NATIONAL LEADERSHIP GRANTS FOR LIBRARIES

Sample Application LG-259841-OLS-26
Project Category: Planning

Virginia Tech University Libraries

Amount awarded by IMLS:	$178,994
Amount of cost share:	$0

Virginia Tech University Libraries will plan an investigation of the auditability of Large Language Models for relevance screening in evidence synthesis. This project will run iterative experiments to improve the accuracy and auditability of the artificial intelligence-assisted screening workflow. The project will incorporate the feedback from evidence synthesis library professionals across the nation and will produce feasibility findings, prototype workflows, and implementation guidance that libraries can use to assess how artificial intelligence-assisted screening can be incorporated into evidence synthesis services while preserving established methodological standards. Deliverables include an artificial intelligence-enabled evidence synthesis prototype; an instruction manual; and a report.

Attached are the following components excerpted from the original application.

- Narrative
- Schedule of completion

When preparing an application for the next deadline, be sure to follow the instructions in the most recent Notice of Funding Opportunity for the grant program to which you are applying.

Auditable AI Workflows for Evidence Synthesis Library Services

Bipasha Banerjee, C. Cozette Comer & William A. Ingram

INTRODUCTION

Virginia Tech University Libraries seeks $200,000 from the Institute of Museum and Library Services for a two-year planning project titled *Auditable AI Workflows for Evidence Synthesis Library Services*. The project addresses a growing challenge for research libraries that support evidence synthesis methods such as systematic reviews and meta-analyses, which require researchers to screen large bodies of scholarly literature through labor-intensive and time-consuming processes. Artificial intelligence (AI) tools are increasingly proposed as a way to assist in the screening of large-scale literature, but their integration into evidence synthesis workflows has not been systematically evaluated for transparency or reproducibility, which are defining characteristics of these research methods. This gap, recently identified by the international Responsible Use of AI in Evidence Synthesis (RAISE (Thomas et al., 2024a)) standards, must be addressed before widespread adoption is feasible. This planning project will examine the auditability of AI-assisted relevance screening workflows in library-supported evidence synthesis services by performing methodological analysis, prototype workflow development, and partnership-based evaluation with evidence synthesis experts. The project will produce feasibility findings, prototype workflows, and implementation guidance that libraries can use to assess how AI-assisted screening can be incorporated into evidence synthesis services while preserving established methodological standards. In doing so, the project responds directly to IMLS's call for libraries to creatively and effectively use artificial intelligence to improve academic research practices.

PROJECT JUSTIFICATION

Needs, Challenges, and Opportunities

Evidence synthesis represents a suite of research methods used in the knowledge-to-action life-cycle, often informing policy, practice, and decision-making. As these methods are used to inform real world decision-making and ultimately impact the prosperity of the wider community, those conducting an evidence synthesis review are expected to follow a predefined, transparent, and reproducible protocol. Research teams must identify and screen all potentially relevant evidence sources using explicit criteria, and then systematically extract and synthesize findings to answer a clearly stated research question. As a result of these requirements, evidence synthesis projects often take months or years to complete (Borah et al., 2017; Demetres et al., 2023).

Despite the time and resources required, researchers across disciplines have employed and advanced evidence synthesis methods such as systematic reviews, meta-analyses, and mapping reviews, to rigorously address a plethora of problems that impact policy, practice, and decision-making across fields. Reflecting the cross- and interdisciplinary nature of this work, international collaborations dedicated to the production of high-quality evidence synthesis have been established in social sciences (Campbell Collaboration), environmental management (Collaboration for Environmental Evidence), and health and medicine (Cochrane Collaboration).

Since the launch of AI chatbots like ChatGPT, the evidence synthesis community has explored opportunities to take advantage of this comparatively nuanced technology to improve timeliness and efficiency. However, at its core, evidence synthesis methods are valuable *because* the method is comprehensive, transparent, reproducible, and involves several steps toward reducing the risk of systematic errors (Royal Society & The Academy of Medical Sciences., 2018; Higgins et al., 2022; Kolaski et al., 2024; Daniel et al., 2026).

Community guidance reflects this concern. The recent recommendations for Responsible Use of AI in Evidence Synthesis (RAISE (Thomas et al., 2024a)), which have been adopted by all three major evidence synthesis collaborations (Flemyng et al., 2025), emphasize the need for transparency and methodological accountability when AI tools are incorporated into review workflows. In particular, the RAISE guidance

identifies relevance screening as the stage where AI has received the most attention, while also noting that many existing tools lack publicly available source code and provide limited information about how classifiers are trained (Thomas et al., 2024b). These limitations complicate efforts to evaluate the reliability and reproducibility of AI-assisted screening decisions.

A 2025 study of trends in AI for evidence synthesis found the most promising stage for responsible AI integration is relevance screening (Lieberum et al., 2025). In an AI-supported screening workflow, there are ways to control uncertainty or variability through methods such as controlled prompting, setting optimal hyperparameters, and fine-tuning to help the model learn from specific data. This challenge arises in part because generative AI systems produce probabilistic, non-deterministic outputs. Manual screening processes rely on explicit reviewer judgments and structured comparison of decisions to maintain inter-rater reliability. By contrast, unexamined use of generative models can introduce variability that is difficult to audit or reproduce. Although techniques such as controlled prompting, parameter tuning, and domain adaptation can reduce variability, these approaches have not yet been systematically evaluated within the context of evidence synthesis workflows.

More concerningly, it is possible that many synthesists and evidence synthesis librarians are using some form of AI without fully knowing or transparently reporting. Indeed, between 2017 and 2024 only 5.06% of evidence synthesis studies published through the Cochrane Collaboration, Campbell Collaboration, and the Collaboration for Environmental Evidence explicitly reported the integration of machine learning during the relevance screening stage (Scotti et al., 2025).

Taken together, these developments highlight a clear opportunity. While AI-assisted screening methods show promise for improving efficiency, the field lacks systematic evaluation of the transparency, stability, and auditability of these workflows. Addressing these issues is necessary before AI-assisted screening can be responsibly integrated into evidence synthesis practice.

Relationship to Existing Models, Standards, Scholarship, and Practice

Evidence synthesis has played a fundamental role in scientific progress, dating back to the mid-1700s with James Lind's *Treatise on Scurvy* (Lind, 1753). In recent decades, formal methodological standards have emerged across disciplines through collaborations such as the Cochrane Collaboration for health and medical research (Cochrane, 1993), the Campbell Collaboration for policy-oriented social sciences in 2000 (Campbell Collaboration, 1999), and the Collaboration for Environmental Evidence in 2007 (CEE, 2007). Throughout this period, libraries have served as essential partners in evidence synthesis by providing access to research literature and expertise in systematic searching, reference management, and methodological practice.

Recent advances in systematic review software have introduced automation into parts of the evidence synthesis workflow. Platforms such as Covidence (Walton, 2023) and DistillerSR (Wright, 2020) incorporate machine learning–assisted screening tools intended to reduce the time required for relevance screening. Other emerging tools, including PICO Portal, Nested Knowledge, Laser.ai, and Elicit, similarly promote AI-assisted workflows for literature discovery and review. However, many of these systems rely on proprietary models whose internal parameters and training data are not accessible to researchers. This limits the transparency and auditability of screening decisions, which are essential characteristics of rigorous evidence synthesis.

Recent research has explored the use of large language models to assist with relevance screening during systematic reviews. Several studies demonstrate that LLM-based classifiers can achieve performance comparable to traditional machine learning approaches while reducing the need for labeled training data (Li et al., 2026; Tsai et al., 2024; Li et al., 2024; López-Pineda et al., 2025). These studies establish the feasibility of using LLMs for title and abstract screening tasks. However, most work to date has focused on improving classification accuracy rather than examining the transparency, reproducibility, and stability of AI-assisted screening decisions.

The present proposal builds on this emerging body of work by investigating the reliability and auditability of AI-assisted screening workflows. Rather than focusing solely on predictive performance, this project evaluates how model choice, prompting strategies, and stochastic variability influence screening outcomes. By combining controlled experimentation with practitioner feedback from evidence synthesis librarians, the project examines the conditions under which AI-assisted screening methods can be considered transparent, reproducible, and suitable for adoption within evidence synthesis practice.

The Benefit to Libraries

This work directly supports the IMLS NLG-L program Objectives 1.2 by improving how libraries engage with and serve research communities through emerging analytical methods. Furthermore, this project benefits the broader community by improving the timeliness and quality of the available evidence base for policy and decision-making.

Because the foundation for a successful synthesis is a comprehensive corpus of all available evidence, librarians are uniquely qualified to support and guide synthesis studies. Co-authoring with librarians has been associated with improvements in quality, reliability, transparency, and comprehensiveness (Aamodt et al., 2019; Pawliuk et al., 2024). As such, demand for evidence synthesis support from librarians has only increased across disciplines (Lê et al., 2024).

While traditional librarian roles include the development and execution of a comprehensive search (Spencer and Eldredge, 2018), there are opportunities to further improve rigor and efficiency through librarian expertise. For example, accelerating the selection of relevant material through machine learning (Cawley and Cawley, 2022).

PROJECT WORK PLAN

The project is divided into two phases, with each phase taking approximately one year to complete. Table 1 displays the tasks that will be completed within each phase of the project to ensure successful project completion.

Planning Objectives

This planning project will investigate how AI-assisted relevance screening workflows might be incorporated into library evidence synthesis services while preserving the methodological principles of transparency, auditability, reliability, and validity. The project will pursue two primary planning objectives:

1. **Workflow feasibility:** How can an AI-assisted screening workflow be developed for responsible and strategic integration into library evidence synthesis services while preserving the principles of transparency, auditability, reliability, and validity?
2. **Professional evaluation:** How do library evidence synthesis professionals evaluate whether AI-assisted relevance screening workflows can be documented and communicated in ways that support reproducible, reliable evidence synthesis results?

To improve the timeliness of evidence synthesis production without deteriorating the inherent value of these methods, we envision the integration of an AI-assisted relevance screening workflow that can be documented and communicated transparently, such that the process can be audited or reproduced. Toward this end, we propose a dual-stage evaluation of model transparency based on computational metrics and feedback from library evidence synthesis professionals. The following work plan is divided into two phases that span over two years. Table 1 shows the tasks that will be carried out in each phase of the project. In Phase 1, we will craft a prototype of an LLM-empowered screening workflow and a draft of a non-technical instruction manual intended for library evidence synthesis professionals. To inform Phase 2, Phase 1 will conclude with a preliminary focus group with evidence synthesis professionals at the University Libraries at Virginia Tech. In Phase 2, we will refine the prototype and instruction manual using feedback from the

Table 1: Planning project phases and key activities

Phase	Tasks
Phase 1	1. Data Collection
	2. Baseline Experimentation & Iterative Prototyping
	3. Draft Instruction Manual
	4. Preliminary Focus Group
Phase 2	5. Process Preliminary Focus Group Results
	6. Controlled Experiments
	7. Instruction Manual Refinement
	8. Focus Group

Phase 1 preliminary focus group. Phase 2 will conclude with a second focus group with representation from evidence synthesis librarians at universities nationwide.

Phase 1

Phase 1 of our project will be conducted in the first year of the funding period. The phase comprises four tasks as depicted in the first half of Table 1. This phase of the project prioritizes developmental activities that will be built upon in Phase 2. The first task is to collect the data that will serve as case studies. We then establish the experimental setup to integrate AI within the screening workflow using various openly available LLMs. This will become the prototype AI workflow. During prototype development, we will also begin drafting a non-technical instruction manual. This phase will conclude with a preliminary focus group of library evidence synthesis professionals who will pilot the prototype and instruction manual.

Task 1: Data Collection. During this stage, we will identify and prepare datasets that will be used to develop and evaluate an AI workflow for relevance screening. To accelerate experimentation, as evidence synthesis reviews can take years to complete, we will use title and abstract screening results from completed systematic reviews conducted in partnership with the Evidence Synthesis Services at Virginia Tech University Libraries. Because this service supports research across multiple disciplines, the data will include reviews from fields such as education, environmental sciences, agricultural sciences, veterinary medicine, public health, medicine, and food studies. Using cases drawn from multiple disciplines allows us to evaluate the transferability of the workflow across different evidence synthesis contexts.

The research team has already examined two completed reviews as preliminary cases. In these reviews, literature was collected from seven and eighteen databases, yielding 6,475 total (2,872 unique) and 5,938 total (3,195 unique) records, respectively. The research questions addressed public health topics, including environmental impacts on pregnancy outcomes and healthcare resource use among elderly populations. As documented in the letters of support, faculty collaborators at Virginia Tech have expressed a strong interest in contributing additional screening datasets to the project.

Informed by our prior experience (Banerjee, 2024), we will preprocess the datasets to prepare them for experimental analysis. For each of the case studies, the metadata will undergo the following steps: *(1)* text normalization and whitespace removal, *(2)* removal of duplicate records and empty fields, and *(3)* correction of metadata anomalies such as invalid date values or truncated abstracts. The resulting datasets will be stored in institutional file storage protected by university authentication controls.

Task 2: Baseline Experimentation and Iterative Prototyping. In this task we establish an experimental pipeline that serves as the baseline for developing the AI-assisted relevance screening prototype. The pipeline will evaluate the use of openly available language models with modest computational require-

ments, including Llama-3 (Llama Team, 2024) and Mistral (Jiang et al., 2023). We begin with smaller parameter models and scale model size as needed during experimentation.

Building on our prior work (Banerjee et al., 2024; Ingram et al., 2024, 2025b), the models will be prompted to classify scholarly articles into the two classes used in evidence synthesis screening: "relevant" and "non-relevant." Each model will receive the inclusion and exclusion criteria associated with the review case and will classify an article as relevant if all inclusion criteria are satisfied and no exclusion criteria are met. Classification will be performed using only the title and abstract, reflecting the initial screening stage used in systematic reviews.

Human screening decisions from completed evidence synthesis reviews will serve as the ground truth for evaluation. Model performance will be assessed using precision, recall, F1 score, and overall accuracy (Rijsbergen, 1979). Because evidence synthesis requires comprehensive identification of relevant literature, achieving high recall for the "relevant" class will be prioritized to minimize the risk of excluding relevant studies.

During this stage we will conduct iterative experimentation to examine how prompt formulation and model parameter settings influence classification performance. The best-performing configuration will be used to develop the initial screening prototype. More extensive prompt optimization will be conducted in Phase 2, Task 6 of the project.

Task 3: Instruction Manual. As the prototype stabilizes, we will document the AI-assisted screening workflow in an instruction manual describing how the system can be implemented and used in evidence synthesis projects. Written for a non-technical audience of evidence synthesis librarians and information specialists, the manual will describe the workflow, system requirements, and operational steps needed to deploy the prototype. The documentation will include a step-by-step demonstration illustrating how the system can be configured and applied to a screening dataset.

Task 4: Preliminary Focus Group. Toward the end of Phase 1, we will conduct an informal preliminary focus group in which the affiliates of the Evidence Synthesis Services at Virginia Tech will test the AI-assisted screening workflow. The purpose of this task is to evaluate the usability, transparency, and practical utility of the prototype and accompanying instruction manual, complementing the quantitative experiments conducted in Task 2. This task evaluates the explainability of AI decisions, which is a major concern in evidence synthesis workflows.

Participants will attend an initial training session during which the PI and Co-PIs will introduce the AI-assisted prototype, demonstrate the workflow, and explain how to interpret the screening results. Participants will then receive the draft instruction manual and apply the workflow to one or more of the project's case study datasets. They will be given a minimum of two weeks to complete these tasks and will also be asked to record semi-structured notes describing their experience. Participants will be asked to reflect on the clarity of the instruction manual, the ease of using the workflow, and the transparency and interpretability of the AI-generated screening decisions. Feedback collected during this task will inform revisions to the prototype and workflow in later phases of the project.

Phase 2

Task 5: Process Preliminary Focus Group Results. This task analyzes feedback collected during the preliminary focus group described in Phase 1, Task 4. The purpose of this analysis is to identify practical issues and opportunities for improving the AI-assisted screening prototype and workflow before further development. We will use an inductive coding approach to analyze the written responses provided by participants. Through this process we will identify recurring themes in the participants' responses. Key themes and trends will be summarized and recorded in an aggregate document separate from the individual responses. These findings will inform revisions to the prototype, documentation, and experimental design in subsequent phases of the project.

Task 6: Controlled Experiments. In this task we extend the experiments conducted in Phase 1 by performing structured controlled experiments to refine the AI-assisted screening prototype. The experimental design will also incorporate insights from the preliminary focus group conducted in Phase 1, Task 4. For each model selected during Phase 1, Task 2, we will conduct a series of controlled experiments to examine sources of variability in AI-assisted screening decisions.

1. **Prompt Sensitivity.** Within each model we will evaluate different prompting strategies, including zero-shot, few-shot, and chain-of-thought prompting. Prior research has shown that model predictions can be sensitive to prompt formulation (Errica et al., 2025; Sorin et al., 2025). We will systematically vary prompts and document how prompt changes affect screening predictions in order to identify strategies that minimize variability and improve reproducibility.
2. **Model Divergence.** We will measure divergence in prediction results across models when identical datasets and prompts are used. Differences in training data, architecture, and context sensitivity can lead to variation in LLM outputs. Building on our prior work on learning from LLM disagreement (Ingram et al., 2025a), we will quantify inter-model agreement using metrics such as Cohen's Kappa inter-rater reliability (Cohen, 1960).
3. **Run-to-Run Variance.** Because LLM outputs can be non-deterministic, we will evaluate prediction stability by running the same model and prompt configuration multiple times. Variability across runs will be measured using standard deviation and related statistical measures (Wasserman, 2004). Lower variance across repeated runs indicates greater stability and improved reproducibility of AI-assisted screening decisions.

These controlled experiments will allow us to document how model choice, prompting strategies, and stochastic variability influence the reproducibility and transparency of AI-assisted relevance screening workflows. The results will provide empirical evidence about the stability of AI-assisted screening methods and identify conditions under which such workflows can be considered reliable for evidence synthesis practice.

Task 7: Refine Instruction Manual. During the controlled experiments described in Task 6, we will refine the instruction manual to incorporate feedback and insights from the Phase 1 preliminary focus group (Task 4) and to reflect adjustments made to the prototype. These revisions will ensure that the documentation accurately describes the finalized workflow and provides clear implementation guidance for evidence synthesis practitioners.

Task 8: Focus Group. In the conclusion of Phase 2, we will conduct a final focus group with evidence synthesis library professionals from institutions across the U.S. Drawing on lessons from the Phase 1 preliminary focus group (Task 4), we will refine the structure of the session, including the format, supporting instructional materials, and reflection question prompts used during the discussion. As indicated by letters of support from evidence synthesis professionals at the University of Minnesota, Colorado State University, the University of Tennessee Health Sciences Center, and Carnegie Mellon University, our colleagues are enthusiastic about participating in and assisting with recruitment for this focus group.

Reporting and Sharing (All Phases):

During each phase of the project, we will document results and disseminate knowledge to the evidence synthesis and broader library communities. Instructional materials and implementation guidance developed throughout the project will be shared to support practitioner evaluation and adoption of the workflow. By involving evidence synthesis library professionals at multiple stages of the project, we will incorporate practitioner feedback to ensure that the resulting methodology aligns with established evidence synthesis practices. Detailed dissemination activities are described in the Project Results section.

Evaluation and Demonstrating Impact:

Feedback from the preliminary focus group (Phase 1, Task 4) and the Phase 2 focus group (Task 8) will be used to evaluate the auditability, transparency, and practical utility of the AI-assisted screening prototype and workflow. In addition to qualitative evaluation from practitioner feedback, we will make the prototype and instruction manual publicly available to enable further testing and use by the evidence synthesis community. Project impact will be assessed through dissemination and scholarly engagement, including publications describing the methodology and results, as detailed in the Project Results section. Indicators of impact will include citation and usage metrics for these outputs, participation in project workshops and demonstrations, and evidence of practitioner interest in testing or adapting the workflow within their own evidence synthesis services.

Project Team

The research will be conducted at Virginia Tech University Libraries. Responsibilities will be shared between PI Banerjee and Co-PIs Comer and Ingram. The project integrates an team of experts with in the library with PI and Co-PIs with experiences in computational science, evidence synthesis and library and information sciences.

Dr. Bipasha Banerjee will serve as the Principal Investigator and Project Director. Dr. Banerjee is the AI Research Scientist at the University Libraries at Virginia Tech. She will lead the analytical part and the design and implementation of LLM experimentation. Her research lies at the intersection of natural language processing, machine learning, information retrieval, AI and digital libraries. She got her Ph.D. in Computer Science at Virginia Tech in 2024 and is a trained library faculty member with a well-established background in computer science and digital libraries. She will also oversee the progress and ensure that the team adheres to the milestones outlined in the timeline. She is also responsible for ensuring that the project adheres to stipulated guidelines with IRB and other regulatory bodies and the overall financial aspects of the project. She will co-supervise the GRA, leveraging her interdisciplinary expertise to create a conducive environment for research and learning.

C. Cozette Comer will serve as the Co-Principal Investigator on this project. Comer is the Assistant Director of Evidence Synthesis Services with nearly a decade of evidence synthesis experience, having served as methodological expert, search expert, content expert, and project manager across a variety of projects and fields. She holds a M.S. in Sociology and is pursuing a Ph.D. in Public Administration and Public Affairs, both at Virginia Tech. Comer will lead coordinate and lead focus groups, as well as the development of the instruction manual. Throughout the project, she will be responsible for ensuring adherence with evidence synthesis guidelines and emerging best practices regarding the development of AI workflows in evidence synthesis. She will also co-supervise the GRA working on this with PI Banerjee.

William A. Ingram will serve as the Co-Principal Investigator on this project. Ingram serves as Associate Dean and Executive Director for Information Technology and Associate Professor at the Virginia Tech University Libraries and as Director of the Center for Digital Research and Scholarship. His research examines the use of artificial intelligence, natural language processing, and large-scale information retrieval to analyze scholarly corpora and support computational approaches to research synthesis. In this project, Ingram will oversee research design, conceptual framing, and will guide the development of auditable workflows and evaluation frameworks for AI-assisted evidence synthesis. He will also ensure that project methods and artifacts remain transparent, reproducible, and reusable, enabling research libraries to evaluate the feasibility of integrating AI-assisted evidence synthesis into future library services.

Collaborative Structure: The project benefits from the interdisciplinary collaboration of the PI and Co-PIs, whose combined expertise spans evidence synthesis methodology, artificial intelligence, digital libraries, and research services. Together, the investigators will guide the development of AI-assisted evidence synthesis workflows, ensuring that proposed methods remain auditable, methodologically rigorous, and responsive to researcher needs. Banerjee and Comer will engage with stakeholders, including evidence synthesis librarians at Virginia Tech and at peer institutions, to gather feedback on emerging workflows and service models and to inform feasibility assessments for future implementation. All investigators will jointly monitor project progress to ensure alignment with user needs and established academic standards, while overseeing risk management and contingency planning to support the project's adaptability and long-term viability.

Graduate Research Assistant: A Ph.D. in Computer Science at Virginia Tech with interests in machine learning and information retrieval will be supported through this project. The GRA will contribute to technical development and experimentation, including data organization and preprocessing, development of AI-assisted workflows for evidence synthesis, and implementation of project code and algorithms. The student will also assist with analysis of focus groups insights, data processing, and AI model training. The GRA will maintain detailed records of experimental design, methodologies, and results, and contribute to project reporting and dissemination activities, including article preparation, conference presentations, and workshops. The student will participate fully in weekly project meetings and collaborative research activities. Mentorship for the GRA will follow structured graduate mentoring practices modeled on NSF mentoring plans (NSF, 2024), including regular advising meetings, training in responsible and reproducible research practices, opportunities for interdisciplinary collaboration, and support for professional development. Working in the library will provide the student with unique interdisciplinary experience at the intersection of artificial intelligence, digital libraries, and research services.

The Evidence Synthesis Services at Virginia Tech: The Evidence Synthesis Services (Comer, 2022) at the University Libraries at Virginia Tech will participate in the preliminary focus group (Phase 1, Task 4), providing instrumental insight into the development of this novel AI-assisted screening workflow.

The Evidence Synthesis Services provide training and support for evidence synthesis reviews across disciplines. In addition to workshops, classroom instruction, and consultations, the team co-authors many evidence synthesis reviews as the search expert and methodological guide. These services are also dedicated to fostering the global evidence synthesis librarian community and profession through several initiatives, such as co-founding and co-convening an international library evidence synthesis conference.

Collectively, the core team of three full time faculty (not including Co-PI Comer) has over 30 years of experience in evidence synthesis librarianship. The services also facilitate a volunteer group of other library employees at Virginia Tech who support evidence synthesis intermittently. By leveraging both highly experienced and consistently engaged evidence synthesis professionals, as well as more those who engage with evidence synthesis more casually, the preliminary focus group will elucidate trends among these groups.

Campus Resources: PI Banerjee and Co-PIs Comer and Ingram are associated with the Center for Digital Research and Scholarship (CDRS, 2024) at the University Library, which gives this project access to interdisciplinary context and expertise. For the research environment, we will make use of local servers with GPUs at the library alongside the University's Advanced Research Computing Cluster, a high-performance computing cluster with access to sophisticated GPUs. The proposed study, which focuses on developing AI-workflows informed by the needs of evidence synthesis librarians, is aligned with the Evidence Synthesis Service initiatives at Virginia Tech and the IMLS goals of championing life learning and strengthening community engagement.

PROJECT RESULTS

Through this planning project, we will work toward the development of an experimental prototype through continuous refinement based on feedback with library evidence synthesis professionals. We will share our project findings through an instruction manual, accompanying slides, and scholarly peer-reviewed publications. The deliverables for the project are as follows:

1. **AI-enabled ES Prototype**: As a result of the experiments conducted during this planning proposal, we will have a prototype for performing transparent and auditable title and abstract screening using an LLM-assisted workflow. All prototype software and code will be openly available for download and reuse. All technical products will also be accompanied by detailed instructions and a README to guide installation and use.
2. **Instruction Manual**: This non-technical instruction manual will accompany the AI-workflow prototype and will also be openly available for download and reuse.
3. **Scholarly Dissemination**: Our work is aimed at impacting stakeholders and academia alike.
   (a) ***Sharing with stakeholders:*** Our stakeholders comprise two distinct groups: (1) researchers who conduct evidence synthesis, and (2) evidence synthesis library professionals who provide services to these researchers. Aggregate findings from focus groups will be shared with participants of the focus groups. Overall findings regarding the prototype and workflow will be shared with researchers and library professionals who participate in the focus groups.
   (b) ***Sharing with academic research community:*** We plan to share our findings with the academic community. We will submit our findings to top-ranking conferences and high-impact journals such as Journal of the Association for Information Science and Technology (JASIST). We will also submit our findings to information retrieval and digital library conferences such as ACM/IEEE Joint Conference on Digital Libraries (JCDL), Theory and Practice of Digital Libraries (TPDL), ACM Special Interest Group on Information Retrieval (SIGIR). A presentation abstract regarding this project has already been submitted to the Cochrane Colloquium, an international conference hosted by the Cochrane Collaboration. We will also submit to other evidence synthesis oriented events such as the What Works Global Summit.
4. **Comprehensive Report**: We will develop a comprehensive report outlining our key findings from the project phases.
5. **Graduate Learning**: Our budget has allocated funds to support a GRA. The GRA will help achieve the project goals by performing the tasks outlined in the narrative. It also offers the student an exposure to interdisciplinary research, learn research methodologies, and thereby grow their academic career. Participating in publishing also helps the student gain research visibility and helps the project demonstrate its impact.

Auditable AI Workflows for Evidence Synthesis Library Services

Bipasha Banerjee, C. Cozette Comer & William A. Ingram

Schedule of Completion

	Year 1 (2026–2027)											
Activities and Milestones	**S**	**O**	**N**	**D**	**J**	**F**	**M**	**A**	**M**	**J**	**J**	**A**
Administrative Tasks												
Launch Project *Lead: Banerjee, Comer, Ingram*												
Hire Graduate Research Assistant *Lead: Banerjee & Comer*												
Review Digital Product Plan *Lead: Banerjee*												
Develop And Review Project Work Plan *Lead: Banerjee*												
Review Monthly Budget Reports *Lead: Banerjee*												
Create and Update Project Website/Documentation *Lead: GRA*												
Annual Reporting *Lead: Banerjee*												
Phase 1												
Data Collection *Lead: Banerjee & Comer*												
Baseline Experimentation and Iterative Prototyping *Lead: Banerjee*												
Draft Instructional Manual *Lead: Comer*												
Preliminary Focus Group *Lead: Banerjee & Comer*												
Phase 2												
Process Preliminary Focus Group Results *Lead: Banerjee & Comer*												
Controlled Experiments *Lead: Banerjee*												
Reporting and Sharing												
Draft and Submit Conference Poster/Short Paper *Lead: Banerjee, Comer & Ingram*												

	Year 2 (2027–2028)											
Activities and Milestones	**S**	**O**	**N**	**D**	**J**	**F**	**M**	**A**	**M**	**J**	**J**	**A**
Administrative Tasks												
Review Digital Product Plan *Lead: Banerjee*												
Review And Update Project Work Plan *Lead: Banerjee*												
Review Monthly Budget Reports *Lead: Banerjee*												
Update Project Website/Documentation *Lead: GRA*												
Final Reporting *Lead: Banerjee, Comer, & Ingram*												
Project Wrap Up *Lead: Banerjee*												
Phase 2												
Process Preliminary Focus Group Results *Lead: Banerjee & Comer*												
Controlled Experiments *Lead: Banerjee*												
Refine Instruction Manual *Lead: Comer*												
Focus Group *Lead: Banerjee & Comer*												
Reporting and Sharing												
Draft and Submit Conference/Journal Paper *Lead: Banerjee, Comer & Ingram*												
Draft and Submit Conference Workshop Proposal *Lead: Banerjee & Ingram*												
Draft and Share Final Instructional Guide *Lead: Comer*												