Reference Services Review

Understanding academic library virtual reference at scale: a multi-institutional content analysis of chat transcripts

Margaret Sinclair Brown-Sica and Rebecca A. Croxton
CSU Libraries, Colorado State University, Fort Collins, Colorado, USA, and
Jennifer Church-Duran
UA Libraries, University of Arizona, Tucson, Arizona, USA

Received 29 April 2026
Revised 22 June 2026
Accepted 27 June 2026

Abstract

Purpose – This study aims to examine the types of questions and requests submitted through academic library chat reference services, how transaction types and complexity vary across institutions and which interactions may be appropriate for automated or artificial intelligence (AI)-assisted response.

Design/methodology/approach – This study employs qualitative content analysis of 34,977 anonymized chat transcripts (2021–2024) from five US R1 universities. A hybrid stratified random sample (n = 1,746) was manually coded using a collaboratively developed 42-code codebook derived through deductive and inductive methods. Four coders applied up to three codes per transcript, with two rounds of intercoder reliability testing (Krippendorff's α, Fleiss' κ and Cohen's κ) to establish codebook stability. Transaction complexity was operationalized using tertiles based on interaction duration and word count (simple, moderate, complex). Descriptive statistics were used to analyze code frequencies, complexity distributions and cross-institutional variation.

Findings – A small set of transaction types account for most interactions. Three categories, finding known journal articles, identifying relevant sources and interlibrary loan, appear consistently across all institutions and are predominantly complex. Overall, 38% of interactions were classified as complex, indicating substantial demand for high-level research support. Institutional variation was observed, particularly in access-services transactions.

Originality/value – No large-scale, multi-institutional, human-coded analysis of academic library chat transactions using a granular transaction taxonomy currently exists in the published literature. This manuscript introduces a replicable transaction taxonomy and complexity framework to inform evidence-based service design and AI integration.

Keywords Academic libraries, Virtual reference services, Chat reference, Content analysis, Transaction classification, Intercoder reliability, Research support services, User behavior, Service design, Artificial intelligence (AI), Automation, Complexity analysis

Paper type Research article

Introduction

Chat reference has become a primary service channel at academic libraries. It is available at all hours and accessible from anywhere, giving users quick access to library services without having to be at a particular place, make an appointment, or visit a reference desk. Its adoption in academic libraries is now effectively universal. A recent survey of academic library websites found that the vast majority offered live chat services (Guy *et al.*, 2023), and the ACRL Academic Library Trends and Statistics survey consistently shows that virtual reference is among the most frequently reported service interactions at academic institutions (ACRL, 2024). Despite this widespread adoption and use, the empirical landscape of what

Reference Services Review
Emerald Publishing Limited
e-ISSN: 2054-1716
p-ISSN: 0090-7324
DOI 10.1108/RSR-04-2026-0078

actually occurs in these interactions, what users need, how complex the interactions are, and how they vary and are similar across institutions is surprisingly incomplete. Most previous transcript studies have been at single institutions and on a relatively small scale, which makes it challenging to see what is similar across all libraries versus what is particular to one or a few.

The need to fill this gap has increased as AI tools are increasingly considered and, in some cases, used as potential solutions to provide chat reference services. Library administrators face growing pressure to automate some reference services. However, planning these services is difficult without substantial empirical evidence. A recent large-scale study by Luo and Brissett (2025) classified over 88,000 academic library chat transcripts using a natural language processing approach, advancing understanding of chat reference at scale and demonstrating the promise of computational methods for this kind of analysis. The present study takes a complementary approach, applying human coding and intercoder reliability testing across five institutions using a 42-code taxonomy designed collaboratively by library practitioners. Where Luo and Brissett prioritized scale through computational classification, this study prioritizes granularity, cross-institutional comparison, and expert validation, offering a different empirical foundation for evidence-based service design.

Expanding prior work in this area, this study applies human coding and explores the use of AI-based coding across a large-scale dataset of more than 34,000 interactions drawn from five public research universities.

Three research questions guide the study:

(1) What types of questions and requests do users submit via chat reference services at academic research libraries, and how frequently does each transaction type occur?

(2) How do transaction types and complexity vary across institutions, and what factors may account for that variation?

(3) Which transaction types, based on their frequency and complexity profiles, are most appropriate candidates for automated or AI-assisted response?

The article begins with a review of relevant literature, followed by a description of the methodology, findings, and a discussion of implications for service design and AI integration.

Literature review

The evolution of chat transcript analysis

For decades, chat transcript analysis has served as a foundational method for assessing library services. As a well-documented and widely adopted medium, chat services provide a comprehensive, longitudinal record of librarian-patron interactions. This data offers a unique ethnographic window into user behavior that is often unavailable or difficult to capture in spontaneous face-to-face interactions.

While many scholars categorize these platforms under the umbrella of “Virtual Reference Services” (VRS), research consistently demonstrates that a significant portion of user inquiries falls outside the traditional “ready reference” or in-depth research framework. Instead, patrons frequently utilize chat for circulation issues, policy clarifications, technical troubleshooting, and general community inquiries. For instance, De Leon and Viray (2022) found that 81% of inquiries at De La Salle University over a two-year period pertained to technical support and library policies rather than traditional reference transactions. Because users expect a seamless experience regardless of internal service labels, understanding these usage patterns (both globally and locally) is essential for service optimization. This understanding is particularly critical as libraries transition toward technology-based supports, such as scripted automated chatbots or retrieval-augmented generation (RAG) enabled AI assistants.

Methodological trends in content analysis

Research in this domain typically begins with deductive content analysis, using question-type coding as a baseline. This foundation is often supplemented by incorporating additional analysis through multifaceted variables. Early efforts often paired transcript coding with external data points, such as post-chat surveys, quantitative usage statistics, and patron demographics to establish benchmarks for these then-nascent services (Arnold and Kaske, 2005; Broughton, 2003; Foley, 2002; Kibbee *et al.*, 2002; Lankes *et al.*, 2003; McClure *et al.*, 2003; Sears, 2001).

As the service matured, the variables analyzed grew in complexity. Scholars began incorporating behavioral performance and accuracy (Maximiek *et al.*, 2010), qualitative focus groups (Rourke and Lupien, 2010), and linguistic patterns in patron behavior (Rourke and Lupien, 2010). More recent work has combined question type coding analysis with such assessments as pre-chat surveys and metadata about the chat (Logan *et al.*, 2019) as well as optional exit surveys, focus groups, and individual student or patron interviews (Hervieux, 2021; Jacoby *et al.*, 2016; Kathuria, 2021; Mawhinney and Hervieux, 2022).

Other studies have leveraged the Reference and User Services (RUSA) Guidelines as a tool to evaluate staff performance (Zhuo *et al.*, 2006) or utilized the Reference, Effort, Assessment and Data (READ) Scale to quantify the intellectual difficulty of transactions (Maloney and Kemp, 2015; Mavodza, 2019). There has also been critical focus on the need to develop instructional strategies for assisting patrons in reference-based chat services (Devlin *et al.*, 2008; Hervieux and Tummon, 2018; Oakleaf and VanScoy, 2010; Valentine and Moss, 2017). Most recently, research has shifted toward establishing best practices for staff performance (Everdeen Moore *et al.*, 2026) and optimizing referral workflows to subject specialists (Dempsey *et al.*, 2024).

The challenge of standardization and granularity

Despite the volume of literature, the lack of a shared, standardized coding scheme remains a persistent challenge for conducting comparative research. As early as 2003, Marsteller and Mizzy advocated for the creation of a common taxonomy to facilitate cross-institutional comparisons. However, no such universal tool has emerged. In its absence, various tiered models have been proposed.

Marsteller and Mizzy (2003) utilized four levels ranging from "directional" to "reference." Houlson *et al.* (2007) developed a 10-category schema with robust subcategories focusing heavily on material formats and item types, though it largely excluded non-library inquiries. Wan *et al.* (2009) and Matteson *et al.* (2011) employed tiered approaches that first distinguished between reference and non-reference before sub-categorizing (e.g. policy, technical, or directional).

Youngbar's (2012) study remains foundational because its schema successfully integrated non-library and technical queries. Unlike the format-heavy hierarchy of Houlson *et al.* (2007), Youngbar (2012) prioritized the nature of the task and interaction complexity. Nevertheless, recent studies (Everdeen Moore *et al.*, 2026; Meert-Williston and Sandieson, 2019; Radford and Connaway, 2013) continue to favor smaller, more manageable code sets.

Addressing coding constraints and the "single intent" gap

The tendency to use restricted code sets is largely a response to the labor-intensive nature of manual content analysis. Reviewing unstructured chat data is time-consuming and resource-heavy (Kavitha and Kavitha, 2023). To manage this cognitive workload, researchers often limit the size of the coding schema and restrict transaction classifications to a to a primary question, purpose or intent, or a single code per transcript, regardless of how many distinct questions a patron might ask.

While efficient, this "single intent" approach unintentionally flattens the complexity of the conversation. It overlooks the dynamic conversational pathways and the interplay of diverse

needs within a single interaction. To address these limitations, we introduce a 42-item codebook that accounts for the significant volume of non-reference activity in chat interactions. Furthermore, by allowing up to three distinct codes per transcript, the approach captures the multifaceted nature of chat interactions, generating a more nuanced dataset to inform the development of next-generation AI support tools.

Methods

This study employed a qualitative, descriptive–exploratory research design to examine academic library chat reference interactions across multiple institutions. Qualitative content analysis was used to develop and apply a structured 42-item codebook capturing user needs, transaction types, and interaction characteristics within chat transcripts. Although qualitative in its primary analytic orientation, the study incorporated quantitative analytic techniques at multiple stages to support methodological rigor and interpretation. Specifically, frequencies and percentage distributions were calculated to identify the most frequently applied codes and to characterize levels of interaction complexity across the coded sample. In addition, quantitative statistical measures were used during codebook testing and application to assess intercoder reliability and codebook stability prior to full coding.

Positionality of the researchers

The research team consisted of four researchers representing two R1 institutions in the United States: three affiliated with Colorado State University and one affiliated with the University of Arizona. Three team members are academic librarians holding graduate degrees in library and information science, each with more than 2 decades of professional experience in academic libraries; the fourth member is a retired higher-education professor whose career spans subject areas of expertise outside of librarianship. Two team members hold doctoral degrees with expertise in quantitative analysis and research design, one of whom also brings specialized expertise in mixed-methods research.

The shared professional backgrounds in academic librarianship of the team's three librarian-researchers informed both the design and interpretation of this study. As practitioners with extensive experience providing and evaluating reference services, the librarian-researchers brought insider knowledge of chat reference environments, institutional policies, and common patron behaviors. This familiarity supported a nuanced interpretation of user intent and transaction characteristics, but it also introduced the potential for disciplinary assumptions and normalization of routine library interactions. To mitigate this risk, the team employed iterative codebook testing, formal intercoder reliability assessment, and collaborative discussion to surface and resolve divergent interpretations. The inclusion of a non-librarian team member provided an additional external perspective, prompting critical reflection on category boundaries, analytic assumptions, and interpretive framing.

Institutional context and data

This study draws on chat reference transcripts contributed by five large, public R1 university libraries in the United States. Universities that contributed data include the University of Arizona, Colorado State University, the University of Kansas, Utah State, and the University of Utah. Universities were selected based on their Carnegie classification and existing interest in collaborative assessment work. Institutional Review Board approval was obtained at Colorado State University, where the project was classified as exempt. Several contributing libraries also had IRB approvals in place at their own institutions. Collectively, these institutions supplied transcripts spanning August 15, 2021, through May 15, 2024. As shown in Table 1, participating libraries contributed a total of 34,977 transcripts, with contributions ranging from 392 (Utah State) to 17,533 (Arizona).

Table 1. Total transcripts per year by institution

Institution	Total transcripts per year 2021	2022	2023	2024	No date	Total Transcripts	% of total transcripts
Arizona	3,064	5,865	5,522	2,902	–	17,353	50%
Colorado State	1,776	1,387	1,700	1,175	41	6,079	17%
Kansas	1,586	3,379	3,062	1,141	–	9,168	26%
Utah	1,027	958	–	–	–	1,985	6%
Utah State	–	–	265	126	1	392	1%
Total	7,453	11,589	10,549	5,344	42	34,977	100%

Institutions supplied transcripts in XML, CSV, or XLSX formats, which were standardized into a common XLSX (Excel version 2601) structure and stripped of system-generated fields not relevant to the analysis. To protect patron privacy, each institution replaced local identifiers with anonymized institutional codes prior to transfer. All files were uploaded to an encrypted Amazon Web Services (AWS) S3 environment and processed with AWS Comprehend to identify and redact personally identifiable information (PII). Redacted files were then stored in a restricted OneDrive directory maintained by Colorado State University.

After redaction, all institutional files were merged into a single combined dataset. Each transcript was tagged with its source institution, assigned a unique study ID, and given a randomly generated number [Excel function = RAND()] to enable random sampling and ordering in subsequent analyses. The consolidated file was stored alongside the individual institutional files in the secure OneDrive directory and served as the dataset for all subsequent analyses.

Sampling strategy

Because the full dataset contained nearly 35,000 transcripts, manually coding all records was impractical. Instead, the team targeted a 5% sample (approximately 1,748 transcripts) to balance analytical rigor with coding capacity. To ensure representation across institutions, several stratified sampling approaches were evaluated (see Table 2). A proportional 5% stratified sample would have mirrored the population distribution but produced very small sample sizes for Utah and Utah State ($n = 99$ and $n = 20$, respectively). An equal-sized sample (349 per institution) would have addressed this imbalance but substantially overrepresented small contributors and underrepresented the largest (e.g. Arizona).

To mitigate the limitations of both approaches, a hybrid stratified sampling strategy was selected. Utah and Utah State were each assigned 50 transcripts, while the remaining sample

Table 2. Stratified sampling options

University	Total transcripts	Proportional stratified sample	Equal-sized sample	Hybrid stratified sample	Hybrid stratified sample % of total
Arizona	17,353	867	349	877	50.2%
Colorado State	6,079	304	349	307	17.6%
Kansas	9,168	458	349	462	26.5%
Utah	1,985	99	349	50	2.9%
Utah State	392	20	349	50	2.9%
Total	34,977	1,748	1,745	1,746	100%

was allocated proportionally across Arizona, Colorado State, and Kansas – the institutions contributing the largest number of transcripts. This approach yielded a final sample of 1,746 transcripts, approximately 5% of the full dataset. Within each institutional stratum, transcripts were selected randomly using the preassigned random numbers in the consolidated file, ensuring unbiased selection.

Codebook development

Codebook development proceeded through an iterative, multistage process that combined both deductive and inductive analytic strategies. All working versions of the codebook and related materials were maintained collaboratively by the research team in a shared spreadsheet environment. This structure supported version control, iterative refinement, and transparent documentation of analytic decisions.

The development process began with the creation of a set of deductive codes derived from a targeted literature review of empirical studies that examined academic library chat reference services. The research team extracted commonly reported categories, interaction types, and thematic patterns from prior scholarship and synthesized them into an initial list of deductive codes. These codes established the conceptual foundation for subsequent analysis.

To complement this deductive approach, the research team engaged in qualitative data familiarization and human inductive coding. First, an exploratory review of 100 randomly selected chat transcripts identified emergent patterns, recurrent user needs, and unanticipated interaction types. This open-coding process generated a preliminary set of inductive codes grounded in the behavioral and contextual characteristics of actual chat exchanges.

Next, ATLAS.ti's AI Summaries feature (https://atlasti.com/) and Insight7 (an AI-powered conversational intelligence platform - https://insights7.com/) were utilized on a small sample of transcripts. Both tools provided supplementary and confirmatory input rather than executing an independent layer of inductive coding. Although machine-generated, these exploratory summaries served purely as a data-familiarization mechanism to sense-check and refine our human-generated categories by providing concise overviews that ensured the framework captured high-level conceptual trends before full-scale deductive coding began.

The research team then engaged in a systematic consolidation process. Deductive codes, inductive codes from the transcript sample, and AI-generated summaries were compared for conceptual overlap, merged where appropriate, and refined through collaborative discussion. This process yielded Codebook Version 1, which included operational definitions and sample questions to support consistent application during subsequent coding phases.

Codebook testing and revision

Intercoder reliability (ICR) procedures were conducted in two iterative rounds to evaluate, refine, and confirm the clarity and stability of the coding scheme. Krippendorff's α (nominal) was selected as the primary reliability statistic because it provides a chance-corrected measure that accommodates multiple coders, nominal codes, and incomplete coder overlap, all of which categorize this study's design (Krippendorff, 2019, Ch. 12). Fleiss' *K* was examined as a supplemental indicator of macro and per-code agreement, while Cohen's *K* indicated pairwise agreement (between coder pairs). Commonly accepted thresholds (see Table 3) guided interpretation, with an $\alpha \geq 0.80$ considered reliable and 0.67–0.80 acceptable for tentative conclusions (Hayes and Krippendorff, 2007; Marzi *et al.*, 2024).

Round 1 involved four coders independently coding 25 randomly selected transcripts (five per institution). Krippendorff's α was 0.564 (weak), and macro-averaged Fleiss' *K* across all transcripts was 0.457 (intermediate to good) when constant (never-used) codes were excluded. Pairwise Cohen's *K* ranged from 0.480 to 0.674, indicating weak to moderate agreement between each coder pair. Coding agreement (Krippendorff's α), as outlined in Supplemental Table A revealed strong agreement in transactional categories such as Study Rooms and Reservations, Renewals, Hold Requests, Database Search Skills, and Finding Relevant

Table 3. Intercoder reliability thresholds

Statistic	Description	Interpretative thresholds Weak	Moderate	Strong
Krippendorff α	agreement across all Study IDs × Codes; provides a single measure of reliability	<0.67	0.67–0.80	≥0.80
Fleiss' *K*	agreement among coders *across the entire set of codes*, averaged over all individual per-code κ values	<0.40	0.40–0.75	>0.75
Cohen's *K*	overall agreement between each pair of coders; used to identify coder-to-coder alignment patterns	<0.59	0.60–0.79	≥0.80

Note(s): Codes that were never assigned to a transcript in each Round were excluded from the Fleiss' *K* calculation

Sources. Coders met and reviewed discrepancies, clarified conceptual boundaries, revised categories, and added five additional codes, resulting in Codebook Version 2.

Round 2 of codebook testing used a new set of 25 transcripts, which were coded by three members of the research team. Reliability improved, though Krippendorff's α (0.610) and macro-average Fleiss' *K* (0.347) remained weak. Of note, Krippendorff's α was strong (>0.80) in several transactional and known-item categories (see Supplemental Table A. Cohen's *K* pairwise was slightly improved in Round 2 (0.517–0.689; weak to moderate). At this point, the research team conducted a consensus reconciliation session. During this work, transcripts were reviewed collectively to resolve remaining differences in interpretations. This process ensured that the team reached 100% agreement on the application of the final 42-item codebook (see Supplemental Table B before proceeding to full coding. Consequently, the full sample ($n = 1{,}746$) was coded using a refined framework grounded in this shared consensus.

Coding process and cross-checking

Round 3 involved full coding of the 5% hybrid random sample of 1,746 transcripts, evenly distributed across the four coders (436 or 437 transcripts per coder). Coder distribution was proportional to institutional transcript volume to ensure balanced representation. Each transcript could be assigned up to three codes. Overall, 2,105 codes were assigned across the 1,746 transcripts in the sample.

High-frequency code cross-checking was conducted in two sequential steps to strengthen coding reliability and ensure consistent interpretation of key themes. Step 1 focused on the most frequently applied codes: each coder independently reviewed all transcripts associated with the top ten codes and flagged any instances they believed were miscoded. Two team members then worked collaboratively to review all flagged instances, resolve discrepancies, and produce a refined set of coded transcripts. Step 2 expanded cross-checking based on overall code frequency: following the initial review, all codes were rank-ordered by total frequency and grouped into quartiles (Q1–Q4), with Q4 representing the most frequently applied codes. To maintain consistent coding quality across high-impact themes, the research team determined that any additional codes in Q3 and Q4 that had not yet undergone cross-checking should also be reviewed. These reviews followed the same procedures used in Step 1.

Complexity classification

A separate analysis was conducted to assign levels of complexity to each transcript. To prepare the data for this analysis, a custom Visual Basic (VBA) macro, created by a member of the research team in Excel, was applied to the raw transcripts to improve readability. The macro attempted to remove embedded HTML artifacts and label speaker turns (Library vs. Guest). Although it did not fully resolve all inconsistencies, it substantially reduced visual noise and produced the standardized "cleaned transcript" field used in all subsequent calculations.

Using the cleaned transcripts, duration (end timestamp minus beginning timestamp, in minutes) and total words for each interaction were calculated in Excel. After removing records with unusable timestamps, 1,647 transcripts remained. Because duration and word-count distributions were strongly right-skewed, equal-frequency tertiles (33rd and 67th percentiles) were constructed rather than relying upon equal-width bins. Duration tertiles were defined as short (≤3.53 min), medium (>3.53–8.30 min), and long (>8.30 min); word-count tertiles followed the same structure (≤100; >100–187; >187 words). Cross-classified combinations were then mapped to Simple, Moderate, or Complex categories, with additional word-based rules applied to the 99 transcripts that lacked duration data (see Table 4). Complexity categories were subsequently applied to the full coded sample of 1,746 transcripts.

Results

This section presents the major empirical findings from the cross-institutional analysis of 1,746 coded chat transactions collected across five R1 universities. The results are organized around two focal areas: overall frequency patterns and complexity of chat interactions across institutions and transaction codes. Together, these findings provide a detailed descriptive account of the types of support users seek through academic library chat services, the level of professional expertise these interactions require, and the ways local service models shape chat activity. All frequency counts, quartile rankings, and complexity classifications referenced in this section are derived from the shared coding framework applied manually by the research team across all institutional datasets.

Table 4. Complexity criteria, designation and frequencies

Criteria	Duration	Words	Designation	Frequency % (n = 1,746)
If Duration = Long AND Words = Long	>8.30 min	>187 words	Complex	22.8% (398)
If Duration = Long AND Words = Medium	>8.30 min	>100 and ≤ 187 words	Complex	6.4% (111)
If Duration = Medium AND Words = Long	>3.53 min and ≤ 8.30 min	>187 words	Complex	8.2% (144)
If Duration is Not Available, AND Words = Long	Missing	>187 words	Complex	0.1% (2)
If Duration = Medium AND Words = Medium	>3.53 min and ≤ 8.30 min	>100 and ≤ 187 words	Moderate	16.1% (281)
If Duration = Short AND Words = Medium	≤ 3.53 min	>100 and ≤ 187 words	Moderate	8.7% (152)
If Duration = Short AND Words = Long	≤ 3.53 min	>187 words	Moderate	0.4% (7)
If Duration is Not Available AND Words = Medium	Missing	>100 and ≤ 187 words	Moderate	0.2% (3)
If Duration = Short AND Words = Short	≤ 3.53 min	≤ 100 words	Simple	22.5% (392)
If Duration = Long AND Words = Short	>8.30 min	≤ 100 words	Simple	2.2% (39)
If Duration = Medium AND Words = Short	>3.53 min and ≤ 8.30 min	≤ 100 words	Simple	7.0% (123)
If Duration is Not Available, but Words = Short	Missing	≤ 100 words	Simple	5.4% (94)

Overall frequency

Across the full dataset (2,105 codes assigned across 1,746 transcripts), a small subset of codes accounted for the majority of activity. The ten most frequently assigned codes, representing just 25% of the overall codebook, constituted 60% ($n = 1,269$) of all coded transactions. As shown in Table 5, these top-ranking, high-volume categories reflect a high degree of concentration in the types of support users most commonly seek through library chat services. A full listing of all codes, their frequencies, and quartile rankings, both overall and by institution, is provided in Supplemental Table C.

The presence of three codes in the top quartile (Q4) at all five institutions: *Find a Known Item: Journals/Articles*; *Finding Relevant Sources*; and I*nterlibrary Loan (ILL)*, is particularly noteworthy. Their consistent prominence across every site suggests a shared core service profile. Supporting known-item article retrieval, assisting with research source discovery, and facilitating ILL requests appear to be foundational functions of academic library chat services, regardless of local institutional context.

At the same time, the institutions also diverged in informative ways. Arizona accounted for the overwhelming majority of renewal transactions (92%, 180 of 195), suggesting that chat may serve as a primary channel for routine, and low-complexity access services at that institution. Similarly, Kansas accounted for a disproportionate share of ILL-related chats (60%; 84 of 139), suggesting deeper integration between chat and ILL workflows than observed elsewhere. Together, these patterns suggest that variation in chat activity across institutions may reflect how a library positions and markets its chat service (e.g. as a research tool versus a quick help desk), rather than underlying differences in user populations.

Code clusters

Each code in the codebook was assigned to one of seven overarching code clusters (Table 6). Across these clusters, the largest share of codes fell within *Access and Circulation* (24%, $n = 498$), followed by *Known Item Retrieval* (19%, $n = 399$). *Discovery and Research Support* ($n = 316$) and *Referral and Administrative* ($n = 326$) each accounted for 15% of the total codes assigned. *Physical and Operational* ($n = 188$), *Technical and Connectivity* ($n = 181$), and *Other* ($n = 197$) each represented 9% of the total coded instances. Overall, the distribution of

Table 5. Top 10 codes, frequencies and complexity

Code*	Overall rank	Count across all codes ($n = 2,105$)	% of all codes	Complexity tendency	Present at all 5?
Renewals	1st	195	9.3%	Simple	No —Q4 at 3 (concentrated at 1 institution)
Find a Known Item: Journals, Periodicals, or Articles	2nd	192	9.1%	Complex	Yes — Q4 at all 5
Finding Relevant Sources	3rd	165	7.8%	Complex	Yes — Q4 at all 5
Interlibrary Loan (ILL)	4th	139	6.6%	Complex	Yes — Q4 at all 5
Find a Known Item: Books	5th (tie)	137	6.5%	Complex	Yes—Q4 at 3
Connectivity/Remote Access	6th (tie)	108	5.1%	Complex	Yes – Q4 at 3
Borrow Technology	6th (tie)	108	5.1%	Simple	Yes –Q4 at 4
Patron Accounts	6th (tie)	108	5.1%	Complex	Yes – Q4 at 2
Library Services	9th	69	3.3%	Mixed	Yes – Q4 at 3
Policies and Procedures	10th	48	2.3%	Mixed	Yes – Q4 at 1
Total	–	1,269	60.2%	–	–

Note(s): *The codes "Abandoned chat" and "Other" are excluded from the ranked list

Table 6. Codes, clusters and frequencies

Coding cluster	Codes in cluster	Total codes assigned (n = 2,105)	% of total codes
Access and Circulation	Borrow Tech | Fees and Fines | Hold Requests | Patron Accounts | Renewals | Lost Items	498	24%
Discovery and Research Support	Citations/Citing | Database Search Skills | Develop Research Topic | Evaluating Info | Finding Relevant Sources | Manage/Organize Info | Research Strategies	316	15%
Known Item Retrieval	Audiovisual | Books | Journals/Periodicals/Articles | Theses and Dissertations | Other | Find Items by Author	399	19%
Physical and Operational	Campus Services | Campus Wayfinding | Library Hours | Library Navigation and Wayfinding | Study Room Reservations | Lost and Found | Noise | Physical Accessibility | Printing and Scanning | Safety and Security | Study Room Reservations	188	9%
Referral and Administrative	Interlibrary Loan | Course Reserves | Library Services | Policies and Procedures | Request a Purchase | Faculty Instructional Support	326	15%
Technical and Connectivity	Connectivity and Remote Access Issues | Software | Tech Support | Website	181	9%
Other	Abandoned Chat | Other | System Test	197	9%
Total		2,105	100%

codes across the clusters suggests that chat services function as a high-use access point for both everyday operational support and higher-complexity research assistance.

Complexity distribution

Using the study's three-level complexity classification framework based on duration and word count, 37% (n = 648) of all transcripts in the sample were classified as simple, 25% (n = 443) as moderate, and 38% (n = 655) as complex. Thus, over one-third of academic library chat interactions are likely to require substantive professional judgment that involves multi-step problem solving, advanced research support, or specialized knowledge. Conversely, just over one-third of interactions fell into the simple category, consisting primarily of routine, highly structured questions with clear and deterministic answers.

Complexity by institution. Table 7 summarizes the complexity distribution for each participating institution and reports the average interaction duration in minutes. A full listing of complexity ratings per code and institution is available in Supplemental Table D. While overall patterns are broadly consistent, substantial institutional variation is evident, as

Table 7. Complexity by institution

Institution	Count of transcripts	Simple	Moderate	Complex	Avg. duration (min)
University of Arizona	877	35%	25%	40%	8.6
Colorado St.	307	47%	22%	31%	7.9
University of Kansas	462	35%	29%	35%	7.3
University of Utah	50	52%	30%	18%	5.6
Utah State University	50	16%	10%	74%	16.8
All institutions combined	1,746	37%	25%	38%	*8.3*
Average Duration (min)	1,746	3.4	4.4	15.0	*8.3*

illustrated in Figure 1. Utah State University had the highest proportion of complex interactions (74%) and the longest average duration (16.8 min). The University of Utah reported the lowest proportion of complex interactions (18%) and a comparatively shorter average duration (5.6 min). The remaining institutions, Arizona, Colorado State, and Kansas, fell between these two endpoints, with complex interactions representing between 31% to 40% of total activity and averaging 8.1 min. When aggregated across all institutions, interactions averaged 8.3 min, with 3.4 min for simple chat interactions, 4.4 min for moderate, and 15.0 min for complex.

Complexity by transaction type. Analysis of transaction complexity revealed substantial variation across high-volume codes, particularly those in the upper frequency quartiles (Q3–Q4) presented in Table 8. Codes associated with research-intensive coding clusters, most notably *Discovery and Research Support* and *Known Item Retrieval*, exhibited a predominantly complex profile. In contrast, codes aligned with *Access and Circulation* and *Physical and Operational* clusters more frequently demonstrated simple or mixed complexity patterns. Overall, the findings indicate a clear distinction between research-focused transactions, which tend to cluster in the complex tier, and logistical or access-related questions, which more often fall within the simple tier. A full listing of complexity distributions across and within institutions is provided in Supplemental Table D.

Across all individual codes and frequency quartiles, *Develop Your Research Topic* exhibited the highest proportion of complex interactions (77%), followed by *Research Strategies* (71%), *Database Search Skills* (67%), *Find Items by Author* (63%), and *Connectivity and Remote Access Issues* (60%), as detailed in Supplemental Table D. Several of these codes correspond to transaction types classified as "complex" in Table 8 and are also concentrated in the higher-volume quartiles (Q3–Q4), highlighting a sustained demand for high-touch research support through chat. However, elevated complexity in these codes was observed even when they appeared less frequently, indicating that complexity is inherent to the nature of the transaction rather than solely a function of volume.

By contrast, individual codes associated with operational and access-related needs demonstrated a stronger tendency toward simpler interactions across quartiles. *Library Hours* had the highest proportion of simple transactions (64%), followed by *Software* (62%), *Printing and Scanning* (58%), *Campus Services* (49%), *Tech Support* (45%), and *Renewals* (45%). *Borrow Technology* (44%) and *Campus Wayfinding* (44%) likewise showed elevated rates of simple interactions, reflecting their largely routine and procedural character rather than concentration in any single frequency quartile.

Results summary

Overall, the cross-institutional analysis of 1,746 human-coded transcripts revealed a concentrated set of user needs dominated by research-focused inquiries that consistently

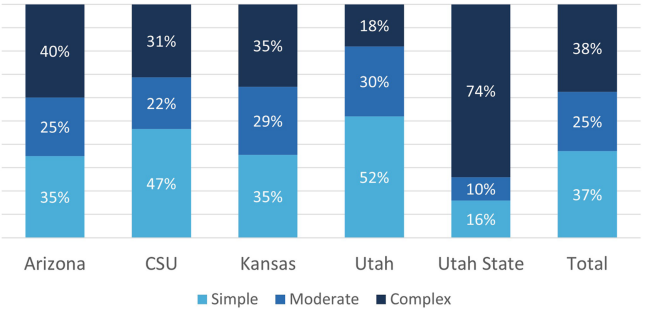

Figure 1. Complexity ratings across institutions. Property of the authors

Table 8. High volume codes (Q4 & Q3) and overall complexity propensity

Code	Code cluster	*N* (2,105 codes)	Quartile	Complexity propensity		
				Simple	Moderate	Complex
Renewals	Access and Circulation	195	Q4	X		
Find a Known Item: Journals/Periodicals/ Articles	Discovery and Research Support	192	Q4			X
Finding relevant sources	Discovery and Research Support	165	Q4			X
Interlibrary Loan	Referral and Administrative	139	Q4			X
Find a Known Item: Books	Discovery and Research Support	137	Q4			X
Connectivity and Remote Access Issues	Technical and Connectivity	108	Q4			X
Borrow Technology	Access and Circulation	108	Q4	X		
Patron Accounts	Access and Circulation	108	Q4			X
Other	Other	74	Q4			X
Library Services	Referral and Administrative	69	Q3	X (mixed)		X (mixed)
Policies and Procedures	Referral and Administrative	48	Q3	X (mixed)		X (mixed)
Fees and Fines	Access and Circulation	47	Q3			X
Database Search Skills	Discovery and Research Support	45	Q3			X
Study Rooms and Reservations	Physical and Operational	43	Q3			X
Tech Support	Technical and Connectivity	42	Q3	X		
Hold Request	Access and Circulation	39	Q3			X
Library Hours	Physical and Operational	36	Q3	X		
Campus Services	Physical and Operational	35	Q3	X		
Find a Known Item: Other	Known Item Retrieval	35	Q3			X

required higher-level expertise across five universities, alongside routine operational questions that varied by local service models. Collectively, patterns in the findings suggest that academic library chat services function both as a reliable source of complex research assistance as well as a mechanism for handling institution-specific logistical and operational support.

Discussion

Cross-institutional patterns

The three-level complexity rating system developed for this study (simple, moderate, and complex) is not solely descriptive but also serves as a framework for planning chat reference staffing and routing. Results indicate that 38% of interactions were complex, 25% were

moderate, and 37% were simple. Simple interactions are typically routine and clearly defined, with predictable responses that do not require professional judgment. Categories such as Library Hours, Printing and Scanning, Software, and Renewals frequently appear in these interactions and are highly represented in the data set. These types of queries are well-suited for simple routing or AI-assisted support. Due to their brevity and straightforward nature, technology tools can effectively address these needs quickly and consistently, while providing 24/7 access to support. Automating these brief, high-volume, low-stakes interactions is unlikely to diminish service quality and will allow human expertise to be redirected towards the 38% of transactions identified as complex.

The moderate complexity rating presents the greatest challenge for determining appropriate handling strategies. These interactions are not sufficiently straightforward for automation, nor can they be easily routed to a librarian without further assessment. Addressing them requires more sophisticated decision-making processes. A hybrid approach may be most effective, beginning with an AI-assisted response and providing a clear mechanism for escalation to a human when necessary. Failure to address these interactions appropriately could negatively impact service quality.

Three categories in the top quartile at all five institutions, Find a Known Item: Journals/Articles, Finding Relevant Sources, and Interlibrary Loan, are classified as complex. This finding underscores the nuanced nature of library reference work, which often requires context-specific responses. Given the high degree of context dependence, any tool developed to address these interactions must be trained on data that accurately reflects this complexity; otherwise, there is a significant risk of incorrect routing or responses. This study provides an initial empirical foundation for developing such a tool.

Institutional differences introduce further complexity in developing new approaches to chat reference. The variation in complexity profiles, ranging from Utah State University's 74% complex interaction rate to the University of Utah's 18%, demonstrates significant diversity, though sample size and service context necessitate caution in interpreting direct comparisons. Some context is helpful for understanding the institutional differences observed in this study, though the descriptive design does not allow causal claims. The variation in sample size is real: Arizona accounts for 877 coded transcripts; hybrid stratified sampling limited Utah and Utah State to 50 transcripts each. This means that the complexity distributions for institutions with smaller sample sizes are more sensitive to individual transcript characteristics and should be interpreted with caution. Additionally, the time periods spanned by the transcripts vary across institutions. The University of Utah's transcripts are from 2021–2022, and Utah State University's transcripts are from 2023–2024. The three other institutions' transcripts are taken from all four years (2021–2024). In addition, each library can be assumed to differ in the way it organizes and provides services. Examples of possible variation include which staff members answer reference questions, whether they use a vendor for staffing chat reference, and the hours during which they provide these services. User populations also vary across institutions in the distribution of undergraduate students, graduate students, and faculty, as well as in their academic disciplines. Taken together, these considerations suggest that any approach to automating or restructuring chat reference services must be adapted to local institutional context rather than derived solely from cross-institutional patterns.

Data use for service design and diagnostic tool

One of the most interesting and useful unexpected uses for this data has been the opportunity to use it for service planning and as a service diagnostic tool. The patterns that surface during this type of analysis show problems upstream in the service ecosystem. Signals such as high complexity in ILL transactions, an increase in connectivity issues, and a concentration of user account questions can show what is not working elsewhere. This makes these problems visible and actionable.

The complexity profile of a university library may say more about its service model than its users. Despite similar models and missions, complexity ratings varied significantly across the universities, ranging from 18% to 74%. Before implementing automation in virtual reference, it may be helpful to ask what the data suggest about the information we have available elsewhere, and how we communicate about our reference services.

Complexity is predictable, and that makes it actionable. Transaction complexity in library chat isn't random. It clusters by type in consistent ways across five institutions. The predictability of these clusters may help guide service design. With this information, it is possible to make evidence-based decisions about where automation fits, where it does not, and when a handoff is needed.

Limitations

Data distribution and institutional bias

The dataset is not uniform across participating institutions, which limits the confidence of certain comparative conclusions. Specifically, the University of Arizona contributed 50% of the total transcripts; consequently, aggregate findings are heavily weighted toward that institution's specific data. In contrast, only 50 transcripts each were included in the sample from the University of Utah and Utah State University, per the hybrid stratified design. These smaller sample sizes result in less stable complexity ratios that are highly sensitive to individual transcript variations. Because of these discrepancies, direct comparisons between institutions should be interpreted with caution.

Time period variance and technological limitations

The timeframe of data collection also varies by institution. While most institutions provided transcripts spanning a four-year period, the University of Utah data covers 2021–2022, and Utah State University data covers 2023–2024. Although localized shifts in library service patterns are typically minor over such a short duration, this programmatic inconsistency introduces a minor variable into the comparative findings.

A more significant temporal limitation is that the study's overarching data collection window (2021–2024) directly spans a period of unprecedented, disruptive technological change in higher education. Transcripts from the earliest phase of this study (2021–2022) precede the public release of ChatGPT, capturing an environment of virtually zero student adoption of generative AI. While student adoption of AI tools rose rapidly through 2023 and into 2024, the full dataset primarily captures a pre-AI baseline and the initial, turbulent emergence of student and faculty adoption of these tools.

Consequently, while these findings provide a robust, highly detailed analysis of user behaviors and reference needs during a crucial transitional era, this dataset reflects a specific historical snapshot. Attempting to use these data points to predict contemporary library chat volume or current inquiry behavior may fall short, as academic libraries now operate in an ecosystem where generative AI has stabilized as an invisible infrastructure for student and faculty research.

Statistical limitations

Institutional comparisons should be interpreted with caution. The complexity rating distributions provided are strictly descriptive. Because the variance between institutions was not tested for statistical significance, observed differences may reflect sampling limitations rather than meaningful distinctions between universities. Consequently, these trends may not be generalizable beyond this specific sample. Complexity was operationalized using interaction duration and word count, practical and replicable proxies, but indirect measures of professional demand.

Coding methodology and depth

Assigning a single qualitative code to multi-topic chat transcripts is a standard practice in the field. It facilitates quick, high-level analysis and reduces the labor involved in coding. However, it can result in a loss of granularity and context and obscure the diversity of user needs. To mitigate this risk, our methodology allowed for up to three codes per transcript. However, the cognitive workload and analytical fatigue associated with this analysis may have limited our application of secondary and tertiary codes. As a result, these findings capture predominant service patterns but may underrepresent the latent complexity of the interactions. Future research should utilize high-granularity analysis to capture intent density more fully.

Environmental and demographic variables

While this analysis provides a broad overview of service interactions, it excludes several key contextual variables. First, anonymization prevents the inclusion of user demographics such as academic programs, level of study, or baseline information literacy. Second, the dataset reflects a heterogeneous environment with diverse staffing (ranging from staff to faculty librarians to external service partners), with each following different training protocols. Finally, the data does not account for time of year or other environmental factors (e.g. peak exam periods or localized technical outages), which can skew chat volume and complexity. These unobserved variables represent opportunities for future longitudinal research into the correlation between user identity and service outcomes.

Future direction for research

The constraints of manual qualitative coding, including the necessity of prioritizing a primary intent for categorical efficiency, highlight a significant methodological opportunity. While the 5% sample analyzed in this study provides a noteworthy window into institutional trends, human-only baselines may inadvertently oversimplify the complex, multi-layered activity inherent in library chat services. To address this, the validated codebook and human-coded transcripts from this study are currently serving as the foundation for a methodological expansion study.

This ongoing research utilizes the existing dataset to train a deterministic AI structured reasoning model designed to audit and code the full corpus of 34,977 transcripts. Unlike traditional pattern-recognition algorithms, this model employs explicit logical frameworks derived directly from the validated codebook. To ensure maximum data integrity, the next phase includes a human-AI collaborative audit of the initial 5% sample; the model will flag coding conflicts for expert human review, refining the model before scaling the analysis across the remaining 95% of the dataset.

By transitioning from primary-intent categorization to exhaustive intent mapping, this computational approach allows for a more granular discovery of professional labor and technical support that is often under-reported in manual assessment models. Ultimately, this expansion will provide both a large-scale validation of the current findings and a comprehensive map of user intents across the entire multi-institutional dataset.

Beyond the methodological extension, this study suggests several additional directions for future research. Because this dataset captures a pre-AI baseline of student research behaviors, search patterns, and struggles, it serves as an essential diagnostic tool to measure how library reliance has begun to pivot. Longitudinal analysis of chat transaction patterns would allow for a direct comparative analysis against contemporary transcripts, mapping exactly how query complexity has evolved in an ecosystem where generative AI has become invisible infrastructure. Such research will be vital to identifying whether advanced AI technologies are successfully absorbing routine inquiries and evaluating how institutional reference demands are shifting toward increasingly complex, high-level research consultations. Applying the complexity framework and codebook to different types of institutions — including smaller colleges, community colleges, and international academic libraries — would test the generalizability of these findings and expand

their usefulness for cross-institutional service design. There are also opportunities to examine relationships between user demographics and complexity profiles; provided data can be obtained in ways that preserve patron anonymity. Finally, incorporating variables such as academic calendar timing, staffing models, and service marketing practices could help explain the institutional variation this study documented but could not fully account for.

Conclusion

This study addresses a gap in the literature related to virtual reference services in academic libraries, providing a large-scale, multi-institutional, human-coded analysis of chat reference. Using a collaboratively developed 42-code taxonomy, two rounds of intercoder reliability testing, and a replicable complexity framework, we analyzed chat transcripts from five large public research universities. The study was designed to establish an empirical baseline and a replicable structure to support the development of high-quality automation for routine virtual reference queries. It also provides a codebook and structure for libraries to analyze their own data, which can help them align an automated system to their unique service model and the needs of their users. The study also provided unanticipated but practically significant contributions, including the capacity to diagnose upstream service problems and inform local service planning decisions.

The findings showed a consistent core of chat codes across five different academic libraries. Three transaction types, Find a Known Item: Journals/Articles, Finding Relevant Sources, and Interlibrary Loan, were all in the top quartile of all the institutions, which confirms that there is commonality among libraries and users of chat services. The finding that 38% of interactions are complex indicates that many questions require higher-level human assistance. The 37% of simple transactions suggests that a significant number could be automated in some fashion. In addition, the outcomes point to patterns in complexity that may be associated with local chat positioning and marketing practices, rather than substantive differences in user populations.

The taxonomy is designed to be replicable. The 42-item codebook is available (see Supplemental Table B and can be applied to any other set of chat reference data. The complexity framework adds substantial analytical value (see Tables 4 and 5 and Supplemental Table D. This analysis can help diagnose problems in a library's service ecosystem, including connectivity issues, ILL delays, and difficult-to-use user accounts. The complexity framework shows which types of interactions can be automated with the best outcome for users, which need to be routed to professionals and where a handoff can occur.

The codebook and human-coded transactions are already serving as the foundation for continued analyses, with AI-assisted coding currently being applied to the full 34,977-transcript dataset. This ongoing work will provide a large-scale validation of these findings and a more granular map of user intent across the full corpus. Crucially, because this study captures the foundational, pre-AI baseline of student research behaviors, it provides the essential diagnostic framework required for future longitudinal research to map how query complexity pivots as generative AI becomes a ubiquitous background utility for student and faculty research. Applying this validated framework to contemporary transcripts and diverse institutional contexts (including smaller universities and international settings) will allow library leaders to track these ongoing structural shifts. Ultimately, this study offers library leaders an empirical foundation for decisions that have too often lacked one, presenting a cross-institutional baseline of chat reference users' needs and a replicable framework for translating that evidence into future service design.

AI statement

During the preparation of this work, the authors used AI tools to copy edit and improve the language, readability, and clarity of portions of the manuscript, including sections of the literature review. Specifically, Gemini 2.0 Flash (Google), Microsoft 365 Copilot based on the GPT-4o chat model (Microsoft), and Claude (Anthropic) were used for this purpose. Additionally, the exploratory AI

summary features of Insight7 and ATLAS.ti were utilized on a small, preliminary subset of transcripts solely as a computational data-familiarization tool to assist the research team during the early stages of codebook development. After using these tools, the authors reviewed and edited all content and take full responsibility for the accuracy, integrity, and content of the publication. Generative AI was not used to create, draft, or generate new research content, findings, or analysis.

Supplementary material

The supplementary material for this article can be found online: https://hdl.handle.net/10217/244256

References

Arnold, J. and Kaske, N.K. (2005), "Evaluating the quality of a chat service", *Portal: Libraries and the Academy*, Vol. 5 No. 2, pp. 177-193, doi: 10.1353/pla.2005.0017.

Association of College and Research Libraries (2024), *2024 ACRL Academic Library Trends and Statistics Survey*, American Library Association, available at: https://acrl.libguides.com/stats/surveyhelp.

Broughton, K.M. (2003), "Usage and user analysis of a real-time digital reference service", *The Reference Librarian*, Vol. 38 Nos 79/80, pp. 183-200, doi: 10.1300/j120v38n79_12.

De Leon, S. and Viray, M.R. (2022), "Examining the online reference service user experience of an academic library through qualitative content analysis of chat transcripts in the COVID-19 era", *Paper presented at the 2022 Conference on Library and Information Studies*, Manila, Philippines, available at: https://animorepository.dlsu.edu.ph/conf_clis/2022/Schedule/8/.

Dempsey, P.R., Insua, G.M., Armstrong, A.R., Hudson, H.J., Caragher, K. and McGregor, M. (2024), "Challenges and opportunities in teaching and learning in AskALibrarian chat: differences across subject domains", *Reference Services Review*, Vol. 52 No. 3, pp. 420-449, doi: 10.1108/rsr-05-2024-0024.

Devlin, F., Currie, L. and Stratton, J. (2008), "Successful approaches to teaching through chat", *New Library World*, Vol. 109 Nos 5/6, pp. 223-234, doi: 10.1108/03074800810873579.

Everdeen Moore, K., Ray, T.A. and Rogers, E. (2026), "Put your best chat forward: analyzing chat reference for best practices", *Georgia Library Quarterly*, Vol. 63 No. 1, p. 8, doi: 10.62915/2157-0396.2872.

Foley, M. (2002), "Instant messaging reference in an academic library: a case study", *College and Research Libraries*, Vol. 63 No. 1, pp. 36-45, doi: 10.5860/crl.63.1.36.

Guy, J., Pival, P.R., Lewis, C.J. and Groome, K. (2023), "Reference chatbots in Canadian academic libraries", *Information Technology and Libraries*, Vol. 42 No. 4, doi: 10.5860/ital.v42i4.16511.

Hayes, A.F. and Krippendorff, K. (2007), "Answering the call for a standard reliability measure for coding data", *Communication Methods and Measures*, Vol. 1 No. 1, pp. 77-89, doi: 10.1080/19312450709336664.

Hervieux, S. (2021), "Is the library open? How the pandemic has changed the provision of virtual reference services", *Reference Services Review*, Vol. 49 Nos 3/4, pp. 267-280, doi: 10.1108/rsr-04-2021-0014.

Hervieux, S. and Tummon, N. (2018), "Let's chat: the art of virtual reference instruction", *Reference Services Review*, Vol. 46 No. 4, pp. 529-542, doi: 10.1108/rsr-07-2018-0060.

Houlson, V., McCready, K. and Pfahl, C.S. (2007), "A window into our patron's needs: analyzing data from chat transcripts", *Internet Reference Services Quarterly*, Vol. 11 No. 4, pp. 19-39, doi: 10.1300/J136v11n04_02.

Jacoby, J., Ward, D., Avery, S. and Marcyk, E. (2016), "The value of chat reference services: a pilot study", *Portal: Libraries and the Academy*, Vol. 16 No. 1, pp. 109-129, doi: 10.1353/pla.2016.0013.

Kathuria, S. (2021), "Library support in times of crisis: an analysis of chat transcripts during COVID", *Internet Reference Services Quarterly*, Vol. 25 No. 3, pp. 107-119, doi: 10.1080/10875301.2021.1960669.

Kavitha, C. and Kavitha, K.P. (2023), "A chatbot system for education NLP using deep learning", *2023 Eighth International Conference on Science Technology Engineering and Mathematics (ICONSTEM)*, IEEE, pp. 1-7.

Kibbee, J., Ward, D. and Ma, W. (2002), "Virtual service, real data: results of a pilot study", *Reference Services Review*, Vol. 30 No. 1, pp. 25-36, doi: 10.1108/00907320210416519.

Krippendorff, K. (2019), *Content Analysis: An Introduction to its Methodology*, 4th ed., SAGE Publications, Los Angeles, CA, doi: 10.4135/9781071878781.

Lankes, R.D., Janes, J., Smith, L.C. and Finneran, C.M. (2003), *The Virtual Reference Desk: Creating a Reference Future*, Neal-Schuman, New York, NY.

Logan, J., Barrett, K. and Pagotto, S. (2019), "Dissatisfaction in chat reference users: a transcript analysis study", *College and Research Libraries*, Vol. 80 No. 7, pp. 925-944, doi: 10.5860/crl.80.7.925.

Luo, J. and Brissett, A. (2025), "Decoding virtual chats: NLP insights into academic library services", *Library and Information Science Research*, Vol. 47 No. 1, 101344, doi: 10.1016/j.lisr.2025.101344.

Maloney, K. and Kemp, J.H. (2015), "Changes in reference question complexity following the implementation of a proactive chat system: implications for practice", *College and Research Libraries*, Vol. 76 No. 7, pp. 959-974, doi: 10.5860/crl.76.7.959.

Marsteller, M.R. and Mizzy, D. (2003), "Exploring the synchronous digital reference interaction for query types, question negotiation, and patron response", *Internet Reference Services Quarterly*, Vol. 8 Nos 1/2, pp. 149-165.

Marzi, G., Balzano, M. and Marchiori, D. (2024), "K-Alpha calculator–Krippendorff's alpha calculator: a user-friendly tool for computing Krippendorff's alpha inter-rater reliability coefficient", *MethodsX*, Vol. 12, 102545, doi: 10.1016/j.mex.2023.102545.

Matteson, M.L., Salamon, J. and Brewster, L. (2011), "A systematic review of research on live chat service", *Reference and User Services Quarterly*, Vol. 51 No. 2, pp. 172-190, doi: 10.5860/rusq.51n2.172.

Mavodza, J. (2019), "Interpreting library chat reference service transactions", *The Reference Librarian*, Vol. 60 No. 2, pp. 122-133, doi: 10.1080/02763877.2019.1572571.

Mawhinney, T. and Hervieux, S. (2022), "Dissonance between perceptions and use of virtual reference methods", *College and Research Libraries*, Vol. 83 No. 3, pp. 503-525, doi: 10.5860/crl.83.3.503.

Maximiek, S., Rushton, E. and Brown, E. (2010), "Coding into the great unknown: analyzing instant messaging session transcripts to identify user behaviors and measure quality of service", *College and Research Libraries*, Vol. 71 No. 4, pp. 361-374, doi: 10.5860/crl-48r1.

McClure, C.R., Lankes, R.D., Gross, M. and Choltco-Devlin, B. (2003), *Statistics, Measures, and Quality Standards for Assessing Digital Reference Library Services: Guidelines and Procedures*, Syracuse University, Syracuse, NY.

Meert-Williston, D. and Sandieson, R. (2019), "Online chat reference: question type and the implication for staffing in a large academic library", *The Reference Librarian*, Vol. 60 No. 1, pp. 51-61, doi: 10.1080/02763877.2018.1515688.

Oakleaf, M. and VanScoy, A. (2010), "Instructional strategies for digital reference: methods to facilitate student learning", *Reference and User Services Quarterly*, Vol. 49 No. 4, pp. 380-390.

Radford, M.L. and Connaway, L.S. (2013), "Not dead yet! a longitudinal study of query type and ready reference accuracy in live chat and IM reference", *Library and Information Science Research*, Vol. 35 No. 1, pp. 2-13, doi: 10.1016/j.lisr.2012.08.001.

Rourke, L. and Lupien, P. (2010), "Learning from chatting: how our virtual reference questions are giving us answers", *Evidence Based Library and Information Practice*, Vol. 5 No. 2, pp. 63-74, doi: 10.18438/B87K7F.

Sears, J. (2001), "Chat reference service: an analysis of one semester's data", *Issues in Science and Technology Librarianship*, No. 32, doi: 10.29173/istl1868.

Valentine, G. and Moss, B.D. (2017), "Assessing reference service quality: a chat transcript analysis", in Mueller, D.M. (Ed.), *At the Helm: Leading Transformation: Proceedings of the ACRL 2017 Conference*, Baltimore, Maryland, March 22-25, 2017, Association of College and Research Libraries, Chicago, IL, pp. 67-75.

Wan, G., Beals, N., Siders, A. and Siders, G. (2009), "Academic libraries' chat reference: the role of answering questions", *The Journal of Academic Librarianship*, Vol. 35 No. 2, pp. 119-127.

Youngbar, A. (2012), "Questions by keystroke: an analysis of chat transcripts at Albert S. Cook Library", *Library Student Journal*, Vol. 7.

Zhuo, F., Love, M., Norwood, S. and Massia, K. (2006), "Applying RUSA guidelines in the analysis of chat reference transcripts", *College and Undergraduate Libraries*, Vol. 13 No. 1, pp. 75-88, doi: 10.1300/J106v13n01_09.

Corresponding author
Margaret Sinclair Brown-Sica can be contacted at: meg.brown-sica@colostate.edu