INSTITUTE of Museum and Library SERVICES

NATIONAL LEADERSHIP GRANTS FOR LIBRARIES

Sample Application LG-259721-OLS-26
Project Category: Applied Research

University of North Texas College of Information

Amount awarded by IMLS:	$490,396
Amount of cost share:	$0

The Department of Data Science in the College of Information at the University of North Texas will enhance access to and public engagement with oral history collections through the development of ORAL-Artificial Intelligence (AI), a responsible and privacy aware framework for navigating multimodal oral digital archives. This project will evaluate the limitations of emerging multimodal AI systems in oral history contexts and develop a structured benchmark dataset and modular, open-source framework that supports cross-modal retrieval, evidence-grounded question answering, and narrative summarization while preserving archival integrity. By centering ethical stewardship and human-centered design, ORAL-AI hopes to expand meaningful public access to oral histories nationwide.

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.

ORAL-AI: High Efficacy Content Representation and User Engagement of Oral Digital Collections with Multimodal LLMs

1 Introduction

The University of North Texas, in collaboration with the University of Illinois Urbana Champaign, seeks $490,396 for a 3-year Applied Research grant project that aims to expand meaningful public access to large-scale oral history collections by developing and evaluating a responsible, privacy-aware multimodal AI framework that supports cross-modal semantic retrieval, evidence-grounded question answering, and contextual narrative summarization. By strengthening libraries' capacity to responsibly integrate emerging AI technologies into digital collections, the project advances National Leadership Grants for Libraries (NLGL) Goal and **Objective 3: Provide broad access to and preservation of information and collections through libraries and archives**. The project will pilot and evaluate the framework across three nationally significant collections—the Densho Digital Repository (Densho collection), the Library of Congress Civil Rights History Project (Civil Rights collection), and the Veterans History Project (Veterans History collection)—representing diverse historical contexts, multimedia formats, and narrative complexity. Building on the team's prior empirical studies of oral history collections and large language model evaluation, the project will investigate three research questions: **RQ1:** What are the capabilities, limitations, and risks of current multimodal large language models in oral digital collections? **RQ2:** How can multimodal retrieval and generation systems be designed to mitigate identified limitations while preserving interpretive integrity and professional stewardship? **RQ3:** How does AI-assisted multimodal access affect user engagement, comprehension, and professional practice compared to traditional discovery systems?

2 Project Justification

2.1 Statement of Need

Oral history collections represent one of the largest bodies of multimedia cultural heritage content stewarded by U.S. libraries and archives. The Veterans History Project alone contains more than 110,000 veteran collections, while institutions nationwide preserve tens of thousands of additional interviews documenting civil rights activism, migration, war, and community memory. The Densho collection and the Civil Rights collection further exemplify nationally significant, publicly accessible oral history archives.

Although digitization has expanded availability, meaningful access remains limited: interviews often span hours, transcripts extend dozens of pages, and metadata depth varies widely across institutions. Traditional keyword search rarely captures the nuance of spoken narrative, where meaning is conveyed through storytelling, tone, memory, and context rather than standardized terminology. In transcript-poor collections, meaningful discovery may require listening to extensive recordings without clear entry points. As a result, educators, researchers, students, and community members face significant barriers when attempting to locate relevant stories or themes.

Recent IMLS-funded initiatives (e.g., LG-256565-OLS-24; LG-256703-OLS-24) highlight growing interest in generative AI for digital collections. Yet few studies have rigorously evaluated multimodal large language models in oral testimony contexts or examined interpretive risks such as hallucination, demographic bias, and privacy governance (Perks, 2025). Oral histories are especially vulnerable to distortion. Research demonstrates substantial ambiguity even among trained annotators interpreting oral testimony (Gref et al., 2022), and scholars caution that generative AI may oversimplify historically situated or culturally sensitive narratives without safeguards (Lee-Talbot, 2025; Dueck, 2024). From a socio-technical systems perspective, the integration of AI into cultural heritage environments requires balancing technological capabilities with human expertise, institutional values, and ethical stewardship (Baxter & Sommerville, 2011). Libraries therefore require applied research that systematically benchmarks model performance, embeds evidence-grounded constraints, and integrates professional oversight into system design. **ORAL-AI** addresses this national need by evaluating multimodal AI across three nationally significant collections, developing a privacy-aware and human-centered framework, and measuring its impact on public engagement.

2.2 Related Work and Research Gaps

2.2.1 Advances in AI for Oral History and Cultural Heritage Collections

Recent scholarship shows growing momentum in applying large language models and related AI technologies to cultural heritage materials. Studies have explored text classification, sentiment analysis, thematic modeling, and knowledge graph construction in oral history archives (Cherukuri et al., 2025; Pessanha & Salah, 2021; Sun et al., 2024; Yi & Yin, 2025). Speech technologies have improved transcription and accessibility, and generative systems have been proposed to enhance interaction and metadata enrichment across archival collections (Draxler et al., 2024). Large-scale initiatives indexing war testimony archives and Indigenous oral histories further demonstrate technical feasibility and institutional interest (Dueck, 2024). However, most existing systems remain modality-specific, focusing primarily on transcript processing or metadata enhancement (Humphries et al., 2025; Reusens et al., 2025) rather than integrating transcripts, audio, video, and images within a unified multimodal alignment and retrieval framework.

Research Gap 1: The field lacks a validated benchmark and implementation framework for multimodal cross-modal retrieval and evidence-grounded summarization tailored specifically to oral history collections.

2.2.2 Interpretive Ambiguity, Bias, Privacy, and Ethical Concerns

Oral histories present distinctive interpretive and ethical challenges. Research demonstrates substantial ambiguity even among trained annotators analyzing emotion and sentiment in oral testimony, underscoring the contextual and subjective nature of narrative interpretation (Gref et al., 2022). Unlike structured records, oral testimonies embed layered memory, emotion, and culturally situated meaning that resist reduction to standardized categories. Scholars have cautioned that generative AI systems may oversimplify historically situated narratives, introduce hallucinated interpretations, or amplify demographic bias if deployed without safeguards (Lee-Talbot, 2025; Dueck, 2024). Privacy further complicates implementation (Mohan, 2024). Many commercial AI systems rely on cloud-based processing, raising concerns about data governance and institutional control—particularly for collections involving veterans, Indigenous communities, immigrants, or other vulnerable populations (Oldenburg & Papyshev, 2025). Even publicly accessible materials require careful stewardship when subjected to automated representation.

Research Gap 2: There is insufficient applied research measuring interpretive fidelity, bias, hallucination risk, and privacy implications when deploying multimodal generative AI in culturally sensitive oral history collections.

2.2.3 Public Engagement and Professional Practice in Oral History Archives

AI-driven conversational interfaces and knowledge graph systems suggest new models of interaction with archival collections (Zhuang & Ma, 2025). Industry initiatives demonstrate scalable indexing of large testimony archives (Gustman et al., 2002). However, most systems emphasize technical capability rather than evaluated impact. Few studies rigorously assess how AI-assisted access affects user comprehension, narrative understanding, perceived transparency, trust, or librarian workflow (Neamatollahi & Danesh, 2026; Regona et al., 2026). From a user-centered design and participatory information services perspective, evaluating technology adoption requires examining how systems reshape interpretive practices and professional roles (Dervin, 1998; Wilson, 1999). Professional organizations have recently called for evaluated case studies of AI in oral history practice, reflecting a field-wide demand for evidence-based guidance[1,2].

Research Gap 3: There is a lack of user-centered, practice-grounded research measuring how multimodal AI systems influence engagement, comprehension, and professional archival practice.

2.3 Preliminary Work

In a large-scale study of the Densho Digital Repository, we analyzed transcript structure, thematic patterns, and sentiment using natural language processing, demonstrating both the analytical potential and interpretive complexity of computational approaches to oral testimony (Chen et al., 2024). Building on this work, we conducted one of the first empirical evaluations of large language models applied to oral history transcripts, identifying measurable annotation ambiguity and output variability (Cherukuri et al., 2025). These findings directly motivated the need for structured benchmarking, hallucination auditing, and human-in-the-loop safeguards—core components of ORAL-AI. Our research on LLM-based summarization developed reference-free evaluation and reflective improvement methods and examined

[1] https://oralhistory.org/ai/
[2] https://www2.archivists.org/publications/book-publishing/AIstatement

question-answering–based summarization strategies (Nguyen et al., 2024; Ding et al., 2024), producing validated metrics for assessing narrative fidelity and grounding in long-form testimony. In parallel, we developed multimodal modeling techniques for complex document representation and cross-modal alignment (Ying et al., 2024), establishing the technical foundation necessary for transcript–audio–video integration. Beyond research outputs, the team has played a leadership role in advancing national dialogue on AI in cultural heritage, including organizing the JCDL 2024 workshop on AI/ML for archival collections, convening an ASIS&T 2024 panel on AI in GLAM institutions, and launching a field-wide special issue on large language models and generative AI for GLAM collections. In addition, we also conducted a structured audit of the Densho collection, Civil Rights collection, and Veterans History collection to assess transcript completeness, metadata richness, audiovisual structure, and access constraints. The audit revealed substantial variation across collections, reinforcing the need for flexible alignment mechanisms and professional oversight workflows.

3 Work Plan

ORAL-AI follows a deliberate progression from empirical assessment to responsible system development to validated public deployment, which is informed by socio-technical systems theory (Baxter & Sommerville, 2011) and user-centered design frameworks (Norman, 2013). The project employs a mixed-methods design that integrates quantitative benchmarking, qualitative interpretive analysis, and user-centered evaluation. Rather than beginning with implementation, we first establish a rigorous evidence base regarding the capabilities, limitations, and ethical risks of multimodal large language models in oral history contexts. Findings from this assessment directly inform system architecture and governance design. The resulting framework is then evaluated through structured usability testing, surveys, and interviews to determine whether it improves access, comprehension, and engagement without compromising interpretive integrity.

3.1 Data Collections

To evaluate and refine the ORAL-AI framework, we will pilot its methods across three nationally significant, publicly accessible oral history collections that differ in scale, subject matter, narrative structure, and metadata richness: (1) **The Densho collection** documents the incarceration of Japanese Americans during World War II through fully transcribed video interviews, photographs, and related archival materials. Its structured metadata, thematic indexing, and culturally sensitive testimony make it an ideal testbed for evaluating cross-modal retrieval, evidence-grounded responses, and hallucination mitigation within a well-curated multimedia corpus. (2) **The Civil Rights collection** preserves firsthand accounts of activists and participants in the Civil Rights Movement. The collection's interpretive depth, diverse narrators, and historically complex narratives allow assessment of how multimodal systems handle nuanced testimony, thematic comparison across interviews, and context-sensitive summarization. (3) **The Veterans History collection** contains a large-scale corpus of oral histories, letters, diaries, and photographs from U.S. veterans spanning multiple conflicts. Its demographic breadth and multimodal diversity enable evaluation of system performance in heterogeneous collections, including grounding accuracy, bias detection, and cross-modal alignment at scale.

3.2 Research Objective 1: Systematically Assess the Capabilities and Limitations of Multimodal LLMs for Oral Digital Collections (RQ1, Sep 2026 – Apr 2027)

Research Objective 1 will establish an empirical foundation for responsible multimodal AI integration in oral history collections. Using a triangulated evaluation framework, we will systematically assess the performance, interpretive fidelity, bias patterns, and privacy implications of leading multimodal large language models (MLLMs) across three selected oral history collections. This benchmarking approach draws on evaluation research and algorithmic auditing frameworks, which emphasize systematic, reproducible assessment of model performance, bias, and failure modes in real-world contexts (Mitchell et al., 2019).

3.2.1 Task 1.1 Design Benchmark and Develop Queries

The first task is to construct a domain-specific benchmark tailored to oral digital collections. We will develop 10 structured user queries per collection (**30 total**), designed to reflect authentic research and public engagement scenarios. Queries will span three task types: (1) cross-modal semantic retrieval (e.g., locating relevant video/audio segments across interviews), (2) evidence-grounded question answering (e.g., identifying historically specific responses within testimony),

and (3) narrative summarization (e.g., generating contextualized summaries of thematic segments). Queries will vary in complexity, including factual historical questions, interpretive narrative questions, cross-interview thematic comparisons, and emotion-centered prompts. Each query will be designed to require alignment across transcript text, audio/video content, and metadata where applicable.

3.2.2 Task 1.2 Evaluate Commercial and Open MLLMs

Using the benchmark, we will systematically evaluate three widely used multimodal models: ChatGPT, Google Gemini, LLaMA, and others if recommended by the advisory board. Each model will be tested under three prompting strategies: Zero-shot prompting, Few-shot prompting, and Chain-of-thought prompting. This design enables analysis of both model capability and prompting sensitivity. For each query-model-prompting combination, we will collect outputs and evaluate across four dimensions: (1) Factual correctness and grounding: precision of retrieved segments, citation alignment to source transcripts, and evidence traceability. (2) Hallucination rate: unsupported factual claims, fabricated details, misattributed quotations, cross-modal inconsistencies. (3) Bias across demographic narratives: differences in summarization tone across racial, ethnic, and veteran identities; emotional mischaracterization patterns; overgeneralization or stereotype amplification. (4) Privacy and data governance risks: whether model interactions require external data transmission; data retention disclosures; risks of unintended memorization or content exposure. This evaluation will provide one of the first structured empirical comparisons of MLLM behavior in oral history contexts.

3.2.3 Task 1.3 Synthesize Limitations of Current Systems

Building upon the empirical findings from Tasks 1.1–1.2, this task will systematically synthesize the technical and ethical limitations of current multimodal large language models when applied to oral digital collections. Rather than treating evaluation results as isolated performance metrics, we will conduct a structured comparative analysis across models, prompting strategies, and collections to identify recurring patterns of failure and risk. This synthesis will focus on five critical dimensions: (1) Core failure modes in cross-modal alignment, (2) Conditions under which hallucination increases, (3) Prompting strategies that reduce interpretive distortion, (4) Privacy vulnerabilities in external API usage, and (5) Demographic disparities in summarization or retrieval performance.

This synthesis will produce: (1) A formal limitations report documenting technical, interpretive, and governance risks, (2) A comparative model performance matrix across collections and prompting strategies, (3) A de-identified, publicly shareable benchmark dataset and evaluation protocol (subject to collection policies), and (4) Evidence-grounded guidelines for human-in-the-loop safeguards and professional review workflows.

3.2.4 Task 1.4 Generate Design Requirements

Following the identification of limitations, this task translates empirical findings into actionable system design requirements. The goal is to define clear technical and governance specifications for a responsible multimodal framework tailored to oral digital collections. Using the evidence generated in Task 1.3, we will articulate design requirements across four domains: (1) Cross-modal alignment architecture, (2) Hallucination mitigation and evidence grounding, (3) Bias monitoring and interpretive safeguards, (4) Privacy-aware deployment specifications. These design requirements will be reviewed with advisory board members and partner institutions to ensure feasibility within library environments.

The outputs of Task 1.4 will include: (1) A technical and ethical design specification document for ORAL-AI. (2) A privacy-aware deployment blueprint for local MLLM integration. (3) Implementation guidelines aligned with professional standards in libraries and archives.

3.3 Research Objective 2: Design and Implement a Responsible, Privacy-Aware Multimodal ORAL-AI Framework (RQ2, May 2027 – Nov 2027)

Building on Objective 1 findings, the second objective focuses on system design and algorithmic development. ORAL-AI is designed not simply as an AI-powered search tool, but as a responsible infrastructure for engaging oral digital collections. As illustrated in Figure 1, the framework integrates transcript text, audio, video frames, and metadata into a structured cross-modal pipeline that supports: cross-modal semantic retrieval, evidence-grounded question answering, context-sensitive narrative summarization, transparent provenance tracking, human-in-the-loop feedback, and privacy-preserving deployment. We apply socio-technical design principles (Baxter & Sommerville, 2011; Amershi et al., 2019)

and human-AI collaboration theory to ensure that system architecture integrates technical safeguards with professional oversight and institutional governance.

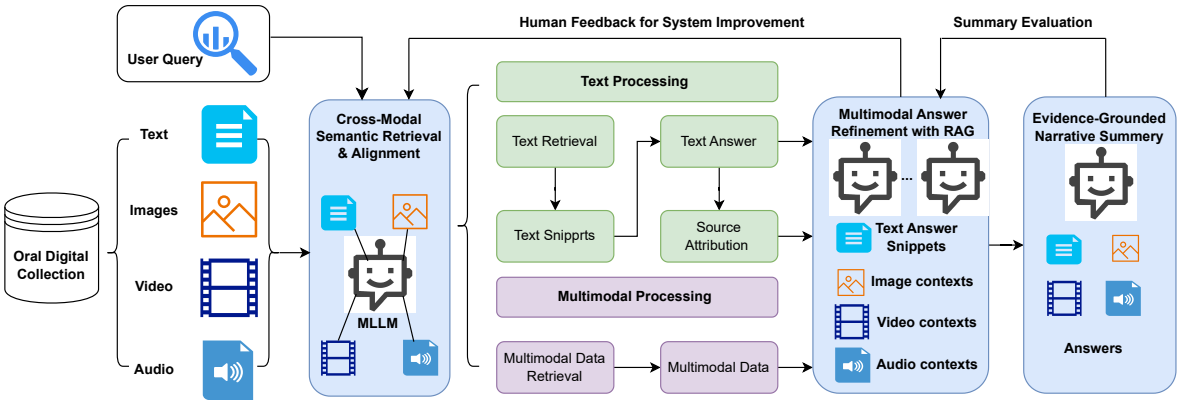

Figure 1: The ORAL-AI Conceptual Implementation Framework.

3.3.1 Task 2.1 Develop Cross-modal Alignment and Retrieval Mechanisms Tailored to Oral Histories

Oral history interviews are long-form, narratively complex, and temporally structured. Conventional multimodal systems often fragment testimony into disconnected segments or fail to preserve contextual continuity. Drawing on failure modes identified in Objective 1; this task designs a retrieval architecture tailored specifically to oral collections. We will construct a cross-modal embedding and indexing pipeline that:

- Segments interviews into semantically coherent narrative units rather than arbitrary time slices.
- Aligns transcript passages with precise audio timestamps and corresponding video frames.
- Integrates structured metadata (e.g., collection context, subject tags, interviewee demographics) into retrieval representations.
- Enables retrieval at multiple granularities: sentence-level, segment-level, interview-level, and cross-collection thematic level.

To implement this architecture, we will evaluate and integrate three candidate models: Llama 3.1–8B, MiniCPM-o 4.5, and Qwen3-VL, which are the state-of-the-art open source MLLMs. In addition, they are computationally feasible on our GPU infrastructure and capable of multimodal reasoning without mandatory cloud dependency. Performance will be evaluated using cross-modal retrieval precision and recall, temporal grounding accuracy, alignment fidelity between transcript and audiovisual segments. This task ensures that semantic retrieval respects narrative structure rather than flattening testimony into decontextualized fragments.

3.3.2 Task 2.2 Integrate Evidence-linking Strategies to Reduce Hallucination

Objective 1 will quantify hallucination rates and identify conditions under which unsupported claims increase. This task directly addresses those risks through architectural safeguards. ORAL-AI will incorporate a retrieval augmented generation (RAG) design in which generative outputs are constrained by retrieved evidence segments. Key features include: (1) Mandatory citation linking between generated responses and transcript timestamps, (2) Structured response templates requiring explicit evidence anchors, (3) Segment-level confidence scoring, and (4) Restriction of generation to retrieved content windows. Effectiveness will be measured through: (1) Reduction in hallucination rate relative to baseline models, (2) Expert scoring of interpretive fidelity, and (3) Percentage of responses with traceable evidence support.

3.3.3 Task 2.3 Implement Human-in-the-loop Review Workflows

Oral history interpretation is inherently contextual and culturally sensitive. Fully automated responses risk flattening nuance or misrepresenting lived experience. This task ensures that professional oversight remains central to system operation. We will design review workflows that allow librarians, archivists, and historians to:

- Inspect retrieved segments and generated summaries
- Edit, annotate, or contextualize AI outputs
- Flag culturally sensitive content

- Approve responses prior to public display where required

The system will incorporate: (1) review dashboard with evidence visualization, (1) version tracking and edit history, (3) logging of reviewer feedback for iterative refinement, and (4) automated triggers for mandatory review on high-risk queries. Usability studies will assess time required for review, perceived interpretive control, trust and transparency ratings, and impact on professional workload.

3.3.4 Task 2.4 Deploy Open-source Multimodal LLMs

Privacy and data sovereignty are central concerns for oral history collections, especially those involving veterans, Indigenous communities, and civil rights testimony. Many commercial AI systems require cloud-based processing that transmits data externally. ORAL-AI will instead deploy selected models locally within institutionally controlled computing environments. Llama 3.1–8B, MiniCPM-o 4.5, and Qwen3-VL will be evaluated for local inference performance and integrated using containerized infrastructure. Local deployment will ensure:

- No automatic transmission of collection content to external services
- Encrypted storage of embeddings and indexed segments
- Transparent logging and audit trails
- Configurable access controls based on collection restrictions

Comparative testing will assess latency and GPU resource requirements, alignment performance, hallucination mitigation effectiveness, sustainability for mid-sized institutions. By prioritizing open-source models, the project supports equitable adoption across libraries that lack commercial AI budgets while strengthening long-term institutional control.

3.3.5 Task 2.5 Incorporate Professional Standards of Archival Stewardship into System Architecture

Technical safeguards alone are insufficient without governance alignment. This task embeds archival and oral history principles directly into system design. ORAL-AI will integrate respect for access restrictions, metadata-preserving retrieval pipelines, clear labeling of AI-generated content, distinction between original testimony and machine-generated summary. Advisory board members will review governance components to ensure alignment with professional standards in archival management and oral history practice. System documentation will include ethical deployment guidelines, community consultation protocols, and adoption toolkits for libraries and archives.

3.4 Research Objective 3: Evaluate Usability, Interpretive Fidelity, and Impact on Public Engagement (RQ3, Sep 2027 - Oct 2028)

Objective 3 evaluates whether the system meaningfully improves access, interpretation, and engagement in practice. The user evaluation design is guided by information behavior and sense-making theory (Wilson, 1999; Dervin, 1998), which conceptualize information interaction as a contextual process of meaning construction rather than simple retrieval. Through case studies using the Densho collection, Civil Rights collection, and Veterans History collection, we will conduct mixed-method user studies comparing ORAL-AI with traditional keyword-based access systems. The goal is not only to measure technical accuracy, but to assess interpretive fidelity, professional trust, and public engagement outcomes.

3.4.1 Task 3.1 Develop and Test Training Materials and Survey Instruments

To ensure consistent and rigorous evaluation, we will first develop structured training materials and research instruments, the details are as follows:

- Training materials
  - A user guide explaining system capabilities and limitations
  - Ethical use guidelines emphasizing interpretive responsibility
  - Demonstration tasks illustrating cross-modal retrieval and evidence-grounded responses
- Measurements of survey and instruments
  - Retrieval precision and grounding accuracy (user-rated and expert-validated)
  - Quality and interpretive fidelity of generated summaries
  - User comprehension of narrative content
  - Engagement outcomes (depth of exploration, follow-up queries)
  - Perceived transparency and trust in AI-assisted responses

    - Impact on user workflows (time, cognitive load, interpretive control)
- Instruments design methods
    - Likert-scale survey items
    - Task-based scoring rubrics
    - Structured reflection prompts

3.4.2 Task 3.2 Recruit Evaluators

We will recruit approximately 80 evaluators across three groups: librarians and archivists (n=20), educators and researchers (n=20), public users and students (n=40). Participants will represent diverse disciplinary backgrounds and demographic perspectives. Recruitment will occur through partner institutions, professional networks, and public outreach channels. All participation protocols will receive Institutional Review Board (IRB) approval.

3.4.3 Task 3.3 Conduct Usability Testing

Participants will complete structured tasks using both: (1) A traditional keyword-based search interface and (2) The ORAL-AI multimodal interface. Tasks will reflect realistic scenarios, such as: (i) Identifying testimonies addressing a specific historical theme, (ii) Locating emotionally resonant passages, and (iii) Generating contextual summaries for classroom use. For each system, we will measure task completion time, retrieval success rate, evidence identification accuracy, summary quality ratings, cognitive workload, interaction depth (number of queries, session duration). This comparative design enables quantitative assessment of whether ORAL-AI improves accessibility and efficiency without sacrificing interpretive integrity.

3.4.4 Task 3.4 Interview Evaluators

Following usability testing, a subset of participants (n = 30) will participate in semi-structured interviews, which will provide qualitative depth to complement quantitative metrics. Interviews will explore:

- Perceived trustworthiness of AI-generated responses
- Interpretive concerns or distortions
- Clarity of evidence linkage
- Perceived loss or preservation of narrative voice
- Professional implications for archival practice
- Comfort with privacy-preserving features
- Workflow integration feasibility (for library and archive professions)
- Time savings versus review burden (for library and archive professions)
- Confidence in maintaining stewardship standards (for library and archive professions)

3.4.5 Task 3.5 Analyze Testing and Interview Results

We will use mixed methods to analyze the data to measure improvements in accessibility and engagement. Quantitative analysis will include paired t-tests, retrieval precision comparisons, hallucination reduction validation, workload score differences, and trust and transparency scale analysis. Qualitative analysis will use thematic coding to identify patterns related to interpretive fidelity, bias concerns, narrative preservation, and professional acceptance.

3.4.6 Task 3.6 Iteratively Revise the System based on Evaluation Results

Findings from usability testing and interviews will directly inform system refinement. We will implement the following revisions for potential improvement:

- Adjusting evidence citation formatting
- Strengthening hallucination constraints
- Improving cross-modal visualization
- Modifying review triggers
- Refining interface clarity

A second round of targeted validation testing will confirm that revisions improve performance and user satisfaction.

3.5 Research Objective 4: Develop Transferable Guidelines, Benchmarks, and Training Resources for Libraries (RQ1, RQ2, RQ3, Nov 2028 – Aug 2029)

3.5.1 Task 4.1 Write and Review Guidelines

Drawing on empirical findings from Objectives 1–3, the project team will develop a comprehensive set of best-practice guidelines (Table 1) for responsible multimodal AI integration in oral digital collections. Draft guidelines will undergo structured review by advisory board members, partner institution librarians, external oral history practitioners. Feedback will be incorporated to ensure clarity, feasibility, and professional relevance.

Table 1 The Guideline Components for Multimodal AI Integration in Oral Digital Collections

Guideline Components	Content
Multimodal retrieval and alignment	• Recommended segmentation strategies for long-form interviews • Cross-modal indexing practices • Context-preserving retrieval configurations
Evidence-grounded generation	• Citation-linking requirements • Hallucination mitigation protocols • Structured output formatting
Bias monitoring and interpretive safeguards	• Demographic parity checks • Trigger conditions for human review • Cultural sensitivity considerations
Privacy and data governance	• Local deployment models • Data isolation practices • API risk assessment checklists
Human-in-the-loop workflow design	• Review interface recommendations • Staff oversight models • Documentation and transparency standards

3.5.2 Task 4.2 Organize and Document Benchmarks

In this task, we will refine and release the benchmark developed in Research Objective 1. The information being release include: (1) A curated set of anonymized evaluation queries, (2) Performance scoring rubrics, (3) Ground-truth evidence annotations, (4) Hallucination detection criteria, and (5) Bias assessment methodology.

We will also document technical setup instructions, hardware requirement estimates, reproducibility guidance, ethical use considerations for responsible AI practice. All benchmark materials will be reviewed to ensure compliance with collection policies and privacy agreements prior to release. The benchmark will be archived in an open repository with persistent identifiers to support reuse by researchers and institutions nationwide.

3.5.3 Task 4.3 Develop and Test Training Materials

To support workforce development, the project will develop modular training materials tailored to different professional audiences. Training modules will include:

- Introduction to multimodal AI in oral history contexts
- Risk assessment and ethical deployment practices
- Local model installation and configuration guidance
- Interpreting evaluation metrics
- Designing human-in-the-loop workflows

Materials will include Slide decks, Recorded demonstrations, Hands-on exercise guides, and Case-based scenarios. Training materials will be pilot tested with: LIS graduate students, Partner institution staff, and Workshop participants. Feedback will be used to refine clarity and usability.

3.5.4 Task 4.4 Review all Materials to be Released

Before dissemination, all materials—including guidelines, benchmark documentation, and training resources—will undergo a structured review process to ensure: (1) Accuracy and technical validity, (2) Alignment with professional

standards, (3) Accessibility and readability, (4) Compliance with privacy and licensing agreements, and (5) Clear distinction between recommended practices and experimental findings. Advisory board members will review final materials to confirm alignment with archival ethics and oral history practice. All publicly released materials will include clear documentation of scope, limitations, and responsible-use recommendations.

3.5.5 Task 4.5 Disseminate of the Guidelines, Benchmarks and Training Materials

To maximize national impact, dissemination will occur through multiple channels including professional conferences, webinars and workshops, open-access publication, institutional and GitHub repository, which is detailed in project results section. The project website will serve as a central hub for accessing all outputs.

3.6 Summary of Research Goals and Outcomes

The ORAL-AI project follows a structured, evidence-driven progression from evaluation to implementation to national dissemination. Each research objective builds on the previous one to ensure that technological innovation is grounded in empirical assessment, aligned with professional standards, and validated through real-world use. The research goals and outcomes of each objective are summarized in Table 2.

Table 2: Synthesis of Research Objectives, Activities, and Outcomes

Research Objective	Key Activities	Outcomes	Measurements	IMLS Alignment
RO1: Evaluate Multimodal LLM Capabilities and Risks	Develop 30 cross-collection benchmark queries; evaluate models; assess hallucination, bias, and privacy risks; conduct human validation	Benchmark dataset; limitations report; comparative performance matrix; privacy analysis	Quantified hallucination rates; cross-modal retrieval precision; documented bias disparities; validated evaluation protocol	Goal 3.2: Evidence-based digital access
RO2: Design and Implement ORAL-AI Framework	Develop cross-modal alignment pipeline; integrate evidence-grounded RAG; implement human-in-the-loop review; deploy open-source models locally	ORAL-AI prototype; privacy-aware deployment blueprint; review dashboard; governance guidelines	Reduced hallucination rate; improved retrieval precision; successful local deployment	Goal 3.2: Innovative digital collection access
RO3: Evaluate Usability and Public Engagement	Conduct comparative usability studies; survey trust and interpretive fidelity; interview librarians and public users; iterative refinement	User evaluation report; validated engagement metrics; system revisions	Reduced task completion time; increased evidence identification accuracy; improved trust ratings	Goal 2.1: Promote broad public engagement
RO4: Develop Transferable Guidelines and Training Resources	Write best-practice guidelines; release benchmark; create and test training modules; disseminate nationally	National guidelines; open repository; training curriculum; conference presentations	Adoption toolkit available; workshop participation; national dissemination	Goals 2.1 & 3.2: Workforce capacity & expanded access

3.7 Project Team and Resources

PI Dr. Haihua Chen is Assistant Professor of Data Science and Director of the Intelligent Data Engineering and Analytics Lab at University of North Texas. His research focuses on artificial intelligence, multimodal large language models, information retrieval, and data-centric AI for digital libraries and cultural heritage collections. He has published over 60 peer-reviewed articles in leading venues such as EMNLP, WWW, ACM MM, ICDM, IP&M, ASIS&T, and JCDL. Dr. Chen has conducted empirical studies applying NLP and multimodal learning to oral history archives. He will lead overall project coordination and technical direction, including benchmark design, multimodal LLM evaluation, cross-modal alignment development, evidence-grounded generation strategies, and system implementation.

Co-PI Dr. Jiangping Chen is Professor of Information Sciences at University of Illinois Urbana Champaign. Her expertise spans digital libraries, multilingual information retrieval, information services, and large language model

applications. She has led multiple IMLS National Leadership Grants focused on digital collection access and multilingual information services. In ORAL-AI, she will lead user-centered evaluation design, development of professional guidelines and training materials, and alignment with archival standards. She will also co-lead national dissemination efforts.

Advisory Board. An interdisciplinary advisory board of nationally recognized experts in multimodal AI, oral history, digital libraries, archival practice, and community-based collections will provide strategic and professional guidance throughout the project. Confirmed members include **Dr. Yapeng Tian** (Director, Computer Vision and Multimodal Computing Lab, The University of Texas at Dallas), **Dr. Douglas Boyd** (Director, Louie B. Nunn Center for Oral History; former President of the Oral History Association; University of Kentucky), **Geoff B. Froh** (Deputy Director and CIO, Densho), **Dr. Todd Moye** (Director, Oral History Program, University of North Texas), **Dr. Colin Rhinesmith** (Director, Digital Equity Action Research Lab, University of Illinois Urbana Champaign), and **Steven Kent Sielaff** (Associate Director, Oral History Association, Baylor University). Two-page CVs and formal letters of commitment from advisory board members are included in the project materials. The board will meet quarterly to review benchmark development, system implementation, usability findings, and dissemination outputs, ensuring alignment with professional standards, ethical stewardship, and practical library implementation.

Computing Resources. Dr. Haihua Chen has full administrative access to UNT Department of Data Science's high-performance GPU infrastructure, which provides substantial capacity for large-scale multimodal LLM development and evaluation required for the ORAL-AI project. The core resources include eight R760xa servers (each with 2× NVIDIA H100 GPUs), one XE9680 server with 8× NVIDIA H200 GPUs, and one R760XA server with 4x L40S GPU, enabling distributed training, fine-tuning, and large-scale inference for multimodal oral history collections. These systems are supported by approximately 70TB of high-speed SSD storage for active model development and 300TB of HDD storage for large oral digital datasets. These secured, institutionally supported resources are fully operational and sufficient to complete all computational components of the proposed project.

4 Project Results

ORAL-AI will deliver both methodological advancement and nationally transferable infrastructure for responsible multimodal AI in oral digital collections. At the research level, the project establishes a rigorous, replicable evaluation framework integrating benchmark testing, model comparison, bias auditing, privacy assessment, and user-centered validation. By triangulating technical metrics with human interpretive review and workflow analysis, the project advances evidence-based standards for multimodal systems in culturally sensitive archival contexts. The resulting benchmark, scoring rubric, and evaluation protocol will provide one of the first structured methodologies tailored specifically to cross-modal retrieval and narrative summarization in oral history collections, contributing to digital library research and responsible AI governance.

Technically, ORAL-AI will produce a validated, locally deployable multimodal framework integrating cross-modal alignment, retrieval-augmented generation, structured evidence citation, and human-in-the-loop review. The system is designed to measurably reduce hallucination, improve grounding accuracy, and strengthen interpretive transparency compared to traditional keyword-based access. By embedding privacy-aware deployment and stewardship safeguards directly into system architecture, the project offers a sustainable model that libraries of varying scale can adopt without relinquishing institutional control. Comparative usability testing across three nationally significant collections will generate empirical evidence on improvements in retrieval precision, task efficiency, narrative comprehension, and user trust, informing broader digital collection access strategies.

To promote adoption, the project will produce modular training materials, implementation guidelines, and open benchmark resources for librarians, archivists, and educators. Dissemination will occur through peer-reviewed publications, national conferences (e.g., ALA, ASIS&T, JCDL, SAA, OHA), professional webinars, and open-access repositories, supported by a project website serving as a central access hub.

Beyond immediate outputs, ORAL-AI establishes a scalable research agenda for multimodal AI in cultural heritage institutions, positioning libraries as leaders in shaping ethical, transparent, and sustainable AI practices in the stewardship of cultural memory.

Schedule of Completion

The proposed project will begin September 1, 2026, and conclude August 31, 2029

Activity Year 1	Sep 2026	Oct 2026	Nov 2026	Dec 2026	Jan 2027	Feb 2027	Mar 2027	Apr 2027	May 2027	Jun 2027	Jul 2027	Aug 2027
Project kickoff	x											
Project advisory board meeting 1-1		x										
RO 1: Systematically Assess the Capabilities and Limitations of Multimodal LLMs for Oral Digital Collections												
Task 1.1 Develop Queries and Design Benchmarks		x	x									
Task 1.2 Conduct Baseline Evaluation of Commercial and Open MLLMs			x	x	x							
Task 1.3 Synthesize Limitations of Current Systems						x						
Project advisory board meeting 1-2					x							
Task 1.4 Generate Design Requirements							x	x				
RO 2: Research Objective 2: Design and Implement a Responsible, Privacy-Aware Multimodal Framework (ORAL-AI)												
Task 2.1 Develop Cross-modal Alignment and Retrieval Mechanisms Tailored to Oral Histories									x	x	x	
Task 2.2 Integrate Evidence-linking Strategies to Reduce Hallucination										x	x	
Project advisory board meeting 1-3								x				
Task 2.3 Implement Human-in-the-loop Review Workflows											x	x
Project advisory board meeting 1-4											x	

Activity Year 2	Sep 2027	Oct 2027	Nov 2027	Dec 2027	Jan 2028	Feb 2028	Mar 2028	Apr 2028	May 2028	Jun 2028	Jul 2028	Aug 2028
Project advisory board meeting 2 -1		x										
RO 2: (Continued)												
Task 2.4 Deploy Open-source Multimodal LLMs	x	x										
Task 2.5 Incorporate Professional Standards of Archival Stewardship into System Architecture		x	x									
RO 3: Evaluate Usability, Interpretive Fidelity, and Impact on Public Engagement												
Task 3.1 Develop and Test Training Materials and Survey Instruments	x	x	x									
Project advisory board meeting 2 -2					x							
Task 3.2 Recruit Evaluators			x									
Task 3.3 Conduct Usability Testing				x	x	x						
Task 3.4 Interview Evaluators							x	x				
Project advisory board meeting 2 -3								x				
Task 3.5 Analyze Testing and Interview Results									x			
Task 3.6 Iteratively Revise the System based on Evaluation Results										x	x	x
Project advisory board meeting 2 -4											x	

Activity Year 3	Sep 2028	Oct 2028	Nov 2028	Dec 2028	Jan 2029	Feb 2029	Mar 2029	Apr 2029	May 2029	Jun 2029	Jul 2029	Aug 2029
Task 3.6 Iteratively Revise the System based on Evaluation Results (continued)	x	x										
Project advisory board meeting 3-1		x										
RO 4: Develop Transferable Guidelines, Benchmarks, and Training Resources for Libraries												
Task 4.1 Write and Review Guidelines			x	x								
Task 4.2 Organize and Document Benchmarks				x	x							
Project advisory board meeting 3-2					x							
Task 4.3 Develop and Test Training Materials					x	x						
Task 4.4 Review all Materials to be Released							x	x				
Project advisory board meeting 3-3								x				
Task 4.5 Disseminate of the Guidelines, Benchmarks and Training Materials									x	x		
Project Wrap up and Promotion										x	x	x
Project advisory board meeting 3-4											x	