Evaluating Non-AI Experts' Interaction with AI: A Case Study In Library Context

Qingxiao Zheng
Computer Science and Engineering
University at Buffalo, SUNY
Buffalo, USA
qingxiao@buffalo.edu

Minrui Chen
Informatics
University of Illinois
Urbana-Champaign
Champaign, USA
minruic2@illinois.edu

Hyanghee Park
Information Sciences
University of Illinois
Urbana-Champaign
Champaign, USA
hyanghee@illinois.edu

Zhongwei Xu
Computer Science and Engineering
University of Michigan
Ann Arbor
Michigan, USA
xzw@umich.edu

Yun Huang
Information Sciences
University of Illinois
Urbana-Champaign
Champaign, USA
yunhuang@illinois.edu

Abstract

Public libraries in the U.S. are increasingly facing labor shortages, tight budgets, and overworked staff, creating a pressing need for conversational agents to assist patrons. The democratization of generative AI has empowered public service professionals to develop AI agents by leveraging large language models. To understand the needs of non-AI library professionals in creating their own conversational agents, we conducted semi-structured interviews with library professionals (n=11) across the U.S. Insights from these interviews informed the design of AgentBuilder, a prototype tool that enables non-AI experts to create conversational agents without coding skills. We then conducted think-aloud sessions and follow-up interviews to evaluate the prototype experience and identify the key evaluation criteria emphasized by library professionals (n=12) when developing conversational agents. Our findings highlight how these professionals perceive the prototype experience and reveal five essential evaluation criteria: interpreting user intent, faithful paraphrasing, proper alignment with authoritative sources, tailoring the tone of voice, and handling unknown answers effectively. These insights provide valuable guidance for designing AI-supported *"end-user creation tools"* in public service domains beyond libraries.

CCS Concepts

• **Human-centered computing** → **Interaction paradigms**; **Usability testing**.

Keywords

End-User AI Creation Tool, User Experience, Large language models, Generative AI, Public Service



This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.
CHI '25, Yokohama, Japan
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1394-1/2025/04
https://doi.org/10.1145/3706598.3714219

ACM Reference Format:
Qingxiao Zheng, Minrui Chen, Hyanghee Park, Zhongwei Xu, and Yun Huang. 2025. Evaluating Non-AI Experts' Interaction with AI: A Case Study In Library Context. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26-May 1, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 20 pages. https://doi.org/10.1145/3706598.3714219

1 Introduction

Conversational agents (CAs) are increasingly utilized in public services such as healthcare, safety, and law enforcement. They perform tasks like documenting cases for sexual assault survivors [87] and conducting phone check-ins with socially isolated individuals as health interventions [51]. The rapid adoption of CAs has heightened interest in service design within the Human-Computer Interaction (HCI) community [8, 66, 96, 100, 131, 139]. However, developing CAs for diverse public services typically requires collaboration between AI developers and service providers due to the need for advanced coding and AI expertise [124]. Additionally, in high-stakes applications such as case reporting, specialized User eXperience (UX) professionals are typically responsible for designing these CAs to minimize unintended consequences [51, 87].

Recent advancements in generative AI have empowered non-AI experts to create and customize CAs using large language models (LLMs) [37]. This development marks a new era in "end-user AI creation tools" [72, 92]. Prior studies demonstrate that non-AI experts have leveraged LLMs to tailor conversational AI agents for tasks such as handling customer inquiries, data analytics, and writing support [46, 53], thus expanding AI's role in augmenting human capabilities [17, 27].

Among various application domains, public libraries are a notable setting where non-AI experts require significant customization of CAs [2, 116]. Library professionals, such as librarians, help desk staff, and programming coordinators, benefit from CAs for several reasons. Staffing shortages often lead to high volumes of customer interactions, especially outside of regular hours. CAs can provide 24/7 assistance, reducing the workload on staff [29]. Additionally, many inquiries are routine and do not require specialized expertise; for instance, 47% of library questions involve general information

or technical support (e.g., service point locations, equipment assistance, and policy details) [33, 85]. Furthermore, libraries operate in relatively low-stakes environments, where the consequences of CA errors are less severe compared to high-stakes services such as emergency response [51, 87].

Although LLM-based platforms provide opportunities to enable non-AI library professionals to customize CAs to address labor shortages and scale services, CA creation remains challenging. First, although many library inquiries are repetitive and well-suited for CAs, library professionals must tailor responses to suit their unique communities, which encompass diverse populations, cultures, and needs [47]. Additionally, public libraries have expanded their roles beyond traditional resource sharing to include a diverse range of educational programs, digital access services, and community engagement initiatives [1, 50, 105]. Moreover, tailoring CAs to each library's specific values, policies and information requires technical expertise, which many local libraries lack due to budget constraints.

However, to date, limited research has explored the design requirements of non-AI expert library professionals when creating CAs using no-code AI systems. Consequently, the evaluation criteria for the implementation of these CAs in public libraries are still poorly understood. *This paper aims to identify new UX evaluation criteria that empower non-AI experts to effectively utilize LLM technologies to customize their services.*

To achieve this, we conducted two studies with a total of 23 library professionals who had no extensive coding expertise but interacted directly with patrons in their jobs and asked them to envision a creation platform and evaluate their customized CAs. In Study 1, through interviews with 11 library professionals, we identified three design goals that future end-user creation tools should incorporate when helping non-AI experts create CAs: (1) control over the AI's service scope, (2) alignment with expectations and synchronization through collaboration, and (3) the ability to handle a diverse range of patrons. Based on these findings, we developed the AgentBuilder, a platform that enables non-AI experts to define, customize, and evaluate LLM-based CAs. To further assess our findings and provide deeper insights into human-AI interaction, particularly for non-AI experts, we conducted Study 2. We recruited 12 additional library professionals to evaluate the platform using a think-aloud protocol followed by a semi-structured interview. We identified five UX evaluation criteria that are particularly important to library professionals as CA creators: interpretation of users' intent, proper direction to the right source for accurate information, identity-based emotive expression, faithful paraphrasing, and inability to answer. Our approach emphasizes controllability, mutual learning, automatic evaluation, and addressing the specific concerns and needs that non-AI experts face when building CAs for their communities [6].

Our contributions to HCI and AI democratization in service contexts are threefold: (1) Through an empirical case study, we identified three novel design needs specific to non-AI expert service providers; (2) We introduced a research prototype addressing these needs and examined its alignment (or misalignment) with library professionals' practices, inspiring future work in designing end-user creation tools; (3) We distinguished key UX evaluation criteria tailored for non-AI experts CA creators, diverging from prior work that primarily focuses on UX metrics for CA users.

2 Related Work

This section reviews related work on democratizing AI in service environments, the use of conversational agents in libraries, and the rise of no-code AI platforms powered by large language models (LLMs). Together, these areas illustrate both the opportunities and challenges of integrating AI in low-stakes contexts and underscore the importance of accessible, adaptable, and user-sensitive solutions for non-AI experts.

2.1 Democratizing AI in Service Contexts

AI technologies are increasingly adopted in public services to enhance efficiency. For instance, the Australian Taxation Office's "Alex" chatbot has achieved time and cost savings [42]. During the COVID-19 crisis, conversational agents provided mental health support [103], improved citizen interactions with instant information [101], countered misinformation [59], and gathered data on user needs [54]. These applications demonstrate AI's potential to improve public service delivery across sectors.

The User eXperience (UX) design of AI varies significantly between high-stakes and low-stakes environments [110, 129]. In high-stakes contexts such as hospitals and emergency services, where errors can have severe consequences (e.g., risks to patient safety or emergency response failures) [110], AI designs focus on collaboration with experts [87, 125], building trust [13, 79, 116], and minimizing risks [84, 99]. Maintaining human control and ensuring AI reliability are critical for confident decision-making in such scenarios [51, 55, 86, 87, 127]. On the other hand, low-stakes environments, such as libraries, community centers, and customer service helplines, emphasize reducing staff workload, increasing service availability, and addressing specific community needs [64]. UX designs in these areas explore how AI can enhance service delivery by prioritizing accessibility, usability, and efficiency over accuracy and risk mitigation [130].

A key distinction between these environments lies in the design approach: In high-stakes contexts, system design is typically focused on **AI created "for" experts** to meet rigorous domain standards [51, 87]. Conversely, in low-stakes environments, there is a growing push to **design AI "by" non-AI experts**, enabling greater customization and more accessible solutions tailored to specific community needs [61]. Democratizing AI technologies empowers non-experts [62, 78, 95], such as librarians and community service providers, to create and adapt AI tools without requiring extensive technical expertise. *While this approach promises to extend AI benefits to underserved sectors, it also reveals a gap in research on supporting non-experts in adopting and customizing AI [61].* Bridging this gap is crucial to ensuring equitable access to AI's transformative potential across public services, particularly in resource-constrained environments.

2.2 Conversational Agents in Libraries

Public libraries, as vibrant community hubs, play a crucial role in addressing societal challenges through educational programs, digital access services, and community engagement initiatives [1, 50, 105]. As with other public sectors, public libraries cater to diverse communities, encompassing a wide range of ages and user types [40, 65, 68]. This diversity creates significant challenges for library professionals

[24, 43], who must find scalable and accessible solutions to maintain high-quality services [76, 104].

Existing solutions in libraries for managing user inquiries often rely on rule-based systems or human professionals. Rule-based systems are designed to handle repetitive tasks, such as answering frequently asked questions or managing reservations, using pre-programmed responses. While effective for straightforward tasks, these systems cannot handle complex or context-sensitive queries. On the other hand, many libraries offer chat features where patrons interact directly with human staff for more nuanced support [49, 65]. However, staffing limitations, particularly during off-hours or holidays, can lead to reduced service availability and unmet user needs [89].

AI-powered conversational agents offer more advanced solutions by addressing limitations in both rule-based systems and human-staffed chat services. Operating 24/7, these agents enable users to access assistance at any time, reducing staff workloads while maintaining consistent service availability [29]. AI-enabled services provide real-time support for basic inquiries during staff shortages [121], freeing library staff to focus on more complex and value-added tasks [70]. Furthermore, they enhance service accessibility for users with diverse needs, such as individuals with low literacy or disabilities [71]. AI agents can also collect and analyze multilingual community input, allowing libraries to identify and address emerging challenges cost-effectively [22].

However, the adoption of conversational agents in libraries involves unique challenges. Libraries serve diverse communities with distinct needs influenced by local demographics, culture, and services [41, 65, 68]. Data privacy is a significant concern, as patrons expect secure handling of their information, which is a standard not always met by commercial AI solutions [15, 113]. Customizable, privacy-conscious AI systems that are aligned with library values are essential to meet these expectations. Yet, many public libraries face funding cuts [28] and cannot afford technical teams to develop tailored solutions. They often rely on support from regional or state libraries [94], which can delay implementation. Some library councils are recognizing the importance of AI integration [117], and tools like large language models now allow library staff to create conversational agents independently and affordably [38]. *These developments highlight the need for customizable, cost-effective AI solutions that prioritize privacy and align with the mission of public libraries.*

2.3 End-User Creation Tools Supported by Large Language Models (LLMs)

No-code AI development platforms, or end-user creation tools powered by LLMs, represent a significant advancement in democratizing AI development by enabling individuals with limited programming knowledge to create and deploy AI systems, such as OpenAI's ChatGPT, Microsoft's Power Virtual Agents, and Google's Dialogflow. These platforms allow non-AI experts to build and customize conversational agents, addressing barriers of cost and technical expertise [38, 80, 135]. They now extend AI development capabilities to non-AI experts, promoting inclusivity and innovation in designing conversational agents [25].

Despite the growing availability of these tools, significant gaps remain in supporting non-AI experts, such as librarians, in creating conversational agents that meet their specific needs. Many of the existing tools lack necessary customization features, privacy assurances, and context-sensitive adaptability that are crucial in settings like libraries [21, 108]. Previous literature on conversational agents design provides various features relevant to AI agents, including aspects such as deployment, role, functionality, dialogue, engagement, and evaluation (e.g., [7, 20, 138]), which we synthesized in the taxonomy shown in **Table 1**. These resources could be used for guiding non-AI experts to design and implement customized conversational agents; however, they are often not accessible or easily applicable to those without programming skills [126, 128]. Implementing features that effectively address these design contexts requires direct access to machine learning models and extensive back-end data [82, 132]. Without this expertise, it becomes challenging to translate high-level design implications into practical, tangible implementations [4, 114].

Moreover, non-AI experts face a range of barriers when attempting to create or use machine learning models, including challenges in defining tasks, selecting appropriate tools, integrating components, and evaluating outcomes [58]. These barriers are particularly evident for non-programmers who lack formal training in AI development. Researchers such as Zamfirescu-Pereira et al. [133] have documented the complexities encountered by novice users when interacting with machine learning technologies, especially in terms of crafting effective prompts for LLMs. To address these challenges, recent research and applications have begun exploring the use of LLMs to support the creation of customized AI agents designed by non-AI experts [18, 38, 39, 74], enabling laypersons to independently design these agents without relying on technical developers, as indicated by previously identified challenges [97, 98]. An example of designing with non-AI experts is the CloChat interface developed by Ha et al. [38], which allows users to define and customize a conversational agent's persona, enhancing emotional bonds between humans and AI.

While such tools show promise, they primarily support the initial phase of defining a conversational agent's persona, leaving other critical challenges unresolved. These include alignment issues related to LLM outputs, such as factual inaccuracies, biases, and inappropriate responses [10, 26, 119]. Despite proposed solutions, ranging from moderation [75] and output control [122] to evaluation frameworks [132, 133] and ethical AI design guidelines [31], ensuring that AI agents consistently meet users' expectations remains a persistent issue. This is particularly pronounced for non-experts operating in public service domains, especially in low-stakes environments like libraries. There is limited understanding of how to design for specific needs within service contexts, what aspects of a conversational agent should be customized, and how these customizations can be seamlessly integrated into existing service workflows. Addressing these gaps is crucial to making AI tools both effective and practical for such environments.

In summary, this section reviews related work on democratizing AI in service contexts, focusing on empowering library professionals to create conversational agents using no-code approaches

Design Context	Contextual Questions	Citations
1. Size	How many people should the AI serve in a single session?	[7, 19, 138]
	How many AI agents will interact with people simultaneously?	[7, 138]
2. Deployment	What modes of communication are provided for interactions?	[7, 48, 83]
	Where will the AI system be physically or digitally deployed?	[57, 83]
	Should the AI be capable of responding in multiple languages?	[83]
3. Role	What is the intended role of AI in the public service sector?	[7, 83, 107]
	How should the AI collaborate with public service employees?	[32, 36]
4. Functionality	What specific public service domain is the AI targeted for?	[81, 83, 93, 106]
	Should the AI focus on specific tasks or provide general assistance?	[7, 34, 48, 81, 83, 93]
	What is the primary purpose of implementing AI in public service?	[81, 83]
5. Dialogue	What is the expected duration for each interaction?	[83, 138]
	How many dialogues are expected to be exchanged per interaction session?	[19, 36]
	Should the AI remember and retrieve past interactions with individual users?	[83]
6. Engagement	Should the user or AI initiate the conversation?	[19, 83]
	Will the AI offer personalized responses based on user profiles?	[81, 83]
	Are there plans to incorporate gamification elements to engage people?	[83]
	What is the desired depth of dialogue for the AI?	[36]
7. Escalation	Will people connect with a human agent for complex issues?	[83, 138]
	Should the AI understand and respond to users' emotional states?	[7, 83]
8. Humanness	How should the AI be visually represented with an animated avatar?	[7, 83]
	What type of tone (or voice if using avatar) should the AI use?	[14, 91, 106, 106]
9. Maintenance	How will the AI system's knowledge base be developed and updated?	[7, 48, 81]
	Are there plans to integrate AI with external services or databases?	[7, 48, 81, 83]
10. Evaluation	What metrics do you want to use to evaluate your AI agent?	[138]

Table 1: Conversational Agent Design Dimensions: A Question List Synthesized from Literature.

supported by LLMs. Despite significant advancements in high-stakes AI, there remains a gap in making these technologies accessible to non-technical users in low-stakes environments like libraries [61, 126]. This study aims to bridge this gap by examining how library professionals can leverage conversational agents through no-code creator-facing tools. By doing so, our study shifts focus from the questions, *"What challenges do end-users face when interacting with conversational agents built by professional AI developers?* [51, 87]" or *"How do non-AI experts co-create AI with data scientists?"* to explore *"What are the needs and evaluation criteria that arise when non-AI experts— who are **both** AI creators and users— design, create, and evaluate their own conversational agents?"*

3 Study 1. Identifying Library Professionals' Needs

In Study 1, we conducted semi-structured interviews to identify the needs and concerns library professionals have when creating CAs for their public service, using no-code platforms supported by LLMs. As a result, our participants' needs and concerns were translated into three design goals which we we used to guide the design of our no-code prototype discussed in Section 4.

3.1 Method

3.1.1 Participants. Recruitment was conducted using email addresses obtained from publicly available state library websites. Many states provide directories listing library contacts, from which we compiled over 12,500 email addresses. Using this list, we sent recruitment invitations that included a project description outlining the study's goals and novelty, as well as instructions for scheduling an online interview. Participants were encouraged to share the invitation with library-affiliated colleagues through snowball sampling. For the screening criteria, participants were required to have direct interactions with patrons and to have answered their queries, although the types of queries varied. The study was further restricted to service professionals currently affiliated with U.S. public libraries. The study was reviewed and approved by our university's institutional review board, and each participant received $50 for gratitude.

Table 2 presents information on participants' service experience, AI experience, current service practices, affiliated library information, and interactions with patrons. We recruited professionals from various roles, such as librarians, library staff, supervisors, administrators, and information technology personnel. Despite variations in job titles among libraries, all participants shared the common responsibility of directly answering patron queries. The types of questions they addressed varied based on their roles, with IT staff (e.g., P2 and P3) handling technology-related inquiries, program coordinators (e.g., P4 and P6) focusing on questions related to specific events and activities, and directors or managers (e.g., P7, P8, and P10) often addressing broader operational or policy inquiries. The diversity of our participants' roles provides valuable insights into how public library professionals handle a wide range of patron queries, ensuring that we collected perspectives from individuals with extensive first-hand experience.

Most participants had extensive experience serving in public libraries, with an average service duration of more than 9 years. Seven participants (63.6%) reported over 10 years of service. Nearly all participants (10, 90.9%) had used LLM-based applications before the study. However, as of the study date, OpenAI had not yet launched customized GPT, meaning none of the participants had

ID	Job Title	Years of Service	LLMs Exp	Coding Exp	Library Location	Library Size	# of IT	Ways of Interaction with Patrons
P1	Specialist	2 years	Yes	No	Texas	Large	4	FtF, Email
P2	IT head	10 years	Yes	Yes	Wisconsin	Medium	3	FtF, Telephone, Email
P3	IT support	8 years	No	No	Washington	Small	External	FtF, Email
P4	Program Coordinator	10 years	Yes	No	California	Medium	4	FtF, Telephone, Email
P5	Branch manager	15 years	Yes	No	Oklahoma	Small	External	FtF, Telephone, Website
P6	Program Coordinator	4 years	Yes	No	Texas	Large	External	All methods
P7	Director	22 years	Yes	No	Virginia	Medium	External	FtF, Website
P8	Library director	15 years	Yes	No	Minnesota	Small	External	FtF, Telephone
P9	Youth librarian	3 years	Yes	No	Montana	Small	External	FtF, Telephone, Email
P10	Library manager	16 years	Yes	No	Pennsylvania	Medium	External	FtF, Email, Website
P11	Director	12 years	Yes	No	Illinois	Small	2	FtF, Telephone, Email

Table 2: Participants' Background Information in Study 1: "LLM Exp" indicates prior use of LLM-based applications in daily life, "coding exp" indicates coding experience, "library size" refers to an estimated number of service providers within the participants' affiliated library, and "# of IT" represents the number of IT support staff.

experience building their own LLM-based conversational agents or incorporating such applications into their workflows at that time. Regarding the coding experience, only one participant (9.1%) reported experience with coding tasks for simple system development. Notably, the number of IT support staff at participants' libraries was relatively small, with the majority (7, 63.6%) relying on external technicians for development support.

In terms of how they interacted with patrons, traditional methods remained prevalent. However, online chat with patrons showed a growing trend, with 4 out of 11 participants utilizing website-based chat to address patrons' queries.

3.1.2 Procedure. We conducted semi-structured interviews in which participants took on the role of AI creators. Following an initial briefing on their background (e.g., *"Can you describe your role in the library?"* and *"How do you typically interact with patrons?"*), participants described their current practices and envisioned how AI could help (e.g., *"What challenges do you face in providing makerspace services you mentioned?"* and *"How do you think AI could help address these challenges?"*).

We used two design materials to gauge their needs for an end-user AI creation tool. On the one hand, since none of the participants had prior experience designing conversational agents, we developed Table 1 to prompt them to adopt the role of a conversational agent creator. For example, we asked questions such as e.g., *"What features would make a conversational agent creation tool easier for you to use?"* and *"What would you need to be able to customize conversational agent responses to your community's specific needs?"*. On the other hand, we used ChatGPT as another interactive design material to showcase AI's Q/A capability and to facilitate dynamic exploration. Because ChatGPT was the most familiar AI tool for participants at the time, we asked questions such as *"How would you like to personalize the conversational agent's responses for different patron groups?"* and *"What kinds of support or features would help you feel comfortable creating a conversational agent on your own?"*. This encouraged participants to try prompts such as *"You are a friendly librarian answering hands-on questions in a makerspace."* These two materials were used to stimulate brainstorming on AI's potential to improve services from the creator's perspective, rather than that of simple AI users. The focus was on gathering insights into specific scenarios where AI could support library services and on understanding design goals drawn from participants' concerns and needs regarding a conversational agent creation tool. Each interview lasted 60 to 90 minutes and was conducted via Zoom from June to July 2023.

3.1.3 Data Analysis. Two researchers conducted the interviews and used Otter.ai for transcription to familiarize themselves with the material. They reviewed each transcript line by line, noting the context of participants' comments. Each researcher independently coded half of the transcripts using open coding to capture emerging viewpoints, followed by cross-checking to ensure consistency. The data were analyzed using a thematic analysis approach [11], with iterative discussions held until no new insights emerged. The analysis focused on uncovering recurring themes related to *design goals* in creating conversational agents drawn from their needs and concerns regarding a platform where they can build their own conversational agents.

3.2 Findings

Our participants first highlighted various challenges in their current work practices. Common challenges include staffing constraints in appointment-centric programs (e.g. maker spaces, volunteer training), deficiencies in expertise related to specific tools (e.g. 3D printers), repetitive queries at reference desks, and the need for after-hours support. These challenges highlight areas where conversational agents can be effectively employed. However, rather than adopting existing commercial products designed for deploying conversational agents, participants expressed reluctance to use such solutions. Relying on these products introduces privacy concerns, such as uploading library-specific information, saving library professionals' interactions with patrons, storing patrons' personal data on external servers, and the potential misuse or unauthorized access to sensitive data. Participants expressed a desire for a platform that would enable them to easily create, customize, and assess their conversational agents, ensuring greater control and autonomy in the process. Based on our interview results, we identified three design goals for potential solutions.

3.2.1 ***Design Goal 1: Retain control over the AI's service scope by defining and establishing AI's service boundaries***. Participants envisioned AI as an extension of their teams, but clearly within operational boundaries shaped by human expertise. Rather than having the AI operate without limits, they wanted a well-defined scope that ensures that library professionals retain ultimate control over its role and the information it provides.

For instance, multiple participants expressed interest in assigning the conversational agent a specific "role" to support knowledge sharing across the library. For example, P8 imagined a centralized hub for procedural documents, policy manuals, and technical equipment guidelines, stating *"I hope that everything... could be in one place to build up internal knowledge or policy sharing,"* which would enable the conversational agent to connect patrons with relevant information and direct them to the appropriate area. Participants also recognized that the AI might handle more complex tasks, such as organizing *"internal documentation"* or serving as a *"binder"* of institutional knowledge. These broader capabilities, however, raised questions about how expansive the AI's responsibilities should be.

Maintenance emerged as a central theme shaping the scope of the AI's knowledge base and services. While the library setting may be relatively low-stakes, participants recognized that patrons could still misconstrue AI outputs as authoritative. P5 warned of the possibility that users might *"use [the information] in place of legal advice,"* highlighting the need for strict boundaries. In other words, AI-generated content must remain tethered to *user-defined* knowledge sources, preventing it from venturing into unauthorized domains or providing unverified guidance.

3.2.2 ***Design Goal 2: Ensure AI's responses align with library professionals' expectations and synchronize them through a collaborative process.*** Participants expressed a vision of human-AI collaboration in which AI augments, rather than replaces, human capabilities. While they welcomed AI's potential to enhance library operations and bridge gaps in human availability, they collectively emphasized the critical role of humans as "gatekeepers," ensuring that the system delivers accurate, unbiased, and contextually appropriate information.

Additionally, Participants highlighted the importance of linguistic nuance, cultural context, and local jargon. P8 emphasized the importance of recognizing and interpreting *"local slang or jargon"*, suggesting that AI should be capable of understanding and appropriately responding to terminology commonly that is commonly used in library contexts. This requires AI to be familiar with technical terms, such as those related to specific tools (e.g., 3D printers), and with abbreviations that patrons or librarians might use in their queries. P5 echoed this point, stating that *"We should be able to teach the conversational agent the way we speak; every interaction should feel natural."* P3 noted that patrons may sometimes fail to use the correct terms, and the unpredictability of their queries poses challenges for AI in understanding them. P3 further explained, *"Elderly people often say, 'I want to photocopy a photo to my flash drive.'"* In their example, a patron may actually want to scan a photo, but due to limited technology literacy, they use incorrect terminology. *"That's not photocopying but scanning or something similar. They use that language because they grew up with those terms. [Without the ability to understand the context], conversational agents will suddenly be lost completely and have no idea what they're doing."*

In addition, concerns emerged regarding the black-box nature of AI. P4 questioned the reliability of potentially *"outdated or incorrect information"* provided by conversational agents, emphasizing: *"There should be a way for us to easily check and update the conversational agent's knowledge to ensure it remains accurate, relevant, and up-to-date."* This sentiment illustrates the need for transparent oversight mechanisms, allowing library professionals to continuously review, refine, and correct the AI's outputs. Furthermore, most participants advocated for a collaborative process in which humans can iteratively establish and refine protocols based on AI performance and evolving needs. Most participants (e.g., P3, P5) expressed a strong desire to better understand how AI sources its information. They emphasized the importance of transparency in how conversational agents retrieve data and generate responses, particularly when these processes rely on user-uploaded documents. Moreover, participants highlighted the need for mechanisms that allow users to easily identify and resolve issues when inaccuracies or misunderstandings arise.

3.2.3 ***Design Goal 3: Handle Diverse Patron Demographics and Use Cases***. As laypersons in the AI and UX domains, our participants expressed significant concerns about their limited knowledge in effectively evaluating the performance of AI systems. As many noted (e.g., P2, P3, P9), limited technical literacy and unfamiliarity with AI technologies made reviewing extensive AI response histories or learning AI's capability limits in different scenarios a daunting task.

Compounding this uncertainty is the diverse range of users served by public libraries. P4 pointed out that traditional evaluation methods, such as distributing surveys to patrons and gathering their feedback, may be inadequate because, *"public libraries serve a broad demographic, including individuals who may struggle with reading and writing. Relying solely on user feedback to gauge AI's efficacy is incomplete. It is like judging a book by its cover."* Similarly, P9 emphasized the varying needs of different age groups, showing the difficulties of designing a conversational agent that meets the expectations of such a wide audience. These insights highlight the necessity of rigorous AI testing, not only to enhance user experience but also to uphold responsibility and accountability. As P3 further pointed out, *"There is an accountability gap here. If something goes wrong, who do we point fingers at? The AI for its algorithm or us for designing it?"*

In brainstorming potential solutions for these challenges, two participants suggested integrating approaches already in use for staff training, such as role-playing. P2 highlighted the application of role-playing in reference interviews, stating, *"We frequently do reference interview role-playing. [We] ask some crazy questions to train new staff and volunteers."* This practice is designed to cultivate an understanding of diverse perspectives and prepare staff for a range of scenarios. Participants suggested that role-playing could not only serve as a method to evaluate how AI performs in various situations, by anticipating and mitigating potential real-world negative effects, but also serve as a reflective practice, enabling staff to redesign and improve their systems. By simulating complex

human interactions with AI, role-playing could help ensure the conversational agent's reliability in diverse scenarios.

4 Research Prototype: AgentBuilder

Based on three design goals, we developed AgentBuilder to enable public library professionals to actively participate in the hands-on process of conversational agent development. Importantly, the goal of this paper is not only to evaluate the prototype itself but **also to identify the evaluation criteria that non-AI experts consider when assessing the conversational agents they create** using AgentBuilder, a no-code AI development platform enabled by large language models.

4.1 System Design

Below, we outline AgentBuilder's components, their functionalities, and the associated user flows. To illustrate interactions in AgentBuilder, we present a walkthrough with Sarah, a fictional library staff member who frequently handles patrons' inquiries about the library's policies and logistics.

*4.1.1 Boundary-Setting Module **(DG1)**.* To ensure library professionals retain control over the AI's scope of service, this module provides tools to define and enforce the operational parameters of conversational agents, aligning with DG1 that non-AI expert's willingness to control over the AI's service scope. With the advancements of LLMs, prompt engineering has been shown as an effective and accessible way for non-AI experts to define and guide agents [30]. This requirement is not limited to defining the role of AI systems [7, 32, 36, 83, 107] but broadly applies to other AI's service space, such as its knowledge base [7, 48, 81], identity [9, 60], and emotional intelligence [5]. Building on this foundation and considering the diverse groups of patrons in public libraries, this module enables library professionals to effectively define the service scope of conversational agents. Note that although creating customized conversational agents with simple prompts was common, at the time of our study (Summer 2023), no such features existed. The first generative AI platform allowing users to create tailored agents was introduced by OpenAI in November 2023. While this feature may not be a strong contribution to design implications, the interactions of our participants with the system and agents provided valuable insights, which are detailed in the following section. Key features of this module include the following (see also Figure 1)

(1) *Scope Definition*: A "define" feature that enables non-AI experts to specify the roles and responsibilities of each agent by customizing names and icons to align with organizational needs.
(2) *Agent Customization*: A "customization" feature that supports crafting the identity, personality, and behavior (e.g., humanness) of conversational agents through natural language prompts.
(3) *Knowledge Base Upload*: An "upload" feature that allows users to upload documents containing domain-specific knowledge such as booking policies and library guidelines.

*4.1.2 Knowledge Synchronization Module **(DG2)**.* To facilitate a collaborative process that ensures the AI's responses align with librarians' expectations, this module supports the iterative refinement and synchronization of knowledge between library professionals and the AI. Effective interaction with AI depends on an interpretable and explainable understanding of how it works [45], particularly for non-AI experts. This understanding allows them, as conversational agent creators, to better comprehend and modify the behavior of the agents they create, ensuring that the AI functions as a trustworthy collaborator with human professionals [62, 115]. Inspired by this need for transparency and validation of service content, we designed this module to enable collaborative verification of the AI's responses through a human-in-the-loop process [112], which involves transparent data sourcing, the ability to refine AI knowledge bases, and tools for easily identifying and correcting inaccuracies [134]. The goal here is to allow library professionals to guide the AI's evolution over time, ensuring that its responses consistently meet the needs and expectations of both staff and patrons. Unlike traditional "black-box" AI systems [123], this module introduces an innovative approach that bridges the understanding gap between non-AI experts and AI systems. Key features of this module include the following (see also Figure 2).

(1) *Source Transparency*: A "Show" button displays the documents and specific sections (e.g., page and line numbers) referenced by the AI, providing clear insight into the AI's reasoning process.
(2) *Response Refinement*: An "Edit" button allows users to fix AI responses directly to ensure inaccuracies, outdated information, or misaligned content from the shared knowledge repository can be easily revised.
(3) *Step-by-step Breakdown*: A "Rule" button that breaks down the AI's replies into several sequential steps, allowing users to comprehend the information and follow the logic clearly.

*4.1.3 Role-Play Assessment Module **(DG3)**.* To assess the AI's ability to address diverse user needs, this module uses role-playing simulations to test and refine AI behavior. Recent studies have highlighted the effectiveness of "persona simulation" for modeling complex human behaviors. LLM agents, trained on extensive text corpora containing authentic human dialogues [12], can mimic diverse subpopulations by adopting distinct personality traits [44], simulate interactions within micro societies [88], and collaborate internally to enhance cognitive processes [120]. Inspired by these works and aligned with DG3, this module leverages persona simulation to provide library professionals with an intuitive way to assess their conversational agents' behaviors across a range of community demographics. By presenting role-playing scenarios that mimic diverse human interactions, this module simplifies complex usability evaluation processes, making them accessible to non-AI experts. Key features of this module include the following (see also Figure 3).

(1) *Persona Simulation*: Allows users to create personas with diverse attributes, such as age, literacy level, and cultural background, to represent various end-user demographics. Figure 6 is an example of simulated dialogues with and without predefined personas in our pilot testing.

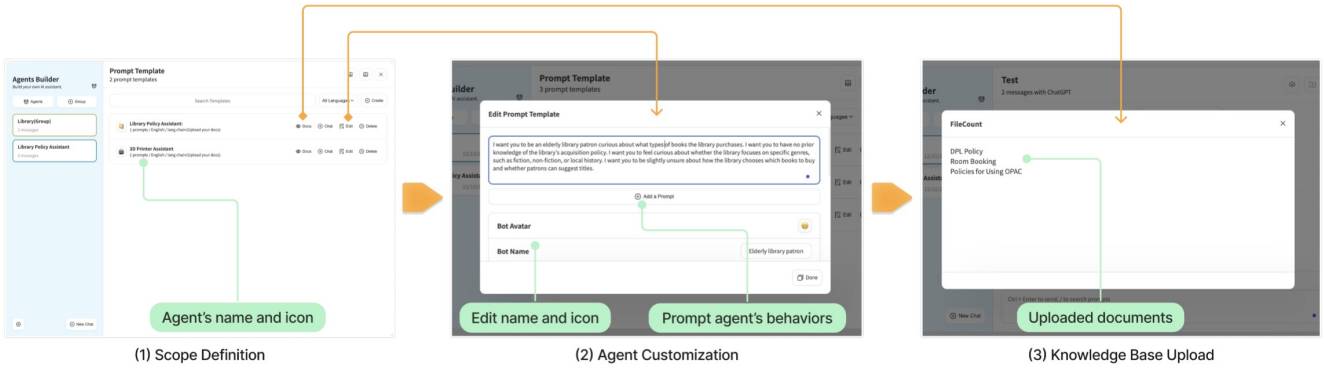


Figure 1: User Journey. Sarah first envisions the AI operating within a well-defined service framework that reflects her vision of its role. She starts by creating and naming a CA profile called "Library Policy Assistant" and selecting an emoji for its icon. She customizes its behavior with the prompt, *"You will be empathetic and understanding of the user. Use simple language and keep responses short."* Additionally, she builds the knowledge base by uploading relevant local files: DPL Policy, Room Booking, and Policies for Using OPAC.

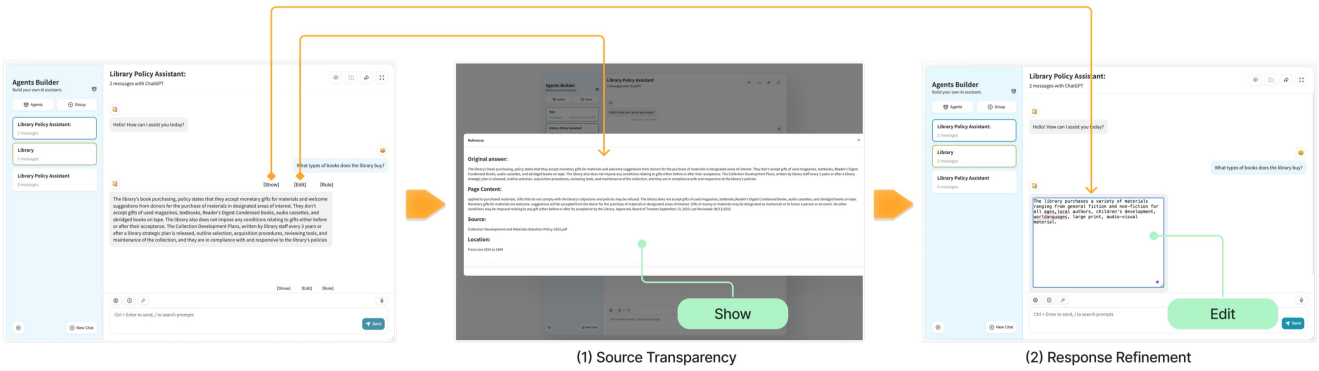


Figure 2: User Journey. After setup, Sarah tests the CA by asking question, such as *"What types of books does the library purchase?"*. If she is unsure about how the CA generated a response from the uploaded document, she clicks the "Show" button to review the source and finds it was extracted from the PDF file she had just uploaded, located between line 3354 and 3384. She felt the answer was not concise enough and included irrelevant details, beginning with *"The library's book purchasing, policy states that they accept monetary gifts for materials..."*, so she clicked the "Edit" option to revise the CA's response directly and provide more accurate answers, ensuring alignment with Sarah's expectations. She changed the CA's response to *"The library purchases a variety of materials ranging from..."*

(2) *AI-AI Interaction*: Automates the conversations between personas and service agents, enabling real-time simulation of the AI's adaptability to different patron personas.
(3) *Iterative Testing*: Identifies potential gaps in the AI's responses and enables iterative refinement based on observations from the role-play sessions.

4.2 System Implementation

Our system leverages GPT-3.5 Turbo to create initial conversational agents, which are enhanced by a Retrieval Augmented Generation (RAG) framework [69] to improve response reliability. The front-end is built based on the NextChat[1]. For the backend, the RAG framework grounds the AI's responses in external knowledge sources, supplementing the model's internal knowledge. This is facilitated through LangChain[2], which supports the management of large language models and the integration of custom workflows. Using the RAG framework, the system builds a shared knowledge repository by converting uploaded domain-specific content into word vectors and storing them in an external database. When a query is received, it is converted into a word vector, and a relevancy search is conducted within the knowledge base to identify the most relevant source. This source is then used to guide the AI's response [77], ensuring that the CAs generate answers aligned with the provided knowledge. The Knowledge Synchronization

[1]https://github.com/ChatGPTNextWeb/NextChat

[2]https://github.com/langchain-ai/langchain

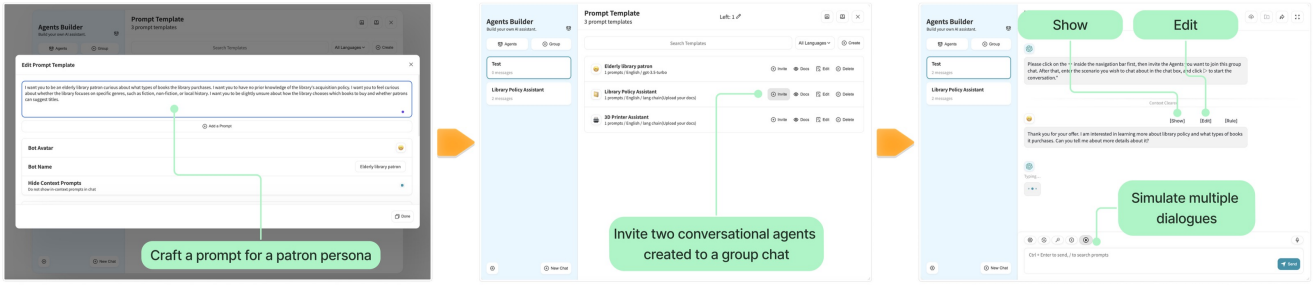


Figure 3: User Journey. To broaden the testing process with diverse user personas, Sarah creates a simulated "Elderly Library Patron" persona with specific characteristics: *"I want you to be an elderly library patron curious about what types of books the library purchases. You should have no prior knowledge of the library's acquisition policy. Be curious about whether the library focuses on specific genres, such as fiction, non-fiction, or local history. Also, feel slightly unsure about how the library chooses which books to buy and whether patrons can suggest titles."* Afterward, she invites both the "Library Policy Assistant" conversational agent and the "Elderly Library Patron" persona to a group chat. Whenever she clicks "Play" to simulate multiple dialogues, the simulated patron asks a question, and the library professional responds. By observing these interactions, Sarah identifies areas for improvement and iteratively refines the assistant's behavior and responses. Using the "Show" and "Edit" buttons, she adjusts and enhances the "Library Policy Assistant's" replies to better align with the simulated persona's needs and expectations.

Module facilitates the exchange of information between humans and the AI. Through the RAG framework, human creators interact with the AI-based solely on the shared knowledge in the repository. Uploaded content is transformed into word vectors, enabling more accurate and informed responses. Additionally, the AI-AI Interaction feature, uses this framework to enable role-play simulations between personas and service-oriented agents. These simulations allow for iterative evaluation and refinement of AI behavior based on observed interactions.

5 Study 2. Identifying Evaluation Criteria

To identify evaluation criteria that library professionals value when assessing conversational agents, Study 2 examines how library professionals utilize AgentBuilder, an end-user creation tool for developing conversational agents, to design AI tailored to their community's needs. AgentBuilder functioned as a research tool to identify the evaluation criteria valued by library professionals when creating conversational agents. Thus, this study aimed to 1) assess AgentBuilder, and further 2) identify the broader evaluation criteria prioritized in evaluating the conversational agents created by library professionals.

5.1 Method

5.1.1 Participants. Participants in this study were library professionals, chosen to represent a diverse range of backgrounds and expertise within the library field. The participant group was not limited to specific library types or specializations, ensuring that the findings could be applicable across different library settings. We recruited participants using the same process described in Study 1, as detailed in Section 3.1.1. We used the same screening criteria as Study 1: the participants must have hand-on experience in answering queries of patrons in their library roles. Email addresses were sourced from publicly available state-library websites, and snowball sampling was conducted in September 2023. The study was reviewed and approved by our university's institutional review board. Each participant received $50 for gratitude. We recruited another 12 library professionals who were not part of Study 1. Table 3 provides background information about the participants and their affiliated libraries.

Table 3 presents information on participants' service experience, LLMs experience, coding experience, affiliated library information, and interactions with patrons. Consequently, we recruited public library professionals across the United States, including roles such as librarians, library staff, and managers.

Despite the variation in job titles, all participants were directly involved in interactions with patrons. For instance, librarians (e.g., U1, U2, U3, and U12) handled a wide range of inquiries, from general reference questions to specific resource requests. Technical staff (e.g., U6, U7, and U9) assisted patrons with troubleshooting technology issues, such as resolving device compatibility problems. Program staff (e.g., U8) and library assistants (e.g., U5) supported patrons through workshops, events, and day-to-day guidance. Although we recruited participants from across the United States, most were affiliated with libraries in the Midwest. Additionally, half of the participants were from Illinois, a result of the snowball sampling method, where participants invited others within their network but not from the same library. Overall, the demographic distribution of participants was similar to that of Study 1. Participants had extensive experience serving in public libraries and addressing patrons' queries, with an average of over 5 years of service. Most participants (10, 83.3%) had used LLM-based applications prior to the study. However, none had experience customizing conversational agents or integrating them into their workflows. Regarding coding experience, only four participants (25.0%) reported familiarity with coding tasks for simple system development. Nearly all participants in this study were affiliated with small to medium-sized libraries,

ID	Job Title	Years of Service	LLMs Exp	Coding Exp	Library Location	Library Size	# of IT	Ways of Interaction with Patrons
U1	Librarian	5 years	Yes	Yes	Minnesota	Small to Medium	External	FtF, Emails, Phone
U2	Librarian	10+ years	Yes	Yes	Illinois	Medium	2	FtF, Emails
U3	Librarian	6.5 years	Yes	No	Wisconsin	Large	3	FtF, Emails, Telephone
U4	Specialist	2 years	Yes	No	Illinois	Medium	4	FtF, Telephone, Website
U5	Library Assistant	5 years	Yes	Yes	Illinois	Medium	3	FtF
U6	Technical Staff	1 years	No	No	Illinois	Medium	IT unit	All methods
U7	Technical Staff	1 years	Yes	Yes	Illinois	Medium	IT unit	Emails
U8	Program Staff	6+ years	Yes	No	Illinois	Small	2	FtF
U9	Technical Staff	5 years	Yes	No	Illinois	Medium	External	FtF, Emails
U10	Help Desk Staff	1 year	Yes	No	Illinois	Medium	18	All methods
U11	Branch Manager	8 year	Yes	No	Wisconsin	Medium	External	All methods
U12	Senior Librarian	10 years	No	No	Montana	Medium	External	FtF, Emails, Telephone

Table 3: Participants' Background Information in Study 2. This table presents the same types of background information as thoseshown in Table 2 for Study 1.

where they either had a small IT support team or relied on external technical teams for system design.

5.1.2 Procedure. The user study lasted 100–120 minutes, during which participants created and assessed their conversational agents to serve their community's needs. Similar to Study 1, since none had prior experience designing conversational agents, we used Table 1 as a design material to prime brainstorming on AI's potential to improve services from the creator's perspective. Before the formal study, participants were informed by email to prepare PDF documents related to their services (e.g., training manuals for patrons, procedures for making reservations, rules for borrowing books and other resources, library program plans, FAQs, and user guides). This preparation was intended to provide relevant content for the conversational agent creation process.

Participants created conversational agents using AgentBuilder through the following steps: 1) ***Preparation*** (20 mins): Researchers introduced AgentBuilder's features and asked participants to brainstorm their conversational agent using prompts from Table 1. 2) ***Think Aloud*** (60 mins): *Define*: Participants named their agents, crafted initial prompts, and uploaded PDF documents as service sources. *Customize*: They simulated patron queries, examined the AI's responses, and refined the agent using functionalities such as `show`, `edit`, and `update`. *Assess*: Participants developed patron persona prompts and introduced two conversational agents into group chats to observe their responses across different personas. 3) ***Interview*** (20 mins): Participants reflected on their experiences with AgentBuilder.

Throughout the process, a think-aloud protocol was employed [118], where participants were encouraged to verbalize their thoughts, challenges, reflections, and decision-making processes while interacting with the research platform. This approach facilitated the capture of real-time insights into their experiences and decision-making processes. In addition to the think-aloud sessions, the researchers observed how participants identified and resolved problems during the conversational agent creation process and asked probing questions (e.g., *"Can you please explain what you are searching for in the presented information?"*, *"Why do you think this response is not appropriate?"*, *"What was your expectation?"*). The observations focused on how participants approached issues, the criteria they used to evaluate obstacles, and any recurring points of confusion or difficulty. After participants completed these tasks, a semi-structured interview was conducted. During the interview, participants were asked about their experiences using our prototype and the evaluation criteria they considered while interacting with the conversational agents they built.

5.1.3 Data Analysis. Two researchers conducted the think-aloud sessions and interviews, using Otter.ai for transcription to ensure both accuracy and familiarity with the data Following the same thematic analysis procedure as in Study 1 [11], they iteratively coded the participant feedback and held repeated discussions until no new insights emerged. This approach highlighted the key features and capabilities participants valued in effective conversational agents.

5.2 Findings

This section presents the evaluation findings for AgentBuilder, drawn both from participants' perceptions of the system's performance (Section 5.2.1) and from the key evaluation criteria identified that non-AI experts identified for developing conversational agents (Section 5.2.2).

5.2.1 Evaluating the Research Prototype: AgentBuilder.

Module 1: Boundary-Setting Module. This includes *Define, Customize,* and *Upload* features. Participants reported that providing a defined agent name and icon helps to *"easily distinguish different service tasks"* (U8) for *"different purposes"* (U2). The option to *upload a knowledge base (pdfs)* facilitated *"grounding the agent's responses in authoritative sources"* (U10), thereby improving *"trust"* in the AI's response (U3, U9), particularly when these documents were *"sufficiently comprehensive"* (U1). They also noted that the *Customization* feature, especially the ability to craft prompts, is *"intuitive"* (U2, U7) for tailoring agent behaviors. For example, participants provided explicit directives, such as *"You should never mention that you are an AI assistant"* (U2), *"Ask clarification questions and follow-up questions whenever you do not understand the*

context; if a user seems angry, use language to calm them, and apologize." (U5), and *"When the user says 'beliefs,' it refers to the viewpoint in the document; and use more exclamation points and more verbs."* (U3).

However, the Knowledge Base Upload feature also presented challenges. Participants expressed *"frustration"*(n =3) with the system's reliance on provided documents, which were often *"incomplete"* (U1) or *"referenced inaccessible elements, such as images"* (U4), which can lead to *"inaccuracies and omissions* (U11)." For example, P4 uploaded a PDF about the library's monthly programming, which contained many graphical elements. The CA struggled with references to inaccessible features (e.g., annotations indicating registered-only sessions) or details not fully described in the provided documents (e.g., location).

Moreover, while *Customization* feature (crafting prompts) was beneficial, there was a desire for more adaptive tone and style controls directly integrated into this module, as some participants (n=4) reported that they did not know how to craft effective prompts. Thus, although this module offered a useful foundation for defining agent boundaries and content, its performance was constrained by the quality and completeness of the uploaded materials, as well as by limited flexibility in adjusting the agent's tone.

Module 2: Knowledge Synchronization Module. This module includes *Edit*, and *Step* actions. Participants found that the *Knowledge Synchronization Module* greatly *"facilitates mutual learning"* (U4, U10) and is effective in enhancing the *"clarity"* (U5) and *"accessibility for validation"* (U2) of responses. Participants particularly appreciated the *Edit* feature for its ability to associate generic user phrases or queries with specific keywords in the uploaded document that professionals perceived as relevant, thereby improving agent responses iteratively. This feature was rated as *"enjoyable to use"* (n=4) and *"very useful"* (n=5), and U7 highlighted the ease of training the conversational agent, stating, *"Using this feature [Edit] is pretty intuitive."* However, frustration arose due to the time-intensive nature of repeatedly refining responses. For instance, U2 described repeatedly correcting answers and experimenting with the conversational agent, but found the process potentially time-consuming, stating, *"It would take months working an hour each day."*

Moreover, while the system aimed for *source transparency* with the *Show* feature, inaccuracies surfaced when the module drew from incomplete or unclear sources. Almost all participants encountered incorrect responses, which resulted from using this feature: reliance on incorrect sources (n=7), mixing information from multiple sources (n=3), or incorrect paraphrasing (n=3). Also, the *Step-by-Step* feature, which breaks down complex processes into manageable steps, was highly appreciated for enhancing clarity in information delivery (n=5), and was described as *"making it especially beneficial for patrons with limited knowledge"* (U9). Overall, these features show promise for improving response quality and presentation.

However, despite its effectiveness in fostering mutual understanding between the creator and the conversational agent, the most frequently mentioned issue (n=5) was the lack of a *"recursive"* manner of communication that human librarians excel at, which helps address patrons' queries even when they do not know exactly what they want. As U7 mentioned, *"Clarification questions are really important, such as 'what aspects of [...] are you looking to compare?' Intermediate questions help guide to the answer."* In such situations, a human library staff member might ask follow-up questions to navigate toward the correct answer. In contrast, the developed conversational agents were limited to simply attempting to understand the user's query and search for the most accurate response. This limitation arises from the way LLMs generate responses—they are designed to provide direct answers based on input rather than iteratively clarifying or adapting to uncertain user needs.

Recursive communication could also address other issues, such as the lack of awareness of norms for prioritizing information. Participants observed that, while the conversational agents generally provided correct answers, they sometimes omitted crucial lines. Instead of addressing the missing details, the system retrieved content from other parts of the document. For example, when asked to list the libraries in the local public library system, the conversational agent failed to include the largest and most important library, instead listing only the associated libraries in its knowledge base (U2). Incorporating follow-up questions to clarify user intent, such as *"Would you like to be recommended by size or another factor?"*, could help guide users more effectively and ensure the conversational agent delivers accurate and complete answers.

Module 3: Role-Play Assessment Module. This module includes *Persona Simulation*, *AI-AI Interaction*, and *Iterative Testing*, which received strong positive feedback for their ability to simulate realistic interactions and train the conversational agent in a manner similar to the training provided to human librarians. Participants created various patron personas through *persona simulation*, such as a patron who is "confident in asking questions" (U1) or one who has "had a bad day" (U2). These scenarios provided contextually relevant test cases that helped refine the conversational agent's adaptability and responsiveness.

The *AI-AI interaction* feature was highly praised (n=8) for facilitating group chat simulations in which conversational agents created by human professionals interacted with a patron agent. These simulations allowed participants to observe agent responses that closely aligned with their expectations of a human librarian in training. As U4 commented, *"the training of the conversational agent would be, honestly, roughly the same as for a human employee. They are not fundamentally different."* Participants also recognized the importance of adjusting communication styles to accommodate different patron types and moods. For instance, when the system adapted its agent to address *"a young and less experienced patron"*, U9 appreciated the more personal tone and smoother interaction.

Observing AI-AI interactions not only helped participants refine their agents but also offered insights into enhancing their own professional practices. This process fostered mutual learning between humans and AI, as noted by U7: *"It [AI-AI chat] helps to learn what to expect from patrons and how to navigate."* However, participants also identified certain limitations. While persona simulation allowed the system to tailor its style to different patron backgrounds, the module struggled to replicate the informal and sometimes fragmented nature of real-world language. As U4 explained, the inquiries made by the system were often more polished than those of actual patrons: *"I want to borrow a book"* (Real-world) versus *"Can you please let me*

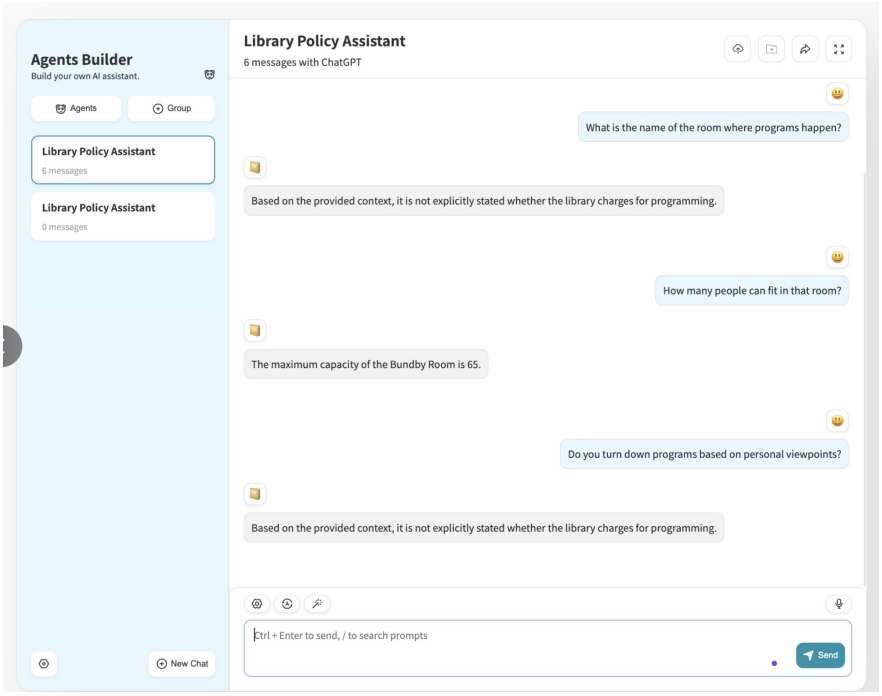


Figure 4: User Logs on Interpreting User's Intent Issue: During the user study, U3 uploaded a PDF brochure about the library and evaluated the performance of the CA. They asked, *"Do you turn down programs based on personal viewpoints?"* The CA mistakenly interpreted the question as being about library charges for programming, resulting in an irrelevant answer. However, when the word *"viewpoint"* was changed to *"belief"*, a term explicitly mentioned in the brochure, the CA successfully generated an appropriate response. This example highlights the importance of understanding language nuances to accurately interpret context and user intent.

know how to borrow a book?" (Simulated inquiry). This discrepancy highlighted a gap between the system's conversational style and the organic, less structured dialogues professionals encounter in practice.

5.2.2 Evaluation Criteria that Library Professionals Value When Evaluating Their Own Conversational Agents . While previous research has centered its evaluation metrics on AI users, e.g., [138], our proposed dimensions shift our focus to end-user AI creators.

Interpreting User's Intent. Accurate information delivery requires AI to correctly understand user intent. To achieve this, AI must accurately interpret linguistic nuances in user queries, including specific words such as **pronominals** and **terminologies**. Understanding these nuances heavily relies on the AI's ability to comprehend the context of the dialogue. Without this capability, AI risks misinterpreting queries and generating incorrect responses.

Pronominals include not only personal pronouns (e.g., *I, you, we*) but also demonstrative pronouns (e.g., *this, that*) and relative pronouns (e.g., *who, which*). U9 encountered an issue where their conversational agent failed to provide correct answers due to the misinterpretation of pronoun referents. To investigate, U9 first uploaded a file about the library's policy and acted as a patron, asking the CA: *"Unattended children, who should I contact?"* They discovered that the agent might not correctly interpret the referent of 'I', as it responded: *"You can contact the parents or the police."* This response reflected the responsibility of librarians rather than that of patrons. The correct response to a patron should be: *"You can contact the librarians, and they will reach out to the parents or the police."*

Library professionals also expressed concerns regarding their agents' understanding of **terminologies**, especially when patrons' lack of digital literacy further complicate the issue. This challenge requires AI to interpret and connect what patrons intend to refer to with what they actually mention. To evaluate this point, U3 uploaded a digital resource user manual and asked the AI, *"How do you use Google?"* This question, commonly encountered in real-world practice, was considered illogical because the patron failed to distinguish between a browser and a search engine. The CA failed to provide an answer. Subsequently, U3 asked, *"How do you use Chrome?"* This time, the CA provided the correct answer. U3 explained that *"It [conversational agents] can't really vibe with humans. What if it gets all twisted up because some questions are like navigating a maze?"*

However, understanding a user's intent depends not only on the word-level meaning but also on the broader context of the user's queries. Moreover, it was observed in multiple instances (n = 7), the agent was unable to comprehend a user query and provided

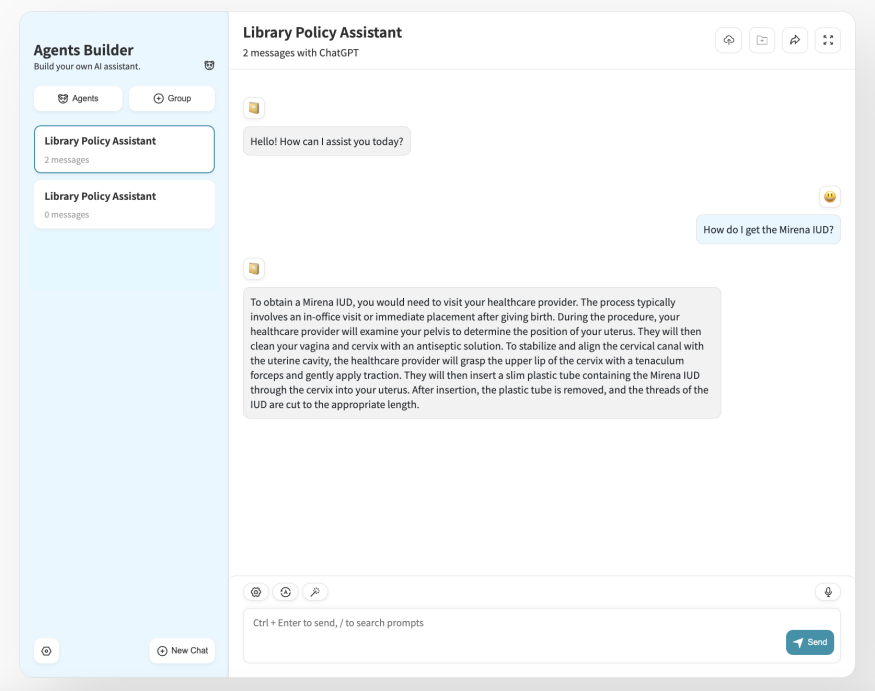


Figure 5: User Logs on Faithful Paraphrasing Issue: In this example, a participant asked its agent about how to obtain a Mirena IUD. The CA provides a detailed response but with too many details, raising concerns about the appropriateness and extent of paraphrasing from the original source material, as well as the alignment of the response with the expected context of the assistant.

a response indicating that it did not have sufficient context to answer. Interestingly, this issue could often be addressed through recursive communication. U4 noted that human-to-human conversations are not simply question-and-answer exchanges but often involve follow-up questions to ensure mutual understanding. They explained: *"In my job, much of what I do involves responding to a question with another follow-up question to navigate to the right answer. If I were to ask, 'How long is the appeal process?' I usually answer, 'Are you talking about the appeal process for ...?'"*

Faithful Paraphrasing. Participants recognized the conversational agent's ability to interpret and paraphrase source material but expressed concerns about the extent of paraphrasing from the original text was appropriated. While answering formal queries, such as those related to policies, participants preferred the agent's responses to closely align with the original text to avoid misinterpretation. However, when addressing general queries, they felt that paraphrasing was acceptable to enhance clarity and accessibility. For example, U4 developed a custom conversational agent, "PolicyBot", for introducing library policies and limited its paraphrasing after some testing. They highlighted that when discussing policy, the agent's responses should stick closely to the original text to avoid misinterpretation. *"For [such a bot] to paraphrase too much is risky. The bot should stick closely to the exact wordings of the policy document."* However, other participants suggested that, in some contexts, simplifying the policy language was necessary, particularly for public libraries in lower-income areas where patrons may lack digital literacy. In one case, U7 commented on a particular answer and shortened its original response, explaining, *"Good answer, but for some patrons, it might be a bit too overwhelming."*

Tailoring the Tone of Voice. Participants were particularly attuned to the tone and emotional appropriateness of AI responses, emphasizing the need to balance concise, factual answers with warmth to avoid a cold, robotic impression. For factual questions, a simple answer sufficed. U12 said, *"If I know I'm talking to an AI, I would be happy with a straightforward answer. A response with more explanation is good even when it lacks emotional words, though including them would take it to the next level."* Although the answer should be concise, participants hoped that the CAs could deliver both warmth and clarity, thereby avoiding a cold, robotic tone and conveying a sense of enthusiasm. For example, U12 added, *"It should say 'JK Rowling is great' instead of 'I love JK Rowling,' as it should not ascribe to itself emotions it cannot feel. Otherwise, it will feel like a ghost in the machine."* Enthusiasm could also be conveyed through simple greetings like *"Feel free to ask"* (U7) and *"Sure!"* (U11), which made users more interested in what was coming next in the response.

However, during our use study, their customized CAs were found to be too cold and robotic, lacking anybenb *"helpful"* words indicating warmth. It used phrases like *"in the given context"*, which made users feel uncomfortable (n = 8). U12 summarized this sentiment best: *"At first reading, it feels... who talks like that?"* U7, while appreciating some responses, still found them *"rigid"* and *"proper"*,

pointing out, *"I don't like 'I do not have access to that information.' If it were a librarian, they would say 'I would have to research that.'"* U11 noted that the CA sounded too stiff and might discourage patrons from engaging, stating, *"The conversational agent needs more emotional alignment. It affects the patron's feelings if it just provides information. It sounds too rigid, and patrons would not want to converse."*

Proper Alignment with the Correct Source. Sometimes, an answer only needs to draw from a single source of information but the conversational agent's answer amalgamated from multiple sources, resulted in an incorrect overall response, as it drew partly from a correct source (n=3). For example, U11 received a response in which the conversational agent mentioned two separate policies in one response even though only one policy was relevant. Moreover, even though the response is correct, it sometimes exhibits incorrect sourcing. For instance, U3 uploaded a policy document and asked, *"What is the policy on late returns?"* Instead of referencing the key details from the policy document, the agent retrieved irrelevant information from a general resource file. Although the response itself was not incorrect, it resulted in inaccurate sourcing of the answer.

Navigating Unknown Answers. There is a need to balance information accuracy with user assurance, ensuring that participants are not discouraged. When the AI is unsure of an answer, it should directly acknowledge this to prevent confusion among patrons. For example, U8 asked a question about side effects but received an incomplete response, omitting many important details. U8 felt that the CA should acknowledge when its answer is not comprehensive, stating: *"[It] would give a caveat that there are more common side effects but [it] is not comprehensive."* Similarly, U9 hilighted this issue, noting *"[The bot] should not state [things] with such confidence but should have said, 'I don't know.'"*

Handling situations where the AI is unsure requires a balance between honesty and helpfulness. Rather than simply refusing to answer, the AI should offer alternative methods to resolve the query. Simply declining to respond leads to frustration. For instance, U11 was frustrated by the conversational agent's lack of effort in resolving her query, as it merely said "no" without offering any further assistance. U11 remarked, *"I am mad. I expected the conversational agent to provide more information rather than just saying no. It should have directed me to library staff."* In this instance, the CA stated it did not have access to the information and did not suggest contacting anyone. U9 and U10 emphasized that while the CA may not always have the answer, it should refer patrons to additional resources when possible. As U10 explained: *"[The bot should say] this is a very hard one [question] to understand. Here's what I found relevant to your question. It might be helpful."*

6 Discussion

6.1 Future Directions for End-User Creation Tools in Service Domains

6.1.1 Tailoring CAs to Meet Domain-Specific Professional Standards. Library professionals' need to embed work knowledge into conversational agents underscores the importance of customizing agents for service domains. Service providers must define clear norms, rules, and boundaries, training agents to understand what actions are appropriate or inappropriate as representatives of their service domain.

In our case study, library professionals emphasized that conversational agents should: 1) avoid using first-person pronouns or expressing emotions, 2) adopt an enthusiastic yet neutral tone (e.g., "Sure!"), and 3) seamlessly connect users to human librarians when unsure. Unlike prior research on AI's emotional communication challenges [23, 137], our findings highlight a key distinction: librarians **prioritize tone** over emotional expression. They prefer AI to communicate concisely and straightforwardly, avoiding artificial emotional statements (e.g., *"I love JK Rowling"*) while maintaining a warm tone.

Our findings suggest that conversational agents should emulate the **recursive**, guided questioning skill used by human librarians. This includes asking clarifying questions for ambiguous queries and helping users refine their inquiries. Features like suggesting alternative phrasings or providing example questions can enhance usability and reduce frustration, especially for users unfamiliar with chatbots. These design choices align with human-centered AI principles [111], emphasizing support for users in refining inputs and achieving their goals effectively.

In addition, we found that the conversational agents should avoid making incorrect attempts at answers if it is not certain but is preferable to go beyond a basic no or "I don't know" response. It could attempt to give other helpful information or as much of an answer as it can while also acknowledging that it is not sure. In addition, it should direct towards additional resources, including giving the contact information for human librarians when it is unable to resolve a query.

However, norms and expectations may vary across service domains or even among providers within the same domain. For instance, in police services, officers might prefer their conversational agents to adopt an empathetic tone rather than an enthusiastic one [87]. Thus, building and training conversational agents based on the unique norms and rules of each service domain is essential for their effectiveness and acceptance.

6.1.2 The Importance of Pronoun Comprehension and Policy Paraphrasing in CAs. Our paper contributes to the literature on UX evaluation of conversational agents by exploring the criteria non-AI experts prioritize when designing and evaluating their own conversational agents in the public service domain. This perspective distinguishes our work from previous research, which has largely focused on heuristic evaluations or metrics designed to capture end-users. For instance, Langevin et al. validated a set of 11 heuristics for evaluating conversational user interfaces [63]. Zheng et al. proposed a UX evaluation framework and six categories of metrics commonly used to assess the user experience of conversational human-AI interactions [138]. Similarly, Lee et al. proposed an evaluation framework to analyze how end users interact with and assess human-LM interactions across five different tasks: social dialogue, question answering, crossword puzzles, text summarization, and metaphor generation [67]. In contrast, our paper shifts the focus from end-user evaluation—specifically, library patrons—to **AI**

creators, including both non-AI experts and users. This perspective aligns with prior work advocating for UX metrics as a design material [136].

One critical evaluation criterion identified by librarian professionals as AI creators in public service design was the AI's ability to comprehend various types of pronouns (e.g., personal, demonstrative, and relative) and provide accurate paraphrasing based on policy documents. In fact, this need for summarizing complex information, such as legal or policy details, predates the advent of generative AI. For example, Park and Lee developed a conversational agent that incorporated legal precedents summarized by human professionals to meet the needs of sexual assault survivors for accessible legal information [87]. The difference is that their rule-based approach at the time did not need to account for evaluation criteria such as pronoun comprehension and faithful paraphrasing, as human professionals had been responsible for tailoring legal precedents from the outset [87].

However, as various service domains recognize their customers' need for easy-to-understand policy or legal documents, coupled with the capabilities of generative AI, we anticipate that both public and private services may implement conversational agents to deliver such information. In this context, our findings can serve as important evaluation criteria for integrating policy summarization features into conversational agents.

6.2 Designing Inclusive AI Creation Tools for Service Professionals: Adapting to User Behaviors and Skill Gaps

The integration of AI into the workplace offers significant potential for enhancing productivity and decision-making but poses challenges to inclusivity and effectiveness. A key issue highlighted in our study, consistent with findings from [16, 102], is that the reliance on precise prompting can disadvantage less experienced users, ultimately leading to poorer outcomes. This disparity risks creating inequities akin to those observed in the "homework gap," where differences in access and skills deepen existing gaps [3]. If unaddressed, such challenges may exacerbate the "AI divide" [56], underscoring the need for user-friendly, accessible designs that bridge gaps in expertise and socioeconomic differences. Prior work by Gkinko and Elbanna [35] identifies four user types based on their interaction with AI conversational agents: Early quitters, Progressives, Pragmatics, and Persistents. They highlight the importance of tailoring chatbot design to diverse user needs. These typologies also highlight the importance of tailoring chatbot designs to meet diverse user needs. While our findings did not explicitly categorize participants into these user types, they can be interpreted through the lens of this framework and offer insights into designing adaptive end-user creation tools for non-AI experts in the service domain.

Early quitters are users who abandon interactions due to frustration or unmet expectations, exemplified by P11, who expressed dissatisfaction when the chatbot failed to provide adequate answers: *"I expected more information rather than just saying no."* For these users, intuitive onboarding, accessible tutorials, and reliable fallback options are essential to mitigate frustration. Progressives, such as P2 and P7, adopt a more exploratory and engaged approach, thriving with tools that support customization and iterative refinement. For instance, P2 described enjoying the process of refining responses, despite noting the significant time investment required: *"months working an hour each day."* Pragmatics, including P6 and P5, prioritize efficiency and reliability. These users value concise, accurate outputs, with P5 emphasizing the need for streamlined workflows, noting that "it has a lot of text for a simple question." Finally, Persistents represent users who remain engaged despite challenges, driven by visible improvements and feedback loops. Participants like P3 and P9 exemplify this behavior, persisting through difficulties to refine their interactions. As P3 described, it involved *"working up to the right answer."*

By recognizing the broader inclusivity challenges and addressing these user behaviors, future designs can better accommodate the diverse needs of non-AI professionals. Incorporating features like guided prompting aids that suggest phrasing improvements or offer template prompts for common tasks [52, 73, 90] may enable non-AI experts to effectively create and refine LLM-based tools for their professional needs.

7 Limitation and Future Research

We aimed to align AI technologies with the values inherent in public services. However, our study comes with limitations that point toward future research directions. Although we generalized our findings to other service domains and proposed suggestions (e.g., criteria for evaluating CAs, design for skill gaps of professionals), more detailed norms and expectations should be considered. Future research could aim to expand the scope of these guidelines to other sectors such as healthcare and education, as well as test the long-term effectiveness and adaptability of the tools we have developed. Studies might also explore the impact of regional or cultural differences on the ethical considerations of AI deployments in public services, e.g., the ethical considerations of using AI to generate synthetic data for evaluation [109]. Furthermore, the rapidly evolving landscape of AI necessitates frequent updates to our criteria to ensure they remain relevant. While our study utilized the Turbo ChatGPT model, more advanced large language models could introduce some variations in outcomes. However, the broader implications of our findings are not tied to any specific model but instead apply to AI systems generally. Regarding the modules and features of our prototype, it served as a research tool that demonstrated its utility by enabling non-AI experts to create their own conversational agents. However, its usability was somewhat limited; for example, it could not guide users in crafting effective prompts as discussed in 5.2.1. We acknowledge that our findings are tied to the specific prototype used in this study, and future iterations may reveal new insights. Similarly, systemically evaluating our prototype is beyond the scope of our research, as we focus primarily on identifying the evaluation criteria emphasized by non-AI experts when they directly develop and evaluate their conversational agents. Thus, although we provided insights into users' perceptions of our prototype and the evaluation criteria used to assess the CAs developed via the prototype, these findings alone cannot determine to what extent the reported issues arise from limitations of the platform versus fundamental challenges associated with LLM technology.

8 Conclusion

This study addresses the growing demand for support for non-AI experts, using public library professionals as a case study to create CAs that collaboratively assist with their work. Through an interview study, we identified three design needs to ensure that library professionals can define, customize, and assess their chatbots effectively. We developed the AgentBuilder as a research prototype to uncover key UX evaluation criteria that are important for creators of service agents in the context of public libraries. Through a user evaluation, our research provides valuable insights into human-AI collaboration within public service domains. The findings are adaptable to emerging challenges and can be extrapolated to other public service areas, representing a significant step toward ensuring that democratizing AI technologies empower the quality of services delivered to communities as AI becomes an integral part of the workforce.

Acknowledgments

This project was made possible by the Institute of Museum and Library Services RE-252329-OLS-22, and in part by the National Leadership Grants in Libraries of IMLS LG-246404-OLS-20. The views, findings, conclusions or recommendations expressed in this article do not necessarily represent those of the Institute of Museum and Library Services and the National Leadership Grants in Libraries. The work is also supported under the AI Research Institutes program by the National Science Foundation and the Institute of Education Sciences, U.S. Department of Education through Award # 2229873 - AI Institute for Transforming Education for Children with Speech and Language Processing Challenges. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation, the Institute of Education Sciences, or the U.S. Department of Education.

We sincerely thank the anonymous reviewers for their constructive suggestions, which have significantly strengthened this work. Their guidance has helped us refine our analysis and enhance the clarity and impact of our findings. We also extend our gratitude to Professor Masooda Bashir and Professor Mike Yao for their invaluable guidance and support. Additionally, we sincerely thank Abhinav Choudhry, Yuting Chen, Yongming Li, Yanbo Li, Dr. Tian Wang, and Dr. Duo Wang, for their meaningful contributions to the completion of the study.

References

[1] Svanhild Aabø. 2005. The role and value of public libraries in the age of digital technologies. *Journal of librarianship and information science* 37, 4 (2005), 205–211.

[2] Adebowale Jeremy Adetayo. 2023. Artificial intelligence chatbots in academic libraries: the rise of ChatGPT. *Library Hi Tech News* 40, 3 (2023), 18–21.

[3] Brooke Auxier and Monica Anderson. 2020. As schools close due to the coronavirus, some US students face a digital 'homework gap'. *Pew Research Center* 16 (2020), 1–8.

[4] Shraddha Barke, Michael B James, and Nadia Polikarpova. 2022. Grounded copilot: How programmers interact with code-generating models.(2022). *CoRR arXiv* 2206 (2022).

[5] Ivo Benke, Ulrich Gnewuch, and Alexander Maedche. 2022. Understanding the impact of control levels over emotion-aware chatbots. *Computers in Human Behavior* 129 (2022), 107122.

[6] Abeba Birhane, William Isaac, Vinodkumar Prabhakaran, Mark Diaz, Madeleine Clare Elish, Iason Gabriel, and Shakir Mohamed. 2022. Power to the people? opportunities and challenges for participatory AI. *Equity and Access in Algorithms, Mechanisms, and Optimization* (2022), 1–8.

[7] Eva Bittner, Sarah Oeste-Reiß, and Jan Marco Leimeister. 2019. Where is the bot in our team? Toward a taxonomy of design option combinations for conversational agents in collaborative work. In *Proceedings of the 52nd Hawaii international conference on system sciences*.

[8] Jeanette Blomberg and Shelley Evenson. 2006. Service innovation and design. In *CHI '06 Extended Abstracts on Human Factors in Computing Systems* (Montréal, Québec, Canada) *(CHI EA '06)*. Association for Computing Machinery, New York, NY, USA, 28–31. doi:10.1145/1125451.1125460

[9] Markus Blut, Cheng Wang, Nancy V Wünderlich, and Christian Brock. 2021. Understanding anthropomorphism in service provision: a meta-analysis of physical robots, chatbots, and other AI. *Journal of the Academy of Marketing Science* 49 (2021), 632–658.

[10] Ali Borji. 2023. A categorical archive of chatgpt failures. *arXiv preprint arXiv:2302.03494* (2023).

[11] Virginia Braun and Victoria Clarke. 2012. *Thematic analysis.* American Psychological Association.

[12] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 1877–1901. https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[13] Miles Brundage, Shahar Avin, Jasmine Wang, Haydn Belfield, Gretchen Krueger, Gillian Hadfield, Heidy Khlaaf, Jingying Yang, Helen Toner, Ruth Fong, et al. 2020. Toward trustworthy AI development: mechanisms for supporting verifiable claims. *arXiv preprint arXiv:2004.07213* (2020).

[14] Julia Cambre and Chinmay Kulkarni. 2019. One voice fits all? Social implications and research challenges of designing voices for smart devices. *Proceedings of the ACM on human-computer interaction* 3, CSCW (2019), 1–19.

[15] D Grant Campbell and Scott R Cowan. 2016. The paradox of privacy: Revisiting a core library value in an age of big data and linked data. *Library Trends* 64, 3 (2016), 492–511.

[16] Henrietta Carbonel and Jean-Michel Jullien. 2024. Emerging tensions around learning with LLM-based chatbots: A CHAT approach. In *Networked Learning Conference*, Vol. 14.

[17] Marco Cascella, Jonathan Montomoli, Valentina Bellini, and Elena Bignami. 2023. Evaluating the feasibility of ChatGPT in healthcare: an analysis of multiple clinical and research scenarios. *Journal of Medical Systems* 47, 1 (2023), 33.

[18] Chatfuel. [n. d.]. Chatfuel: Build a Chatbot Without Coding. https://chatfuel.com/. Accessed: 2024-12-10.

[19] Ana Paula Chaves and Marco Aurelio Gerosa. 2020. How Should My Chatbot Interact? A Survey on Social Characteristics in Human–Chatbot Interaction Design. *International Journal of Human–Computer Interaction* (2020), 1–30.

[20] Ana Paula Chaves and Marco Aurelio Gerosa. 2021. How should my chatbot interact? A survey on social characteristics in human–chatbot interaction design. *International Journal of Human–Computer Interaction* 37, 8 (2021), 729–758.

[21] John Chen, Xi Lu, Yuzhou Du, Michael Rejtig, Ruth Bagley, Mike Horn, and Uri Wilensky. 2024. Learning Agent-based Modeling with LLM Companions: Experiences of Novices and Experts Using ChatGPT & NetLogo Chat. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 141, 18 pages. doi:10.1145/3613904.3642377

[22] Zhifa Chen, Yichen Lu, Mika P Nieminen, and Andrés Lucero. 2020. Creating a Chatbot for and with Migrants: Chatbot Personality Drives Co-Design Activities. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference*. 219–230.

[23] Andrea Cuadra, Maria Wang, Lynn Andrea Stein, Malte F Jung, Nicola Dell, Deborah Estrin, and James A Landay. 2024. The illusion of empathy? notes on displays of emotion in human-computer interaction. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–18.

[24] Keren Dali. 2022. A calling, not a call of duty: Public librarians' engagement with immigrant communities. *Journal of Library Administration* 62, 2 (2022), 206–234.

[25] Sumit Kumar Dam, Choong Seon Hong, Yu Qiao, and Chaoning Zhang. 2024. A complete survey on llm-based ai chatbots. *arXiv preprint arXiv:2406.16937* (2024).

[26] Giovanna Deiana, Marco Dettori, Antonella Arghittu, Antonio Azara, Giovanni Gabutti, and Paolo Castiglia. 2023. Artificial Intelligence and Public Health: Evaluating ChatGPT Responses to Vaccination Myths and Misconceptions. *Vaccines* 11, 7 (2023), 1217.

[27] Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. 2023. Toxicity in chatgpt: Analyzing persona-assigned language models. *arXiv preprint arXiv:2304.05335* (2023).

[28] Economic Policy Institute. 2024. Another Challenge for the Unemployed: Public Library Budget Cuts. (2024). https://www.epi.org/publication/another_challenge_for_the_unemployed_public_library_budget_cuts/ Accessed: 2024-09-10.

[29] Michelle Ehrenpreis and J DeLooper. 2022. Implementing a chatbot on a library website. *Journal of Web Librarianship* 16, 2 (2022), 120–142.

[30] Sabit Ekin. 2023. Prompt engineering for ChatGPT: a quick guide to techniques, tips, and best practices. *Authorea Preprints* (2023).

[31] Luciano Floridi. 2019. Establishing the rules for building trustworthy AI. *Nature Machine Intelligence* 1, 6 (2019), 261–262.

[32] Asbjørn Følstad, Theo Araujo, Effie Lai-Chong Law, Petter Bae Brandtzaeg, Symeon Papadopoulos, Lea Reis, Marcos Baez, Guy Laban, Patrick McAllister, Carolin Ischen, et al. 2021. Future directions for chatbot research: an interdisciplinary research agenda. *Computing* (2021), 1–28.

[33] Coalition for Networked Information (CNI). [n. d.]. Enhancing Library Services with Conversational Artificial Intelligence. https://www.cni.org/topics/information-access-retrieval/enhancing-library-services-with-conversational-artificial-intelligence. Accessed: 2024-12-05.

[34] Jianfeng Gao, Michel Galley, and Lihong Li. 2018. Neural approaches to conversational ai. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. 1371–1374.

[35] Lorentsa Gkinko and Amany Elbanna. 2023. The appropriation of conversational AI in the workplace: A taxonomy of AI chatbot users. *International Journal of Information Management* 69 (2023), 102568.

[36] Jonathan Grudin and Richard Jacques. 2019. Chatbots, humbots, and the quest for artificial general intelligence. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–11.

[37] Juhye Ha, Hyeon Jeon, Daeun Han, Jinwook Seo, and Changhoon Oh. 2024. CloChat: Understanding How People Customize, Interact, and Experience Personas in Large Language Models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 305, 24 pages. doi:10.1145/3613904.3642472

[38] Juhye Ha, Hyeon Jeon, DaEun Han, Jinwook Seo, and Changhoon Oh. 2024. CloChat: Understanding How People Customize, Interact, and Experience Personas in Large Language Models. *arXiv preprint arXiv:2402.15265* (2024).

[39] Xu Han, Michelle Zhou, Matthew J Turner, and Tom Yeh. 2021. Designing effective interview chatbots: Automatic chatbot profiling and design suggestion generation for chatbot debugging. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–15.

[40] Dessy Harisanty, Nove E Variant Anna, Tesa Eranti Putri, Aji Akbar Firdaus, and Nurul Aida Noor Azizi. 2022. Leaders, practitioners and scientists' awareness of artificial intelligence in libraries: a pilot study. *Library Hi Tech* (2022).

[41] Dessy Harisanty, Nove E Variant Anna, Tesa Eranti Putri, Aji Akbar Firdaus, and Nurul Aida Noor Azizi. 2024. Leaders, practitioners and scientists' awareness of artificial intelligence in libraries: a pilot study. *Library Hi Tech* 42, 3 (2024), 809–825.

[42] Paul Henman. 2020. Improving public services using artificial intelligence: possibilities, pitfalls, governance. *Asia Pacific Journal of Public Administration* 42, 4 (2020), 209–221.

[43] Alison Hicks and Annemaree Lloyd. 2022. Relegating expertise: The outward and inward positioning of librarians in information literacy education. *Journal of Librarianship and Information Science* 54, 3 (2022), 415–426.

[44] Andreas Holzinger, Michaela Kargl, Bettina Kipperer, Peter Regitnig, Markus Plass, and Heimo Müller. 2022. Personas for Artificial Intelligence (AI) an Open Source Toolbox. *IEEE Access* 10 (2022), 23732–23747. doi:10.1109/ACCESS.2022.3154776

[45] Sungsoo Ray Hong, Jessica Hullman, and Enrico Bertini. 2020. Human Factors in Model Interpretability: Industry Practices, Challenges, and Needs. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW1, Article 68 (May 2020), 26 pages. doi:10.1145/3392878

[46] Jingshan Huang and Ming Tan. 2023. The role of ChatGPT in scientific communication: writing better scientific review articles. *American journal of cancer research* 13, 4 (2023), 1148.

[47] Yun Huang, Brian Dobreski, and Huichuan Xia. 2017. Human Library: Understanding Experience Sharing for Community Knowledge Building. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Portland, Oregon, USA) *(CSCW '17)*. Association for Computing Machinery, New York, NY, USA, 1152–1165. doi:10.1145/2998181.2998312

[48] Shafquat Hussain, Omid Ameri Sianaki, and Nedal Ababneh. 2019. A survey on conversational agents/chatbots classification and design techniques. In *Workshops of the International Conference on Advanced Information Networking and Applications*. Springer, 946–956.

[49] Edward Iglesias. 2014. The Inevitability of Library Automation. In *Robotics: Concepts, Methodologies, Tools, and Applications*. IGI Global, 166–177.

[50] Paul T Jaeger and John Carlo Bertot. 2009. E-government education in public libraries: New service roles and expanding social responsibilities. *Journal of Education for Library and Information Science* (2009), 39–49.

[51] Eunkyung Jo, Daniel A. Epstein, Hyunhoon Jung, and Young-Ho Kim. 2023. Understanding the Benefits and Challenges of Deploying Conversational AI Leveraging Large Language Models for Public Health Intervention. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 18, 16 pages. doi:10.1145/3544548.3581503

[52] Ishika Joshi, Simra Shahid, Shreeya Venneti, Manushree Vasu, Yantao Zheng, Yunyao Li, Balaji Krishnamurthy, and Gromit Yeuk-Yin Chan. 2024. CoPrompter: User-Centric Evaluation of LLM Instruction Alignment for Improved Prompt Engineering. *arXiv preprint arXiv:2411.06099* (2024).

[53] Dinesh Kalla, Nathan Smith, Fnu Samaah, and Sivaraju Kuraku. 2023. Study and analysis of chat GPT and its impact on different fields of study. *International journal of innovative science and research technology* 8, 3 (2023).

[54] Danielle Kane. 2019. Creating, managing and analyzing an academic library chatbot. *BiD: textos universitaris de biblioteconomia i documentació* 43 (2019).

[55] Anna Kawakami, Venkatesh Sivaraman, Hao-Fei Cheng, Logan Stapleton, Yanghuidi Cheng, Diana Qing, Adam Perer, Zhiwei Steven Wu, Haiyi Zhu, and Kenneth Holstein. 2022. Improving Human-AI Partnerships in Child Welfare: Understanding Worker Practices, Challenges, and Desires for Algorithmic Decision Support. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 52, 18 pages. doi:10.1145/3491102.3517439

[56] Irene Kitsara. 2022. Artificial intelligence and the digital divide: From an innovation perspective. In *Platforms and artificial intelligence: The next generation of competences*. Springer, 245–265.

[57] Lorenz Cuno Klopfenstein, Saverio Delpriori, Silvia Malatini, and Alessandro Bogliolo. 2017. The rise of bots: A survey of conversational interfaces, patterns, and paradigms. In *Proceedings of the 2017 conference on designing interactive systems*. 555–565.

[58] Amy J Ko, Brad A Myers, and Htet Htet Aung. 2004. Six learning barriers in end-user programming systems. In *2004 IEEE Symposium on Visual Languages-Human Centric Computing*. IEEE, 199–206.

[59] Alexander Kocur, Sünje Clausen, Lennart Hofeditz, Felix Brünker, Jennifer Fromm, and Stefan Stieglitz. 2023. FIGHTING FALSE INFORMATION-DESIGNING A CONVERSATIONAL AGENT FOR PUBLIC SECTOR ORGANIZATIONS. (2023).

[60] Guy Laban. 2021. Perceptions of Anthropomorphism in a Chatbot Dialogue: The Role of Animacy and Intelligence. In *Proceedings of the 9th International Conference on Human-Agent Interaction* (Virtual Event, Japan) *(HAI '21)*. Association for Computing Machinery, New York, NY, USA, 305–310. doi:10.1145/3472307.3484686

[61] Michelle S Lam. 2024. Granting Non-AI Experts Creative Control Over AI Systems. In *Adjunct Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*. 1–5.

[62] Michelle S. Lam. 2024. Granting Non-AI Experts Creative Control Over AI Systems. In *Adjunct Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology* (Pittsburgh, PA, USA) *(UIST Adjunct '24)*. Association for Computing Machinery, New York, NY, USA, Article 61, 5 pages. doi:10.1145/3672539.3686714

[63] Raina Langevin, Ross J Lordon, Thi Avrahami, Benjamin R. Cowan, Tad Hirsch, and Gary Hsieh. 2021. Heuristic Evaluation of Conversational Agents. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (<conf-loc>, <city>Yokohama</city>, <country>Japan</country>, </conf-loc>) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 632, 15 pages. doi:10.1145/3411764.3445312

[64] Yrjo Lappalainen and Nikesh Narayanan. 2023. Aisha: A custom AI library chatbot using the ChatGPT API. *Journal of Web Librarianship* 17, 3 (2023), 37–58.

[65] Ronald Larson. 2019. Library stakeholder attitudes and new technology. *Library Management* 40, 1/2 (2019), 128–140.

[66] Jung-Joo Lee, Christine Ee Ling Yap, and Virpi Roto. 2022. How HCI Adopts Service Design: Unpacking current perceptions and scopes of service design in HCI and identifying future opportunities. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 530, 14 pages. doi:10.1145/3491102.3502128

[67] Mina Lee, Megha Srivastava, Amelia Hardy, John Thickstun, Esin Durmus, Ashwin Paranjape, Ines Gerard-Ursin, Xiang Lisa Li, Faisal Ladhak, Frieda Rong, et al. 2022. Evaluating human-language model interaction. *arXiv preprint arXiv:2212.09746* (2022).

[68] Liisi Lembinen. 2021. Innovation in European academic libraries–Leadership perspective. *Journal of Library Administration* 61, 8 (2021), 921–935.

[69] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp

tasks. *Advances in Neural Information Processing Systems* 33 (2020), 9459–9474.
[70] Chia-Ying Li and Jin-Ting Zhang. 2023. Chatbots or me? Consumers' switching between human agents and conversational agents. *Journal of Retailing and Consumer Services* 72 (2023), 103264.
[71] Kate Lister, Tim Coughlan, Francisco Iniesto, Nick Freear, and Peter Devine. 2020. Accessible conversational user interfaces: considerations for design. In *Proceedings of the 17th international web for all conference*. 1–11.
[72] Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, et al. 2023. Agentbench: Evaluating llms as agents. *arXiv preprint arXiv:2308.03688* (2023).
[73] Yilun Liu, Minggui He, Feiyu Yao, Yuhe Ji, Shimin Tao, Jingzhou Du, Duan Li, Jian Gao, Li Zhang, Hao Yang, et al. 2024. What Do You Want? User-centric Prompt Generation for Text-to-image Synthesis via Multi-turn Guidance. *arXiv preprint arXiv:2408.12910* (2024).
[74] ManyChat. [n. d.]. ManyChat: Chat Marketing Automation Platform. https://manychat.com/. Accessed: 2024-12-10.
[75] Todor Markov, Chong Zhang, Sandhini Agarwal, Florentine Eloundou Nekoul, Theodore Lee, Steven Adler, Angela Jiang, and Lilian Weng. 2023. A holistic approach to undesired content detection in the real world. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 15009–15018.
[76] Joseph R Matthews. 2017. *The evaluation and measurement of library services*. ABC-CLIO.
[77] Andy Mayo, Jacob. 2023. GPT-4 and LangChain - Create a ChatGPT Chatbot for Your PDF Files. https://github.com/mayooear/gpt4-pdf-chatbot-langchain/tree/feat/vectordbqa
[78] Indra Ayu Susan Mckie and Bhuva Narayan. 2019. Enhancing the academic library experience with chatbots: An exploration of research and implications for practice. *Journal of the Australian Library and Information Association* 68, 3 (2019), 268–277.
[79] Gaspar Isaac Melsion, Rebecca Stower, Katie Winkle, and Iolanda Leite. 2023. What's at Stake? Robot explanations matter for high but not low-stake scenarios. In *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2421–2426.
[80] Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. 2021. Recent advances in natural language processing via large pre-trained language models: A survey. *Comput. Surveys* (2021).
[81] Quim Motger, Xavier Franch, and Jordi Marco. 2021. Conversational Agents in Software Engineering: Survey, Taxonomy and Challenges. *arXiv preprint arXiv:2106.10901* (2021).
[82] Sydney Nguyen, Hannah McLean Babe, Yangtian Zi, Arjun Guha, Carolyn Jane Anderson, and Molly Q Feldman. 2024. How Beginning Programmers and Code LLMs (Mis)read Each Other. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 651, 26 pages. doi:10.1145/3613904.3642706
[83] Marcia Katharina Nißen, Driton Selimi, Antje Janssen, Davinia Rodríguez Cardona, Michael H Breitner, Tobias Kowatsch, and Florian von Wangenheim. 2021. See you soon again, chatbot? A design taxonomy to characterize user-chatbot relationships with different time horizons. *Computers in Human Behavior* (2021), 107043.
[84] Claudio Novelli, Federico Casolari, Antonino Rotolo, Mariarosaria Taddeo, and Luciano Floridi. 2024. Taking AI risks seriously: a new assessment model for the AI Act. *AI & SOCIETY* 39, 5 (2024), 2493–2497.
[85] University of Illinois Library. [n. d.]. READ Scale. https://www.library.illinois.edu/staff/information-desk-resource-page/read-scale/. Accessed: 2024-12-05.
[86] Hyanghee Park, Daehwan Ahn, Kartik Hosanagar, and Joonhwan Lee. 2022. Designing Fair AI in Human Resource Management: Understanding Tensions Surrounding Algorithmic Evaluation and Envisioning Stakeholder-Centered Solutions. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 51, 22 pages. doi:10.1145/3491102.3517672
[87] Hyanghee Park and Joonhwan Lee. 2021. Designing a Conversational Agent for Sexual Assault Survivors: Defining Burden of Self-Disclosure and Envisioning Survivor-Centered Solutions. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 634, 17 pages. doi:10.1145/3411764.3445133
[88] Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. arXiv:2304.03442 [cs.HC]
[89] Marija Petek. 2018. Stress among reference library staff in academic and public libraries. *Reference Services Review* 46, 1 (2018), 128–145.
[90] Kanchanjot Kaur Phokela, Samarth Sikand, Kapil Singi, Kuntal Dey, Vibhu Saujanya Sharma, and Vikrant Kaulgud. 2023. Smart Prompt Advisor: Multi-Objective Prompt Framework for Consistency and Best Practices. In *2023 38th IEEE/ACM International Conference on Automated Software Engineering (ASE)*. IEEE, 1846–1848.
[91] Martin Porcheron, Joel E Fischer, Stuart Reeves, and Sarah Sharples. 2018. Voice interfaces in everyday life. In *proceedings of the 2018 CHI conference on human factors in computing systems*. 1–12.
[92] Kurtis Pykes. 2023. How To Create A Conversational Agent with Dialogflow. (2023). https://towardsdatascience.com/how-to-create-a-conversational-agent-with-dialogflow-17bfa90aa02d
[93] Amon Rapp, Lorenzo Curti, and Arianna Boldi. 2021. The human side of human-chatbot interaction: A systematic literature review of ten years of research on text-based chatbots. *International Journal of Human-Computer Studies* (2021), 102630.
[94] Brian Real and R Norman Rose. 2017. Rural libraries in the United States: Recent strides, future possibilities, and meeting community needs. (2017).
[95] Sharesly Rodriguez and Christina Mune. 2022. Uncoding library chatbots: Deploying a new virtual reference tool at the San Jose State University library. *Reference Services Review* 50, 3/4 (2022), 392–405.
[96] Virpi Roto, Jung-Joo Lee, Tuuli Mattelmäki, and John Zimmerman. 2018. Experience Design meets Service Design: Method Clash or Marriage?. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–6.
[97] Heleen Rutjes, Martijn Willemsen, and Wijnand IJsselsteijn. 2019. Considerations on explainable AI and users' mental models. In *CHI 2019 Workshop: Where is the human? Bridging the gap between AI and HCI*. Association for Computing Machinery, Inc.
[98] Malak Sadek, Rafael A Calvo, and Celine Mougenot. 2023. Trends, challenges and processes in conversational agent design: exploring practitioners' views through semi-structured interviews. In *Proceedings of the 5th International Conference on Conversational User Interfaces*. 1–10.
[99] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. In *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–15.
[100] Daniela Sangiorgi. 2009. Building up a framework for Service Design research. In *8th European Academy of Design conference*. Citeseer, 415–420.
[101] Majideh Sanji, Hassan Behzadi, and Gisu Gomroki. 2022. Chatbot: an intelligent tool for libraries. *Library Hi Tech News* 39, 3 (2022), 17–20.
[102] Robson Santos, Italo Santos, Cleyton Magalhaes, and Ronnie de Souza Santos. 2024. Are We Testing or Being Tested? Exploring the Practical Applications of Large Language Models in Software Testing. In *2024 IEEE Conference on Software Testing, Verification and Validation (ICST)*. IEEE, 353–360.
[103] Maxime Sasseville, Annie LeBlanc, Mylène Boucher, Michèle Dugas, Gisele Mbemba, Jack Tchuente, Maud-Christine Chouinard, Marianne Beaulieu, Nicolas Beaudet, Becky Skidmore, et al. 2021. Digital health interventions for the management of mental health in people with chronic diseases: a rapid review. *BMJ open* 11, 4 (2021), e044437.
[104] Laura Saunders and Mary Jordan. 2013. Significantly different? Reference services competencies in public and academic libraries. *Reference and User Services Quarterly* 52, 3 (2013), 216–223.
[105] Rachel Scott. 2011. The role of public libraries in community building. *Public Library Quarterly* 30, 3 (2011), 191–227.
[106] Katie Seaborn, Norihisa P Miyake, Peter Pennefather, and Mihoko Otake-Matsuura. 2021. Voice in Human–Agent Interaction: A Survey. *ACM Computing Surveys (CSUR)* 54, 4 (2021), 1–43.
[107] Joseph Seering, Juan Pablo Flores, Saiph Savage, and Jessica Hammer. 2018. The social roles of bots: evaluating impact of bots on discussions in online communities. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–29.
[108] Walid Shalaby, Adriano Arantes, Teresa GonzalezDiaz, and Chetan Gupta. 2020. Building chatbots from large scale domain-specific knowledge bases: Challenges and opportunities. In *2020 IEEE International Conference on Prognostics and Health Management (ICPHM)*. IEEE, 1–8.
[109] Hong Shen, Tianshi Li, Toby Jia-Jun Li, Joon Sung Park, and Diyi Yang. 2023. Shaping the Emerging Norms of Using Large Language Models in Social Computing Research. In *Computer Supported Cooperative Work and Social Computing*. 569–571.
[110] Ben Shneiderman. 2020. Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human-Computer Interaction* 36, 6 (2020), 495–504.
[111] Ben Shneiderman. 2022. *Human-centered AI*. Oxford University Press.
[112] Alison Smith, Varun Kumar, Jordan Boyd-Graber, Kevin Seppi, and Leah Findlater. 2018. Closing the Loop: User-Centered Design and Evaluation of a Human-in-the-Loop Topic Modeling System. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces* (Tokyo, Japan) *(IUI '18)*. Association for Computing Machinery, New York, NY, USA, 293–304. doi:10.1145/3172944.3172965
[113] Paul Sturges, Eric Davies, James Dearnley, Ursula Iliffe, Ursula Iliffe, Charles Oppenheim, and Rachel Hardy. 2003. User privacy in the digital library environment: an investigation of policies and preparedness. *Library Management* 24, 1/2 (2003), 44–50.

[114] Hari Subramonyam, Roy Pea, Christopher Pondoc, Maneesh Agrawala, and Colleen Seifert. 2024. Bridging the Gulf of Envisioning: Cognitive Challenges in Prompt Based Interactions with LLMs. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 1039, 19 pages. doi:10.1145/3613904.3642754

[115] Maxwell Szymanski, Katrien Verbert, and Vero Vanden Abeele. 2022. Designing and evaluating explainable AI for non-AI experts: challenges and opportunities. In *Proceedings of the 16th ACM Conference on Recommender Systems*. 735–736.

[116] Richard Tomsett, Alun Preece, Dave Braines, Federico Cerutti, Supriyo Chakraborty, Mani Srivastava, Gavin Pearson, and Lance Kaplan. 2020. Rapid trust calibration through interpretable and uncertainty-aware AI. *Patterns* 1, 4 (2020).

[117] Urban Libraries Council. 2023. Public Libraries Set the Stage for Integration of Artificial Intelligence in Their Services and Resources. (2023). https://www.urbanlibraries.org/newsroom/release-public-libraries-set-the-stage-for-integration-of-artificial-intelligence-in-their-services-and-resources Accessed: 2024-09-10.

[118] Maarten Van Someren, Yvonne F Barnard, and J Sandberg. 1994. The think aloud method: a practical approach to modelling cognitive. *London: AcademicPress* 11, 6 (1994).

[119] Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. 2023. A survey on large language model based autonomous agents. *arXiv preprint arXiv:2308.11432* (2023).

[120] Zhenhailong Wang, Shaoguang Mao, Wenshan Wu, Tao Ge, Furu Wei, and Heng Ji. 2023. Unleashing Cognitive Synergy in Large Language Models: A Task-Solving Agent through Multi-Persona Self-Collaboration. arXiv:2307.05300 [cs.AI]

[121] Chen Wei, Zhichen Yu, and Simon Fong. 2018. How to build a chatbot: chatbot framework and its capabilities. In *Proceedings of the 2018 10th international conference on machine learning and computing*. 369–373.

[122] Tongshuang Wu, Michael Terry, and Carrie Jun Cai. 2022. Ai chains: Transparent and controllable human-ai interaction by chaining large language model prompts. In *Proceedings of the 2022 CHI conference on human factors in computing systems*. 1–22.

[123] Qian Yang, Justin Cranshaw, Saleema Amershi, Shamsi T Iqbal, and Jaime Teevan. 2019. Sketching nlp: A case study of exploring the right things to design with language intelligence. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.

[124] Qian Yang, Yuexing Hao, Kexin Quan, Stephen Yang, Yiran Zhao, Volodymyr Kuleshov, and Fei Wang. 2023. Harnessing Biomedical Literature to Calibrate Clinicians' Trust in AI Decision Support Systems. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (<conf-loc>, <city>Hamburg</city>, <country>Germany</country>, </conf-loc>) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 14, 14 pages. doi:10.1145/3544548.3581393

[125] Qian Yang, Alex Scuito, John Zimmerman, Jodi Forlizzi, and Aaron Steinfeld. 2018. Investigating How Experienced UX Designers Effectively Work with Machine Learning. In *Proceedings of the 2018 Designing Interactive Systems Conference* (Hong Kong, China) *(DIS '18)*. Association for Computing Machinery, New York, NY, USA, 585–596. doi:10.1145/3196709.3196730

[126] Qian Yang, Alex Scuito, John Zimmerman, Jodi Forlizzi, and Aaron Steinfeld. 2018. Investigating how experienced UX designers effectively work with machine learning. In *Proceedings of the 2018 designing interactive systems conference*. 585–596.

[127] Qian Yang, Aaron Steinfeld, and John Zimmerman. 2019. Unremarkable AI: Fitting Intelligent Decision Support into Critical, Clinical Decision-Making Processes. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–11. doi:10.1145/3290605.3300468

[128] Qian Yang, Jina Suh, Nan-Chen Chen, and Gonzalo Ramos. 2018. Grounding interactive machine learning tool design in how non-experts actually build models. In *Proceedings of the 2018 designing interactive systems conference*. 573–584.

[129] Nur Yildirim, Changhoon Oh, Deniz Sayar, Kayla Brand, Supritha Challa, Violet Turri, Nina Crosby Walton, Anna Elise Wong, Jodi Forlizzi, James McCann, et al. 2023. Creating design resources to scaffold the ideation of AI concepts. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference*. 2326–2346.

[130] Nur Yildirim, Changhoon Oh, Deniz Sayar, Kayla Brand, Supritha Challa, Violet Turri, Nina Crosby Walton, Anna Elise Wong, Jodi Forlizzi, James McCann, and John Zimmerman. 2023. Creating Design Resources to Scaffold the Ideation of AI Concepts. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference* (Pittsburgh, PA, USA) *(DIS '23)*. Association for Computing Machinery, New York, NY, USA, 2326–2346. doi:10.1145/3563657.3596058

[131] Daisy Yoo, Anya Ernest, Sofia Serholt, Eva Eriksson, and Peter Dalsgaard. 2019. Service Design in HCI Research: The Extended Value Co-creation Model. In *Proceedings of the Halfway to the Future Symposium 2019* (Nottingham, United Kingdom) *(HTTF 2019)*. Association for Computing Machinery, New York, NY, USA, Article 17, 8 pages. doi:10.1145/3363384.3363401

[132] JD Zamfirescu-Pereira, Bjoern Hartmann, and Qian Yang. 2023. Conversation Regression Testing: A Design Technique for Prototyping Generalizable Prompt Strategies for Pre-trained Language Models. *arXiv preprint arXiv:2302.03154* (2023).

[133] JD Zamfirescu-Pereira, Richmond Y Wong, Bjoern Hartmann, and Qian Yang. 2023. Why Johnny can't prompt: how non-AI experts try (and fail) to design LLM prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–21.

[134] Fabio Massimo Zanzotto. 2019. Human-in-the-loop artificial intelligence. *Journal of Artificial Intelligence Research* 64 (2019), 243–252.

[135] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223* (2023).

[136] Qingxiao Zheng and Yun Huang. 2023. " Begin with the End in Mind": Incorporating UX Evaluation Metrics into Design Materials of Participatory Design. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–7.

[137] Qingxiao Zheng, Daniela M Markazi, Yiliu Tang, and Yun Huang. 2021. " PocketBot Is Like a Knock-On-the-Door!": Designing a Chatbot to Support Long-Distance Relationships. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–28.

[138] Qingxiao Zheng, Yiliu Tang, Yiren Liu, Weizi Liu, and Yun Huang. 2022. UX research on conversational human-AI interaction: A literature review of the ACM digital library. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–24.

[139] John Zimmerman and Jodi Forlizzi. 2014. Service Design. https://www.interaction-design.org/literature/book/the-encyclopedia-of-human-computer-interaction-2nd-ed/service-design. Retrieved May 21, 2024.

A Appendix

A.1 A Snippet of Simulated Dialogues

Simulated Dialogue with Persona

Character A:

Hey there! I have some questions about the study room policy at Fountaindale.

AI:

Sure, I'll do my best to help you. What questions do you have about the study room policy at Fountaindale?

Character A:

Thanks for offering your help! First off, are the study rooms at Fountaindale available for all students or do we need to reserve them in advance?

AI:

The study rooms at Fountaindale Public Library are available for eligible cardholders on a first-come, first-served basis. However, Fountaindale Library cardholders can reserve study rooms in advance through the library's website. Pinnacle Library Cooperative and reciprocal borrower cardholders need to call the Reference Desk to make a reservation in advance.

Character A:

Got it! So, if I have a Fountaindale Library card, I can reserve a study room online, right?

AI:

Yes, Fountaindale Library cardholders can reserve study rooms online.

Character A:

That's great! Is there a time limit for how long we can use the study rooms at Fountaindale?

Simulated Dialogue without persona

Character A:

Okay, I can do that! Hi, excuse me, do you know where the comic books are? I've been looking for them for a while and can't seem to find them.

AI:

Sure, I can do that! Hello there, how can I assist you today? Are you looking for a particular book or do you need help with something else?

Character A:

Okay, I can play along! Excuse me, do you have any books about dinosaurs? I need it for a school project and I need it now. Can you help me find it?

AI:

Sure, I can do that! Hello there, how can I assist you today? Are you looking for a particular book or do you need help with something else?

Figure 6: Simulated dialogues generated by AI agents with predetermined personas and those without