trust formation, exploration strategies, and perceptions of the system affordances and limitations.

6 Evaluation

6.1 Measures

We adopted a mixed-methods approach combining behavioral observation, usability surveys, Likert-scale questionnaires, and qualitative analysis to evaluate user interaction, trust formation, and exploratory behavior across the two systems.

Behavioral Measures: We captured interaction patterns through screen recordings and observation notes. These included the number and type of queries participants issued, the use of external tools such as Google for verification, interaction with system elements like book nodes and summaries, and instances of query reformulation. We also qualitatively assessed the time participants spent on each task based on observed effort and navigation depth. These data informed our understanding of how participants navigated the interfaces and whether their strategies reflected lookup or exploratory behavior.

Survey Measures: To assess subjective user experience, we administered the NASA Task Load Index (NASA TLX) [10] following each interface session. This instrument measured perceived mental demand, physical effort, time pressure, frustration, overall effort, and perceived task success. We also employed the System Usability Scale (SUS)[5], a ten-item questionnaire evaluating perceived usability, complexity, integration, and learnability. In addition, we designed custom Likert-scale items to gauge perceived trustworthiness, expressiveness, ease of use, and support for exploration. These instruments offered both standardized and task-specific insights into users' cognitive load and satisfaction.

Exploration and Trust Indicators: We inferred exploratory behavior from several indicators, including the use of multi-turn conversational queries, selection and comparison of multiple books, engagement with AI-generated summaries, and the nature of follow-up inquiries. Indicators of trust included participants' self-reported trust ratings, observed behaviors such as fact-checking or cross-referencing using external sources, and verbal expressions of confidence, doubt, or skepticism during think-aloud sessions. These measures helped identify how participants interpreted the system's credibility and how trust shaped their interaction strategies.

6.2 Data Analysis

We analyzed the collected data using a combination of quantitative and qualitative methods, applying a concurrent triangulation design to identify converging or diverging patterns across multiple data sources.

For the quantitative component, we relied on descriptive statistics and comparative visualizations to examine survey responses, including those from the NASA Task Load Index (TLX), the System Usability Scale (SUS), and trust-related Likert items. Given the small sample size (N = 8) for this emerging research, we did not perform inferential statistical tests. Instead, we emphasized interpretive analysis, using medians and distribution plots to characterize participant responses. This approach was appropriate given the within-subjects, exploratory nature of the study. Additionally, we reviewed behavioral data extracted from session recordings, tallying observable actions such as the use of external verification tools, query reformulations, and user interactions with key system features including graph nodes, metadata panels, and summary generation.

Qualitative data were analyzed using a six-phase thematic analysis approach as outlined by Nowell et al. [19]. The process involved initial familiarization with the data, generation of initial codes, iterative theme development and review, clear definition and naming of final themes, and structured reporting. This method was applied to both think-aloud transcripts and post-task open-ended responses. Emergent themes included usability barriers, trust and credibility, cognitive load, emotional reactions, and system affordances. To assist with consistency and efficiency in the coding

process, a lightweight Python script was developed to automatically count and summarize theme occurrences across participant responses.

6.3 ~~Quantitative~~ Results

6.3.1 NASA Task Load Index (TLX). To evaluate participants' subjective experiences across multiple dimensions, the NASA Task Load Index (NASA TLX), the System Usability Scale (SUS), and custom Likert scale questionnaires acted as measurement tools. These measures were selected to capture a comprehensive view of the perceived workload, usability, and specific user attitudes toward each system. The following section presents qualitative results derived from these instruments.

A comparative qualitative analysis of the NASA Task Load Index (TLX) responses for the Athena and Project Gutenberg systems reveals distinct patterns in the perceived workload of participants in physical, mental, and temporal states. Each measure was rated on a scale of 0-10, with higher values indicating workload in addition to greater effort, frustration, and performance.

As visualized in Figure 7, the Project Gutenberg system revealed notably higher average scores in mental demand (M = 5.125), frustration (M = 5.000), ~~performance (M = 5.000)~~, and effort (M = 4.625). These categories skewed toward the upper half of the scale, particularly mental demand, which had a median of 6.5, suggesting that several participants rated this aspect at 7 or higher. The horizontal bar distribution confirms this upward skew, with orange cells concentrated in the 6 to 8+ range.

However, the reported physical demand remained very low (M = 0.875), with responses clustering strongly in the range of 0 to 1, indicating minimal physical strain, which is consistent for digital interfaces. The temporal demand measure showed moderate ratings (M = 3.375), with a wider spread between 2 and 8 on the scale, revealing varied perceptions of time pressure among users.

The distribution for Project Gutenberg reflects a cognitively taxing interface with varied levels of perceived performance success, as the scores were evenly split between the positive and negative ends. The user experience with Project Gutenberg displays mixed response and a lack of intuitiveness in the system design.

Variable	Count	Average	Median	
Physical Demand: How physic...	8	0.875	0.50	
Temporal Demand: How hurrie...	8	3.375	2.50	
Effort: How hard did you have t...	8	4.625	4.00	
Performance: How successful ...	7	5.000	4.00	
Frustration: How insecure, disc...	8	5.000	4.50	
Mental Demand: How mentally...	8	5.125	6.50	
Total (6)	47	3.979	3.00	0 2 4 6 8 10

Fig. 7. Diagram of indicated NASA TLX results for Project Gutenberg.

In contrast to the Project Gutenberg system, the NASA TLX results for Athena shown in Figure 8 depict a system that noted lower cognitive and emotional load across nearly all variables. The average mental demand (M = 2.250), frustration (M = 2.130), and effort (M = 2.75) ratings were all under 3 on the scale of 0-10, with medians supporting this trend. The visual distribution bars were heavily concentrated toward the left, with most responses falling in the 0 to 3 range, representing a generally low perceived task difficulty and emotional strain.

A notable metric, performance, was rated very high (M = 7.880). With the orange distribution showing a major cluster in the 8 to 10 zone, the data suggests participants felt they were highly successful using the Athena system. This stark contrast with Project Gutenberg's more evenly distributed performance ratings suggests Athena may be more user-friendly or intuitive for its tasks.

Variable	Count	Average	Median
Physical Demand: How physic...	8	1.25	0.00
Frustration: How insecure, disc...	8	2.13	1.50
Mental Demand: How mentally...	8	2.25	2.00
Temporal Demand: How hurrie...	8	2.25	1.50
Effort: How hard did you have t...	8	2.75	2.50
Performance: How successful ...	8	7.88	8.50
Total (6)	48	3.[illegible]	2.00

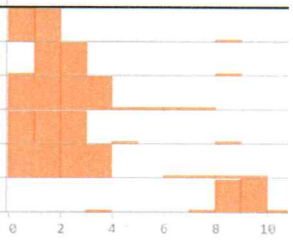

Fig. 8. Diagram of indicated NASA TLX results for Athena.

The data distribution across the TLX dimensions indicates that Project Gutenberg may have posed more cognitive challenges in terms of mental effort and frustration. Whereas, Athena enabled a smoother and more successful user experience. The distribution in the orange cells rightward for Project Gutenberg in demand and effort metrics, and leftward for Athena in the same reinforces this conclusion.

6.3.2 System Usability Scale (SUS). Participants completed the System Usability Scale (SUS) after interacting with each system. The SUS prompts measure aspects such as perceived complexity, ease of use, integration of functions, and anticipated frequency of use. The participant responses range from "Strongly disagree" to "Strongly agree" and reveal notable contrasts between the two systems in terms of user perceptions.

The SUS results for Project Gutenberg demonstrate a wide distribution of responses with a tendency toward neutral or moderate agreement. The prompts associated with complexity and consistency highlighted high levels of concern. Figure 9 showcases that 60% of participants selected "Somewhat agree" for the prompt stating "I found the system unnecessarily complex."The system was reported to largely display inconsistency, suggesting that users encountered confusion or fragmentation while interacting with the interface. Additionally, a substantial proportion of users remained neutral regarding system ease of use and integration of functions, indicating uncertainty or variability in the user experience. Project Gutenberg may require refinement in both interface coherence and cognitive load to better support users.

The results for Athena were completely different from Project Gutenberg, skewing sharply toward positive responses. A majority of participants over 65% in some categories strongly agreed that they would like to use the system frequently, that the system was easy to use, and that they could learn it quickly. In addition, as featured in Figure 10, the participants strongly disagreed with negative statements such as "The system was unnecessarily complex" and "I would need technical support to use the system," which suggests a strong perceived usability and learnability. Athena received many favorable ratings, with over half of users expressing agreement or strong agreement on each prompt. The consistency of positive marks across the overall responses implies a high level of user satisfaction with Athena.

The contrasting distributions in user responses reveal key quantitative insights. Athena's response skew is decisively positive, with consistently high metrics on usability and integration, whereas Project Gutenberg's results revealed

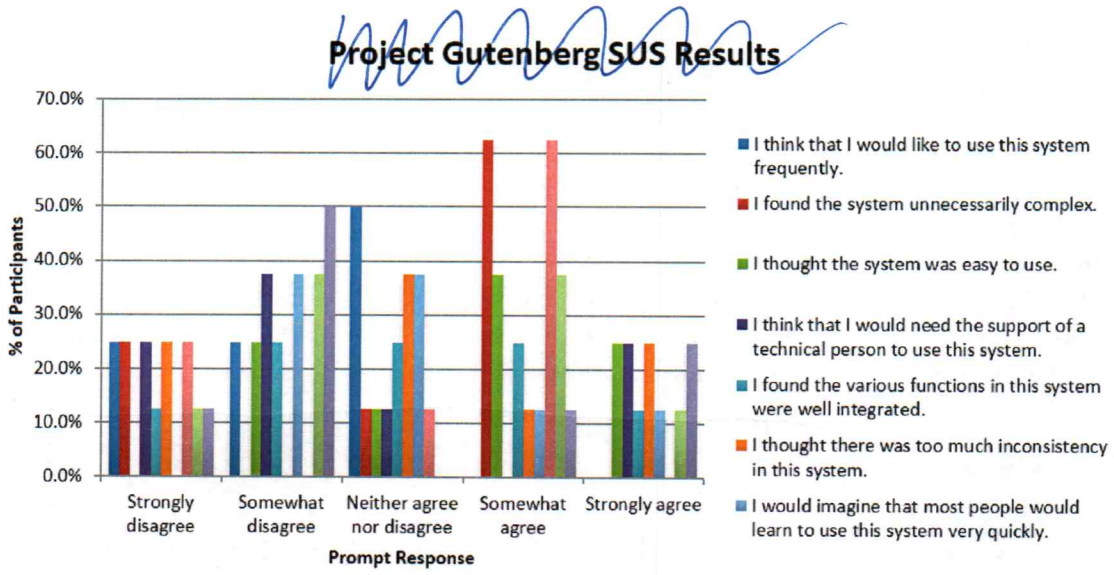

Fig. 9. Bar graph containing results of system usability scale (SUS) for Project Gutenberg.

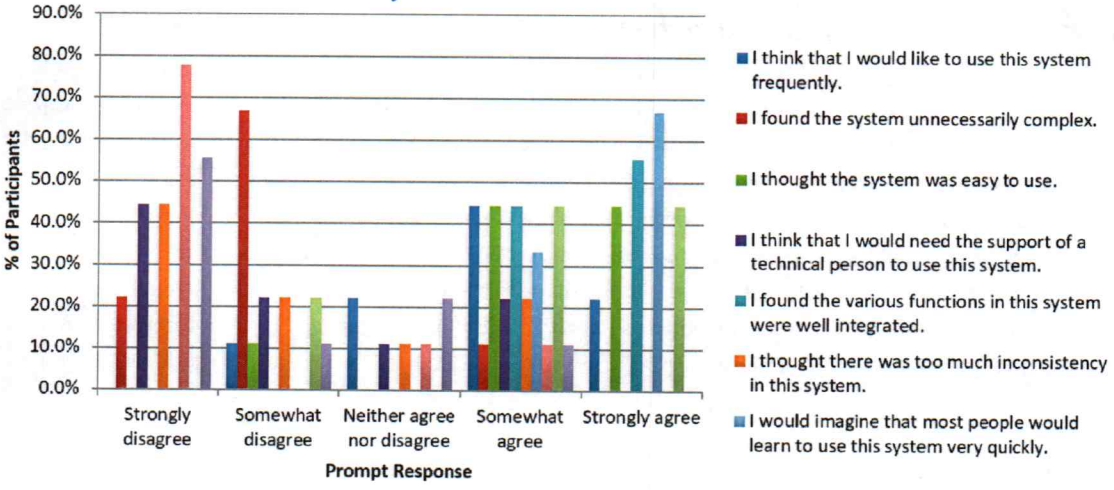

Fig. 10. Bar graph containing results of system usability scale (SUS) for Athena.

concerns over complexity and inconsistency. The difference in perception suggests that Athena delivers a more intuitive and seamless user experience, while Project Gutenberg may benefit from targeted redesigns to address cognitive overload and lacking functionality.

6.3.3 Likert Scale on Trustworthiness. Project Gutenberg, the static digital library system utilized across the course of this study was examined for its perceived trustworthiness by users as found in Figure 11. The data reveals that the majority of users found the system to be highly credible, with 40 percent rating it as "Extremely trustworthy" and another 30% rating it as "Very trustworthy." This indicates that 70% of users had a strong level of confidence in the system. Project Gutenberg's static presentation of literary content presents a reliable interface, particularly for historical and or educational books.

The remaining 30% of responses were divided equally among the lower three categories. "Not at all trustworthy," "Slightly trustworthy," and "Moderately trustworthy," each received 10% of responses from the participants. While the overall perception is to trust Project Gutenberg, a few users had concerns. Regardless, the overall higher trust ratings emphasizes that Project Gutenberg is generally seen as a dependable resource for accessing digital information.

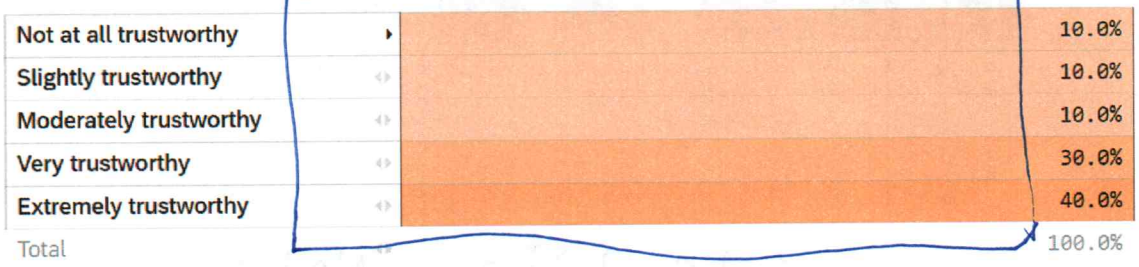

Fig. 11. Diagram of perceived levels of trustworthiness from Project Gutenberg.

Figure 12 displays participant perceptions of the trustworthiness of Athena. Unlike Project Gutenberg, with higher trust ratings, responses for Athena were more centralized across the board with 40% of participants rating the system as "Moderately trustworthy." Users recognized some level of credibility in Athena but there were reservations about the use of artificial intelligence particularly regarding the accuracy or consistency of Athena's responses.

20% of participants found Athena to be "Very trustworthy," and another 20% rated it as "Extremely trustworthy." 40% of users total placed high trust in Athena. The remaining 20% of users were divided evenly between "Slightly trustworthy" and "Not at all trustworthy," representing that Athena does hold some credibility. A notable portion of users may remain cautious, likely due to Athena's unpredictability or AI's general lack of transparency. All in all, Athena's trust ratings reflect a balanced yet skeptical user sentiment compared the higher trustworthiness described for Project Gutenberg.

6.4 Qualitative Results

To analyze qualitative data, we conducted a thematic analysis of open-ended responses from post-task questionnaires and think-aloud transcriptions. This process followed the six-phase method outlined by Nowell et al. [19], including data familiarization, code generation, theme development, theme review, naming, and reporting. Instances of identified themes were tagged and counted across data sources. To support consistency in frequency reporting, we developed a lightweight Python script to automate tallying. The analysis yielded five core themes: usability barriers, trust and credibility, cognitive load, emotional responses, and system affordances. Frequencies of each theme per interface are shown in Table 2.

Fig. 12. Diagram of perceived levels of trustworthiness from Athena.

Table 2. Theme Frequencies by Interface with Descriptions

Theme Name	Code Description	Gutenberg	Athena
Usability Barriers	Challenges interacting with the interface, finding information, or navigating tasks.	23	12
Trust & Credibility	Concerns about the accuracy of information, need for verification, or hallucinations.	6	8
Cognitive Load	Mental effort required to learn, process, or execute tasks using the system.	9	7
Emotional Reactions	Expressed user emotions, such as frustration, confusion, or relief.	10	5
System Affordances	Positive perceptions of interface cues, visuals, or helpful features.	4	7

6.4.1 Usability Barriers and Search Behavior. Usability barriers were the most frequently coded theme, particularly in Project Gutenberg, where participants often struggled with vague or inconsistent search outputs. Search behavior in this static interface was typically keyword-driven, with users copying known titles or author names verbatim (e.g., "Frankenstein by Mary Shelley"). Broader queries often led to confusion. As one participant noted, "I thought the date was when the book was first published, but it's just when they uploaded it." This misunderstanding affected temporal filtering tasks and contributed to frustration.

In several cases, participants used external search engines like Google to troubleshoot (P1, P2, P5), indicating that the static interface offered insufficient guidance. One user explained, "If I couldn't find it, I thought maybe I spelled it wrong. Then I try it again." Participants also found the results redundant or overwhelming, reporting difficulty distinguishing between books with similar metadata. These linear, confirmatory strategies highlighted the absence of system-level support for exploration.

Athena, in contrast, elicited more naturalistic and open-ended interactions. Participants framed prompts conversationally—"Find me a book about dystopian societies"—and took advantage of the multi-turn dialogue system. The graph interface encouraged serendipitous discovery, with one participant likening it to "a virtual bookshelf that made me want to explore more." Multi-book selection, follow-up inquiries, and summarization supported richer, more reflective search behaviors, aligning with the system's design for exploratory use.

However, Athena's interaction design introduced new learning curves. For example, participants sometimes failed to reset the conversational context between tasks, leading to confusing results. As P1 noted, "I was asking about a

new topic, but the AI kept bringing up old stuff." Others forgot to use the "Add Books" function, suggesting that some interface affordances were not immediately discoverable. While Athena facilitated broader exploration, it occasionally suffered from hidden complexity or unclear action pathways.

6.4.2 Trust and Credibility Perceptions. Trust dynamics differed sharply between the two systems. Project Gutenberg generally benefited from the familiarity and perceived neutrality of its static layout. This was also borne out in quantitative results - users reported high baseline trust—70%, rating it as "Very" or "Extremely trustworthy"—despite encountering metadata inconsistencies and a lack of contextual cues. Trust was often reinforced through manual verification: users read book introductions, checked author lifespans, or triangulated information across documents. This kind of active, user-driven validation reflected a default skepticism in the absence of system-generated claims.

Athena evoked more complex trust perceptions. We see this in both our survey data and our observed qualitative data: 40% of users rated it as highly trustworthy, an equal percentage placed it only in the "Moderately trustworthy" category. Several participants voiced skepticism about the completeness or authority of Athena's AI-generated summaries. One asked, "Did it even read the whole book before giving me the answer?" Users noted the absence of source citations and explanations for visual graph relationships. For example, a participant remarked, "There's a line connecting two books, but I have no idea why." These gaps in transparency often prompted users to verify Athena's claims with external resources like Wikipedia.

At the same time, Athena's fluency and interactive design encouraged surface-level trust. The "AI is thinking" animation was perceived as a sign of active processing, not delay. However, this delivery sometimes masked uncertainties in output reliability, illustrating the tension between conversational credibility and grounded transparency in LLM systems.

see mogjan paper

6.4.3 Cognitive Load and Information Processing. Participants in the Project Gutenberg condition described the interface as mentally taxing. The need to formulate precise queries, navigate dense result pages, and decode metadata without semantic clustering contributed to perceived overload. As one user put it, "Too many results and too little organization—it's exhausting." Without scaffolding, cognitive demands shifted to manual sorting and interpretation.

Athena mitigated some of these burdens through structured responses and multimodal navigation. Yet, it introduced new challenges around response density and conversational complexity. Several participants found the generated text blocks verbose or poorly structured. P2 commented, "It's just too much text. I had to dig through it to find what I needed." Additionally, the interplay between chat commands and graph manipulation required users to maintain a mental model of stateful interactions, shifting the cognitive effort from searching to coordinating system behaviors.

6.4.4 Emotional Responses. Emotional reactions were more negative in the Project Gutenberg condition, with participants citing feelings of frustration, confusion, or disengagement. These were especially common during tasks with ambiguous goals or weak search cues. P5 said, "It was hard to know which result was right—everything looked the same."

In Athena, frustration occurred less frequently and tended to relate to expectations rather than outright failure. Users were occasionally disappointed by vague answers or unexplained relationships, but they also expressed appreciation for the interface's responsiveness. One participant noted a sense of relief when a good recommendation emerged: "It actually found something I didn't expect, but I really liked it."

6.4.5 System Affordances. Finally, system affordances—features that signaled available actions—played a crucial role in shaping user experience. In Project Gutenberg, participants appreciated the minimalist layout and the consistent

formatting of metadata, even if search capabilities were rigid. In Athena, the presence of graph visualizations, dynamic summaries, and persistent tags added layers of expressiveness. However, some features were underutilized or misunderstood. For instance, users did not always realize they could select multiple nodes for comparison or trigger follow-up summaries. This suggests that while Athena offers rich functionality, improved onboarding or visual affordance cues may be necessary for full engagement.

To summarize, the thematic analysis of qualitative data revealed that Athena promoted exploratory and expressive interaction patterns, albeit at the cost of interpretive and coordination load. Project Gutenberg offered predictability and transparency but lacked flexibility or support for non-linear discovery. Trust in Athena was more contingent and behaviorally negotiated, with participants balancing fluency against verification. These findings complement the quantitative results and demonstrate trade-offs in conversational interface design for digital libraries.

[Do the] " Gutenberg RAG ♫♫ "

7 Conclusion

We introduced Athena, a Retrieval-Augmented Generation (RAG) conversational system for book discovery in digital libraries, and evaluated its performance against a traditional keyword-based interface using Project Gutenberg. Our within-subjects study revealed that Athena supports more exploratory and engaging search behaviors through natural language interaction, contextual summarization, and knowledge graph visualization.

Participants reported lower cognitive load and higher usability with Athena, as evidenced by TLX and SUS scores. The system's ability to surface thematically relevant books and support multi-turn dialogue encouraged open-ended inquiry and reflective engagement. However, trust in Athena was mixed. While users appreciated its fluency and guidance, some expressed concern over the lack of transparency in the model's reasoning and visual outputs, reflecting broader issues around perceptions LLM hallucination and epistemic trust.

These findings suggest that LLM-powered interfaces can expand the expressive range of digital libraries, but they require careful grounding, interface design, and transparency cues to support sustained user confidence. Future work should explore richer provenance cues, broader user sampling, and long-term engagement patterns to further assess the viability of conversational search in cultural heritage contexts.

8 Acknowledgments

We thank Prof. Frank E. Ritter for his invaluable guidance throughout the development of this project. We are also grateful to our teaching assistant, Noah Gehman, for his support and feedback. Finally, we sincerely appreciate all participants who took part in the evaluation study for their time and insights, which were crucial to this research.

References

[1] Sandeep Avula, Bogeum Choi, and Jaime Arguello. 2022. The Effects of System Initiative during Conversational Collaborative Search. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW1, Article 66 (April 2022), 30 pages. doi:10.1145/3512913

[2] Jürgen Bernard, Debora Daberkow, Dieter Fellner, Katrin Fischer, Oliver Koepler, Jörn Kohlhammer, Mila Runnwerth, Tobias Ruppert, Tobias Schreck, and Irina Sens. 2015. Visinfo: a digital library system for time series research data based on exploratory search—a user-centered design approach. *International Journal on Digital Libraries* 16 (2015), 37–59.

[3] Miriam Boon, Orland Hoeber, Larena Hoeber, Dale Storie, and Veronica Ramshaw. 2024. Exploratory Search and Beyond: A Study of Complex Search Scenarios within a Public Digital Library. *Proceedings of the Association for Information Science and Technology* 61, 1 (Oct. 2024), 44–55. doi:10.1002/pra2.1007

[4] Ted Boren and Judith Ramey. 2000. Thinking aloud: Reconciling theory and practice. *IEEE Transactions on Professional Communication* 43, 3 (2000), 261–278.

[5] John Brooke. 1996. *SUS: A "quick and dirty" usability scale.* Technical Report. 189–194 pages.

where?

[6] Zeljko Carevic, Maria Lusky, Wilko van Hoek, and Philipp Mayr. 2018. Investigating exploratory search activities based on the stratagem level in digital libraries. *International Journal on Digital Libraries* 19 (2018), 231–251.

[7] Yang Deng, Lizi Liao, Zhonghua Zheng, Grace Hui Yang, and Tat-Seng Chua. 2024. Towards Human-centered Proactive Conversational Agents. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Washington DC, USA) *(SIGIR '24)*. Association for Computing Machinery, New York, NY, USA, 807–818. doi:10.1145/3626772.3657843

[8] K. Anders Ericsson and Herbert A. Simon. 1998. How to study thinking in everyday life: Contrasting think-aloud protocols with descriptions and explanations of thinking. *Mind, Culture, and Activity* 5, 3 (1998), 178–186.

[9] Martin Gerlach and Francesc Font-Clos. 2020. A standardized Project Gutenberg corpus for statistical analysis of natural language and quantitative linguistics. *Entropy* 22, 1 (2020), 126.

[10] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Human Mental Workload*, Peter A. Hancock and Najmedin Meshkati (Eds.). Advances in Psychology, Vol. 52. North-Holland, 139–183. doi:10.1016/S0166-4115(08)62386-9

[11] Zexuan Ji, Nayeon Lee, Rita Frieske, Teng Yu, Dan Su, Yanran Xu, Eric Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys (CSUR)* 55, 12 (2023), 1–38.

[12] Ming Jiang, Ryan C Dubnicek, Glen Worthey, Ted Underwood, and J. Stephen Downie. 2022. A prototype gutenberg-hathitrust sentence-level parallel corpus for OCR error analysis: pilot investigations. In *Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries* (Cologne, Germany) *(JCDL '22)*. Association for Computing Machinery, New York, NY, USA, Article 45, 5 pages. doi:10.1145/3529372.3533298

[13] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kulkarni, Sebastian Riedel, Luke Zettlemoyer, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, Vol. 33. 9459–9474.

[14] Ewa Luger and Abigail Sellen. 2016. "Like Having a Really Bad PA": The Gulf between User Expectation and Experience of Conversational Agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 5286–5297. doi:10.1145/2858036.2858288

[15] Gary Marchionini. 2006. Exploratory search: from finding to understanding. *Commun. ACM* 49, 4 (April 2006), 41–46. doi:10.1145/1121949.1121979

[16] Lori McCay-Peet, Anabel Quan-Haase, and Dagmar Kern. 2015. Exploratory search in digital libraries: a preliminary examination of the use and role of interface features. In *Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community* (St. Louis, Missouri) *(ASIST '15)*. American Society for Information Science, USA, Article 70, 4 pages.

[17] Fengran Mo, Kelong Mao, Ziliang Zhao, Hongjin Qian, Haonan Chen, Yiruo Cheng, Xiaoxi Li, Yutao Zhu, Zhicheng Dou, and Jian-Yun Nie. 2024. A Survey of Conversational Search. https://arxiv.org/abs/2410.15576

[18] Alexander Tobias Neumann, Yue Yin, Sulayman Sowe, Stefan Decker, and Matthias Jarke. 2025. An LLM-Driven Chatbot in Higher Education for Databases and Information Systems. *IEEE Transactions on Education* 68, 1 (2025), 103–116. doi:10.1109/TE.2024.3467912

[19] Lorelli S Nowell, Jill M Norris, Deborah E White, and Nancy J Moules. 2017. Thematic analysis: Striving to meet the trustworthiness criteria. *International Journal of Qualitative Methods* 16, 1 (2017), 1–13.

[20] Andrea Papenmeier, Dagmar Kern, Daniel Hienert, Alfred Sliwa, Ahmet Aker, and Norbert Fuhr. 2021. Starting Conversations with Search Engines - Interfaces that Elicit Natural Language Queries. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval* (Canberra ACT, Australia) *(CHIIR '21)*. Association for Computing Machinery, New York, NY, USA, 261–265. doi:10.1145/3406522.3446035

[21] Sachin Pathiyan Cherumanal, Lin Tian, Futoon M. Abushaqra, Angel Felipe Magnossão de Paula, Kaixin Ji, Halil Ali, Danula Hettiachchi, Johanne R. Trippas, Falk Scholer, and Damiano Spina. 2024. Walert: Putting Conversational Information Seeking Knowledge into Action by Building and Evaluating a Large Language Model-Powered Chatbot. In *Proceedings of the 2024 Conference on Human Information Interaction and Retrieval* (Sheffield, United Kingdom) *(CHIIR '24)*. Association for Computing Machinery, New York, NY, USA, 401–405. doi:10.1145/3627508.3638309

[22] Project Gutenberg. 2024. Project Gutenberg. https://www.gutenberg.org/. Accessed: 2025-05-01.

[23] Minjin Rheu, Ji Youn Shin, Wei Peng, and Jina Huh-Yoo and. 2021. Systematic Review: Trust-Building Factors and Implications for Conversational Agent Design. *International Journal of Human–Computer Interaction* 37, 1 (2021), 81–96. doi:10.1080/10447318.2020.1807710 arXiv:https://doi.org/10.1080/10447318.2020.1807710

[24] A. B. Riddell. 2022. Reliable editions from unreliable components: estimating ebooks from print editions using profile hidden markov models. In *Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries* (Cologne, Germany) *(JCDL '22)*. Association for Computing Machinery, New York, NY, USA, Article 24, 5 pages. doi:10.1145/3529372.3533292

[25] Frank E. Ritter, Graham D. Baxter, and Elizabeth F. Churchill. 2014. Cognition: Human–Computer Communication. In *Foundations for Designing User-Centered Systems*. Springer, London, 201–223. doi:10.1007/978-1-4471-5134-0_7

[26] Frank E. Ritter, Graham D. Baxter, and Elizabeth F. Churchill. 2014. Cognition: Memory, Attention, and Learning. In *Foundations for Designing User-Centered Systems*. Springer, London, 123–164. doi:10.1007/978-1-4471-5134-0_5

[27] Anuschka Schmitt, Thiemo Wambsganss, and Jan Marco Leimeister. 2022. Conversational Agents for Information Retrieval in the Education Domain: A User-Centered Design Investigation. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 486 (Nov. 2022), 22 pages. doi:10.1145/3555587

[28] Chirag Shah and Emily M. Bender. 2022. Situating Search. In *Proceedings of the 2022 Conference on Human Information Interaction and Retrieval* (Regensburg, Germany) *(CHIIR '22)*. Association for Computing Machinery, New York, NY, USA, 221–232. doi:10.1145/3498366.3505816

cool. generally [25, ch.5] [25, ch.7]

[29] Rao Shen, Naga Srinivas Vemuri, Weiguo Fan, Ricardo da S. Torres, and Edward A. Fox. 2006. Exploring digital libraries: integrating browsing, searching, and visualization. In *Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries* (Chapel Hill, NC, USA) *(JCDL '06)*. Association for Computing Machinery, New York, NY, USA, 1–10. doi:10.1145/1141753.1141755

[30] Xin Sun, Yunjie Liu, Jan De Wit, Jos A. Bosch, and Zhuying Li. 2024. Trust by Interface: How Different User Interfaces Shape Human Trust in Health Information from Large Language Models. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '24)*. Association for Computing Machinery, New York, NY, USA, Article 344, 7 pages. doi:10.1145/3613905.3650837

[31] Xin Sun, Rongjun Ma, Xiaochang Zhao, Zhuying Li, Janne Lindqvist, Abdallah El Ali, and Jos A. Bosch. 2024. Trusting the Search: Unraveling Human Trust in Health Information from Google and ChatGPT. arXiv:2403.09987 [cs.HC] https://arxiv.org/abs/2403.09987

[32] Ryen W. White. 2024. Advancing the Search Frontier with AI Agents. *Commun. ACM* 67, 9 (Aug. 2024), 54–65. doi:10.1145/3655615

[33] An Zhang, Yang Deng, Yankai Lin, Xu Chen, Ji-Rong Wen, and Tat-Seng Chua. 2024. Large Language Model Powered Agents for Information Retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Washington DC, USA) *(SIGIR '24)*. Association for Computing Machinery, New York, NY, USA, 2989–2992. doi:10.1145/3626772.3661375

Not consistent but nearly correct

A Task Materials

Task #	Prompt	Task Type	Target Behavior
1	Find and access the first chapter of *Frankenstein* by Mary Shelley.	Lookup	Known-item retrieval; title search
2	Find a book written by Jules Verne that was published before 1880 and download it in plain text format.	Lookup	Metadata filtering; structured search
3	You want to read a short science fiction story written in the 1800s. Find one that interests you.	Exploratory	Genre-based exploratory search; soft constraints
4	Find a book from the 1800s that explores the idea of a utopian or dystopian society. Choose one you'd like to read.	Exploratory	Thematic exploration; subjective interest
5	You're putting together a reading list on political philosophy. Find two books that you think would pair well together.	Exploratory	Synthesis of content; reasoning about compatibility
6	Find out who wrote *The Prince*, and briefly explain what the book is about and who it is written for.	Trust	Factual lookup + interpretive summary
7	Find a novel that deals with the effects of poverty in 19th-century society. Who wrote it and when?	Trust	Theme + historical filtering; LLM grounding test
8	What is considered the first feminist novel in English literature? Find it and tell us why it's described that way.	Trust + Exploratory	Bias surfacing; gender inference; summary interpretation

Table: Tasks with Prompts, Task Types, and Target Behaviors

Received 5 May 2025