Gauging Library Needs for Advanced AI-Assisted Cataloging

1. Project Justification

1.1 Vision and Project Goal and Objective Aligned with IMLS NLGL

The current era of unprecedented information proliferation and increasing multilingual diversity challenges libraries' traditional cataloging and resource management processes. Cutting-edge artificial intelligence (AI) tools known as large language models (LLMs), which excel at processing natural language, have the potential to assist librarians in their quest to organize and provide access to their ever-growing collections. By combining the capability of AI with the expertise of catalogers, we aim to create a synergy that will empower catalogers to be as efficient and accurate as possible as they enhance the accessibility and inclusivity of library resources.

The University of North Texas (UNT) Department of Information Science, as the project leader, in collaboration with the UNT libraries, requests ($132,759) from the **National Leadership Grants for Libraries** (NLGL) for a **2-year Applied Research** project to investigate the applicability of LLMs running locally to assist the subject cataloging of digital and print resources. We plan to address two main questions. **RQ1**: How can LLM-based models be developed to generate accurate cataloging results, particularly classification and subject analysis, for both English and foreign language resources? **RQ2**: How can AI models be integrated into cataloging procedures to assist librarians? This project aims to build knowledge for the future development and deployment of LLM-based applications for cataloging, which aligns with NLGL Goal 3.2 (support innovative approaches to digital collection management and improve cataloging and inventory practices).

1.2 Statement of Need

Preliminary interviews the PI conducted at the UNT Libraries revealed that their catalogers had already been considering how AI-assisted cataloging might help their daily work.

1) *Facilitating cataloging, classification, and subject analysis for locally produced born-digital materials such as Electronic Theses and Dissertations (ETDs).* UNT produces 400 to 500 ETDs each year, and the UNT Libraries' Cataloging and Metadata Services staff is responsible for cataloging them. ETDs may use technical terminology or unfamiliar languages; AI-assisted cataloging may help them process these resources more effectively and efficiently.
2) *Enhancing legacy records with new authorized terminology from evolving standards*. Over the last two decades, cataloging standards have been evolving to allow greater and more expressive means of description. Alongside the most recent rules for Resource Description and Access (RDA), for instance, new controlled vocabularies have emerged, designed to capture more pertinent attributes of cataloged materials in ever greater detail. However, these vocabularies cannot improve discovery if they cannot be applied to the whole collection consistently. Updating legacy records with such enhancements is not usually a priority, even though it would have a major positive impact on resource discovery throughout the catalog. LLMs' ability to summarize and generate predictive text could be useful for suggesting applicable terms to help catalogers with this daunting task.
3) *Identifying and updating biased or outdated terminology that may appear in catalog records*. Acceptable terminology changes over time, and librarians strive to ensure that the library catalog is as inclusive and welcoming to its users as possible without hindering resource discoverability. Although librarians can (and do) respond to patron complaints about offensive terms appearing in records, being able to identify such instances ahead of time would help them be more proactive in addressing them. LLMs may be able to process large numbers of records to identify problematic language.
4) *Facilitating cataloging of foreign-language resources*. Libraries' collections include resources in many different languages – for instance, the UNT Libraries' catalog contains >106,000 German resources, >103,000 French resources, >50,000 Spanish resources, and thousands more resources in Latin, Italian, Russian, and hundreds of other languages. Cataloging an object in a language that catalogers do not understand is, of course, difficult, whether they are doing original cataloging or adapting a record for local needs from a source like OCLC. AI-assisted cataloging may assist them in understanding and describing such resources, thereby enhancing the catalog's accessibility and inclusivity.

These needs are not unique to the UNT Libraries. Issues around retrospectively updating legacy records (ALCTS/CaMMS Subject Analysis Committee, Working Group on Full Implementation of Library of Congress Faceted Vocabularies, 2017; Flannery, 2015) and dealing with foreign language resources (ALCTS Task Force on Non-English Access, 2007) are longstanding and widespread. Although researchers have explored automatic generation of metadata for knowledge objects – which may be used to help alleviate these issues – since the early 2000s (Wang, 2009), little of this research has transferred to real-world use in library cataloging. Despite continuous improvement in the accuracy of automatic classification and keyword extraction, it seldom reaches a level of precision that can replace human judgment. Furthermore, there is a lack of related research to determine how automatically generated results might be used by cataloging librarians effectively in practice.

State-of-the-art LLMs have shown superior capability in comprehending human language and assisting in problem-solving compared to traditional machine learning models (Achiam et al., 2023). The key to their success is the fact that they are pre-trained on a large amount of data from diverse domains and do not require task-specific training. Indeed, studies have found that using LLMs to make inferences without conducting additional training – known as "zero-shot" learning (GPT-4 from ChatGPT) – can achieve better results than traditionally trained models (Zhao et al., 2023). However, LLMs *can* be further optimized to improve domain-specific performance with additional training – a process often called "finetuning." For example, an LLM finetuned using a training set of book abstracts and classification codes can achieve better accuracy compared to zero-shot when asked to generate classification codes based on new book abstracts. This project will attempt to harness LLMs' potential, creating LLM-based models using both zero-shot learning and finetuning with library data. We will then explore librarians' perspectives about the performance and utility of our models. We hope to push the boundaries of automated cataloging and create foundational knowledge for further development and implementation of useful AI-assisted cataloging tools.

1.3 Existing Work and Research Gaps

In previous works, various AI methods have been investigated for their potential to aid in the comprehension, summarization, and categorization of textual content in cataloging.

1.3.1 Lack of Studies on the Use of LLMs for Cataloging

While research has been conducted on AI models for cataloging tasks, there is little research that specifically explores using state-of-the-art LLMs in this context. Researchers have, for example, explored machine learning (ML) models to classify digitalized texts into the Universal Decimal Classification (UDC) system (Kragelj and Kljajic, 2021), built deep learning (DL) models to classify patent literature using the International Patent Classification system (Lyu and Han, 2019), and developed ML models to classify books based on the semantic space transformation (Yu, 2020). Although these models have achieved good performance, they have not been accurate enough to have been transformed into real-world use. LLMs, such as GPT-4 (Achiam et al., 2023) and Llama (Touvron et al., 2023), have shown remarkable improvement and exhibit human-level performance on many natural language processing tasks – showing promise for cataloging. However, little work has been done to understand how LLMs can be developed specifically for the cataloging task, which is one goal of our project.

Most of the existing research in the field of library and information science has primarily concentrated on cataloging resources in one's predominant language – e.g., cataloging English resources in the US and Swedish resources by Swedish researchers (Goub et al., 2020). Abdelrahman and Fox (2022) conducted a classification study on Arabic ETDs with deep learning models, but the models were trained with Arabic data. Limited research has been done on how to catalog resources automatically in languages other than the predominant language(s), especially "low-resource languages" – ones that do not have enough samples for model training. LLMs have shown good performance in NLP tasks in cross-lingual settings (Achiam et al., 2023; Zhao et al., 2023). This inspires us to further explore the capability of LLMs in cataloging materials in secondary or low-resource languages.

1.3.2 Lack of Studies on the Applicability of AI-assisted Cataloging

Machine learning and deep learning models have been developed to automate the classification of digitized resources in libraries. However, few have been transformed into practical applications. Little research has been done to understand *how to effectively incorporate LLMs into cataloging librarians' workflows*. Cataloging librarians' perceptions about the applicability of the results will be important for developing and deploying LLM-based applications in real-world settings. The proposed study will conduct surveys and interviews with librarians to explore these questions. Study results will inform future research into the practical applications of advanced AI in assisting library cataloging and digital resource management.

2 Overview of Research Objectives and Expected Outcomes

We propose a two-phased project that includes (1) developing LLM models for cataloging tasks and (2) evaluating the models' applicability to library cataloging. In **Phase 1**, we will conduct experiments to answer *RQ1: How can LLM-based models be effectively developed to generate accurate cataloging results*, including *the development of classification models* (**Task 1.1**), *the development of subject analysis models* (**Task 1.2**), *and the adaptation of models to foreign language resources* (**Task 1.3**). From these experiments we will gain a comprehensive understanding of how the LLM-based models performed in the given tasks along with the required computing resources and response times.

In **Phase 2**, we will provide the findings from Phase 1 to cataloging librarians who will then help us answer *RQ2: How can AI models be effectively integrated into cataloging procedures*. We will first *investigate factors related to librarians' intention to use AI-assisted cataloging* (**Task 2.1**) with surveys to gauge librarians' needs. This will build knowledge to assess when AI-assisted cataloging tools might be ready to deploy. Then we will *examine how to integrate AI models into cataloging workflows* (**Task 2.2**) using participatory design. This task will elicit concrete ways cataloging librarians believe they can integrate AI models into their workflows. The table below summarizes research tasks in each phase, methods to be used, and expected results.

<table>
<tr><th>Phase</th><th>Research Tasks</th><th>Methods</th><th>Results</th></tr>
<tr><td rowspan="3">1</td><td>Task 1.1: Develop LLM-based classification models for library resources.</td><td rowspan="3">Prepare data for classification and subject analysis tasks in English and foreign languages using MARC records and Library of Congress standards as the external knowledge base.

Train and evaluate LLM-based models using various experimental settings.</td><td rowspan="3">Code to train LLM-based models for classification and subject analysis tasks

Wrapped up models that can be used for classification and subject analysis of library resources

Publications in venues focused on digital libraries and computational linguistics</td></tr>
<tr><td>Task 1.2: Develop LLM-based subject analysis models for library resources.</td></tr>
<tr><td>Task 1.3: Explore the adaptability of LLM-based models on cataloging tasks for foreign language resources.</td></tr>
<tr><td rowspan="2">2</td><td>Task 2.1: Investigate determinant factors influencing librarians' intention to use AI models in cataloging.</td><td>Design survey questions based on the performance of LLM-based models for cataloging tasks.

Survey cataloging librarians to collect data.

Conduct factor analysis and structural equation modeling to investigate the association of factors with the intention to adopt AI-assisted cataloging.</td><td>A survey questionnaire to assess librarians' intention to use AI-assisted cataloging and determinant factors

A report summarizing librarians' intentions and significant factors on the intention to use

Publications in venues with focuses on digital libraries and human computer interactions</td></tr>
<tr><td>Task 2.2: Investigate how to integrate AI models into librarians' cataloging workflows.</td><td>Use participatory application design with cataloging librarians based on the performance of LLM-based models in cataloging tasks.</td><td>A report summarizing librarians' perceived ways of integrating AI in cataloging procedures</td></tr>
</table>

3 Preliminary Work by the Team

The team has done preliminary work on classifying books using Llama2 and ChatGLM, which are open-sourced LLMs that can be finetuned locally. We used the Parameter-Efficient Fine-Tuning (PEFT) Low-Rank Adaptation

(LoRA) method (Hu et al., 2021) to finetune the models, training them to use the title and abstract from MARC records to predict the first-level class (e.g., H: Social sciences), the classification code (e.g., HB501), the first-level class and code (e.g., H: Social sciences and HB501), and multi-level class and code (e.g., HB Capital. Capitalism and HB501). Our proposed Phase 1 experiments can easily adapt the code framework from this earlier project.

These initial experiments show that LLM-based models perform better than existing deep and machine learning models. We first used 2,200 samples, about 100 in each first-level class, to train and finetune models. Figure 1(a) demonstrates that finetuned ChatGLM3 and Llama2 have much better accuracy than SVM and Naive Bayes and slightly better than BERT-based models. Figure 1(b) shows that the classification accuracy for multi-level class is 0.786 for finetuned ChatGLM and 0.783 for finetuned Llama2, lower than other tasks. It shows that when the classification task involves more categories, the performance of finetuned LLMs will slightly decrease. This preliminary study broadly demonstrates that the performance of finetuned LLMs is better than traditional models, may generate results sufficient for practical use, and could offer a pragmatic and readily deployable framework for enhancing cataloging processes.

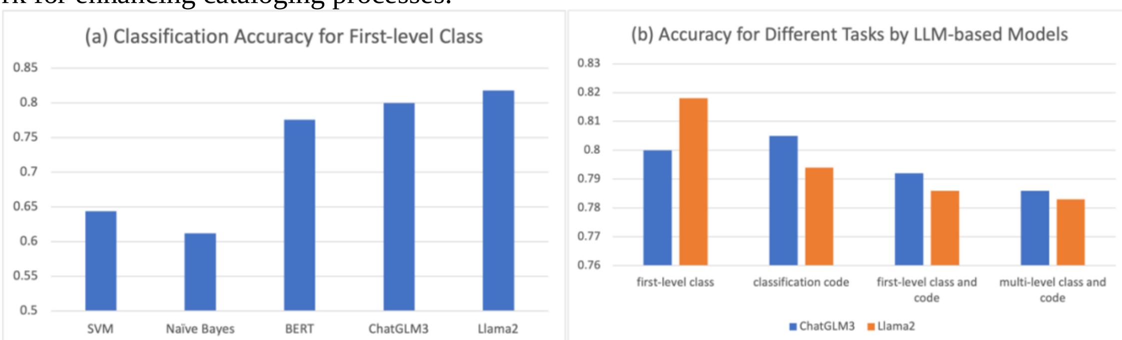

Figure 1. Accuracy of traditional models and LLM-based models on book classification tasks.

4 Project Work Plan

This section details the tasks for each phase. Phase 1 will focus on developing and evaluating LLM-based models for cataloging tasks. Phase 2 will conduct user studies based on the experimental results obtained from Phase 1.

4.1 Development of LLM-based models for Cataloging Tasks (Phase 1)

We have broken Phase 1 into three tasks, each of which involves similar methods for developing and evaluating a set of LLM-based models to accomplish a particular data generation task: classification, subject analysis, and the adaptation for resources in foreign languages. Because these have varying levels of complexity, each requires its own experimental settings for configuring and training the models. Data preparation and model training methods common to all tasks are introduced below, followed by details about each of the three tasks.

Data Preparation Co-PI Thomale has developed tools we will use to extract data from MARC records in the UNT Libraries' Integrated Library System (ILS) and harvest qualified Dublin Core records from the UNT Digital Library. These tools will allow us to extract any necessary metadata, such as title, abstract, description, subject, classification code, language, publication date, etc., as well as the full text for digital resources.

LLM-based Model Training We will experiment with several open-source LLMs that claim to have state-of-the art performance in natural language processing tasks, including LLaMa 2 by Meta, Falcon by the Technology Innovation Institute, and MPT by Mosaic ML. GPT-4 from Open AI will not be used as it is not open source and cannot be locally trained. The GPT-4 training process requires data to be uploaded to the system, which is not appropriate for library data due to data privacy and potential copyright concerns. We will explore Guidance, a programming paradigm that allows using regular expressions to constrain zero-shot learning output for LLMs. The listed models have been integrated into the Hugging Face ecosystem, which allows us to download local copies for model training and saving. We will explore various finetuning methods, including the Parameter-Efficient Fine-Tuning (PEFT) Low-Rank Adaptation (LoRA) and Transformer Reinforcement Learning (TRL).

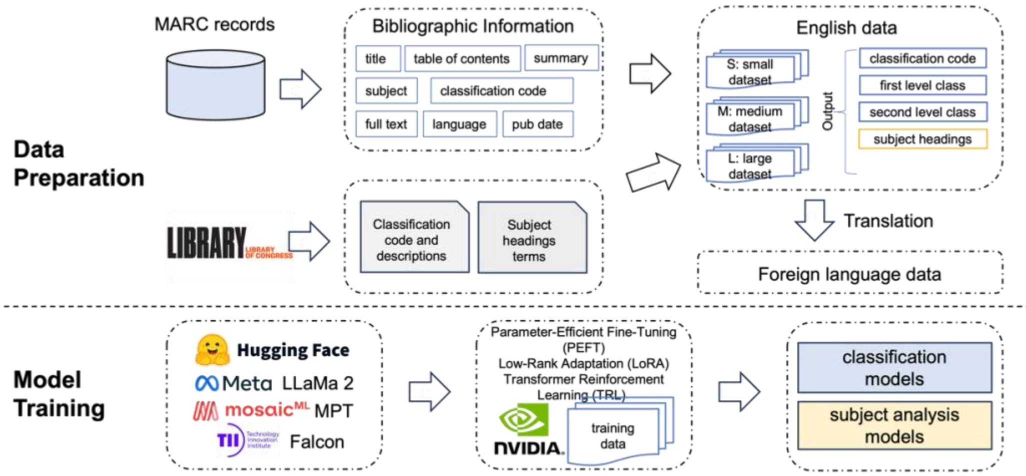

Figure 2: Overview of data preparation and model training.

4.1.1 Develop LLM-based Classification Models (Task 1.1)

How catalogers utilize AI-generated classification results is contingent upon both the format of these results and their accuracy. To glean sufficient contextual information that will help with this assessment, we propose conducting subtasks in which we generate outputs that have different **levels of granularity** for different **types of resources**. An example subtask would be to train a model that can predict multi-level classes for ETDs. For each subtask, we will conduct experiment iterations to help us identify the optimal combinations of training variables that yield the most effective outcomes. Specifically, we will explore the **input**, the **size of training data**, the **inclusion of an external knowledge base**, and the choice of **LLM** in terms of how each impacts performance.

Table 1: Configuration of model training factors and subtasks for classification

Variables	Model Training Factors				Subtasks	
	Information Input	Training Data Size	Knowledge Base	LLM	Types of Resources	Granularity level (example)
Conditions	Title Table of Contents Description Summary PublishYear	0 10 100 500 1000	Yes No	LLaMA 2 Falcon MPT	ETD Book	First-level class (H: Social sciences) Second-level class (HB: Economic theory) Multi-level class (HB Capital. Capitalism) Classification code (HB501)

Table 1 shows example configurations for training-related factors and variables. The *Information Input* will be a combination of the listed fields based on data availability for the classification task; e.g., Title + Summary + PublishYear. *Training Data Size* shows the approximate number of samples per first-level class. We will start with sample size 0 to test the performance with zero-shot learning. The *Knowledge Base* will be the Library of Congress classification code and description, which can be used to add information explaining classification categories and numbers in the training data. This is to explore whether the inclusion of knowledge base information will enhance accuracy. For *LLM*, Table 1 only lists the model family, but each model family also includes multiple model variations. For example, LLaMA 2 includes models such as Llama-2-7b, Llama-2-70b-chat. Additionally, not shown in the table, the hyperparameters of models, such as the learning rate, batch size, and number of epochs, will be finetuned in the model training process.

Each trained model from each subtask will be evaluated based on how well it classifies test data. We will use a combination of computing-based evaluation and human evaluation to assess and validate the accuracy of classification.

- *Computing-based metrics*. We will employ a comprehensive set of performance metrics to assess the classification models: a precision, recall, and F1 score for each individual class and a weighted precision, recall, and F1 score for all classes. This will produce both a granular evaluation of how well a model performs on each category and a holistic measurement we can use to compare different models. We will report accuracy to assess the models' classification correctness.

- *Human evaluation*. Human evaluation will further validate the first-level, second-level, and multi-level class assignments. Because some information resources may well be described using categories *other* than the classification code in the MARC record, human evaluation will tell us whether the classes our tools assign are appropriate and how much the assigned classification diverges from the inherent information contained within the resource. We will hire two Library Science graduate students who will help us evaluate a random selection of 50 samples from each of the models that has the best computing-based metrics in each subtask. The students will annotate results by completing a 5-point Likert scale question about the suitableness of class predictions for each sample based on the sample's complete bibliographic information. We will analyze the annotated results for further insights into the models' accuracy.

4.1.2 Develop LLM-based Subject Analysis Models (Task 1.2)

In Task 1.2, we propose subject analysis subtasks in which we generate outputs with various **numbers of subject terms** for different **types of resources**. Experimenting with different numbers of generated subject terms should help librarians assess whether our models might generate valid and appropriate Library of Congress Subject Headings (LCSH) on their own or are better used to generate candidate terms or headings for cataloging librarians to choose from.

Table 2 shows example configurations for training-related factors for Task 1.2. Factors again include the choice of **input**, **size of training data**, **knowledge base**, and **LLM**. The *Knowledge Base* in this task refers to the LCSH. Without a knowledge base, LLM-based models may generate words outside of LCSH, even when the model is finetuned with metadata that has LCSH terms. Using LCSH as the knowledge base, we can adopt the constrained generation method developed by Kumar et al. (2021) to ensure the outputs only use those terms.

Table 2: Configuration of model training factors and subtasks for subject analysis

Variables	Model Training Factors				Subtasks	
	Information Input	Training Data Size	Knowledge Base	LLM	Types of Resources	Number of Subject Terms
Conditions	Title Table of Contents Description Summary PublishYear	0 10 100 500 1000	Yes No	LLaMA 2 Falcon MPT	ETD Book	3 5 10 no limit

Trained models will be evaluated on the relevance and diversity of subject terms they generate. Like in Task 1.1, we will combine computing-based evaluation metrics and human evaluations.

- *Computing-based metrics*. We will evaluate the relevance of generated terms by comparing them with the reference set of terms in subject headings and measure the amount of overlap. We will assess their comprehensiveness by calculating the proportion of subject terms that are captured by the generated list (recall). The generated terms may not be identical to the terms in the subject headings from the cataloging records even though they may be used to describe the content of the resource. We will further calculate the semantic relevance of generated terms and reference terms as an additional assessment of relevance.
- *Human evaluation*. We will solicit feedback from human evaluators to assess the overall quality and utility of the generated terms. We will hire two Library Science graduate students who will help us evaluate a random selection of 50 samples from each of the models that has the best computing-based metrics in each subtask. The students will rank the coverage, specificity, and accuracy of generated terms using a 5-point Likert scale based on the sample's complete bibliographic information. We will then analyze the annotated results for further insights into the performance of subject analysis.

4.1.3 Adaptability of Classification and Subject Analysis Models to Foreign Language Resources (Task 1.3)

Given a context where English is the predominant language, describing a non-English item using both English and the item's own language makes it accessible to most users regardless of their language proficiency. For a resource described using only a foreign language, we would also want to be able to describe it in English following the appropriate English-language classification or subject standards; e.g., LCSH. Language models can be developed to generate such English classification and subject analysis terms for non-English resources. But foreign-language resources often have limited sample sizes that may be insufficient for directly training an accurate model. One viable approach might be to leverage the resources in a library collection's dominant language to train multilingual large models so the models can be applied to resources in other languages. Many LLMs have been pre-trained with data in multiple languages, such as English, Spanish, French, and Chinese, enabling them to comprehend and process text in various languages simultaneously (Winata et al., 2021). Task 1.3 will study how well the LLM-based models developed in Tasks 1.1 and 1.2 will perform English-language classification and subject analysis for foreign language resources.

Figure 3 shows this task's experimental setting. First, we will take the best-performing model developed in each subtask from 1.1 and 1.2, trained with the metadata of English resources, and we will use it to make inferences about non-English test data. This is to evaluate whether the LLM-based models trained on a predominant language, such as English, can be used effectively for resources in a different language. Then we will experiment with LLMs finetuned using metadata of non-English resources and evaluate the performance of models on the test data in the same language, e.g., Spanish. Non-English samples may be too few for effective model training, so we will also experiment with a data augmentation strategy in which we create synthetic non-English data, translated from English, and add it to the training data.

Finally, we will compare the performance of these models using the same evaluation methods in Tasks 1.1 and 1.2. We will identify the specific non-English languages later for experiments based on the need and availability of metadata in the UNT Libraries. Student assistants in foreign languages will be hired together with LIS students on the human evaluation work.

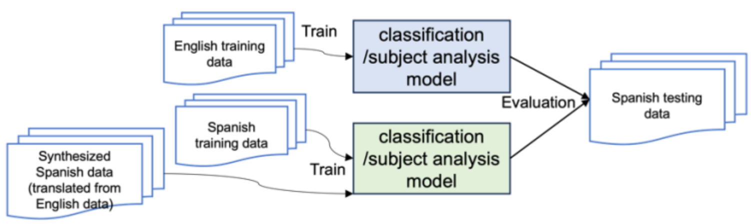

Figure 3. Experimental setting to explore LLM-based models to process foreign language resources, using Spanish data as an example.

4.2 Investigate the Effective Integration of AI Models in Assisting Cataloging Tasks (Phase 2)
Phase 2 will explore cataloging librarians' attitudes and perceptions about AI-assisted cataloging, using results from Phase 1 to elicit concrete ideas about how to build hypothetical AI-assisted cataloging systems.

4.2.1 Investigate Librarians' Intention to Use and Determinant Factors (Task 2.1)
Our first Phase 2 task (2.1) will be to conduct a survey study based on the Unified Theory of Acceptance and Use of Technology (UTAUT) model (Venkatesh et al., 2012) designed to reveal factors influencing librarians' intention to use AI-assisted tools and how those factors interrelate. This will serve to help libraries determine whether there is a need to initiate efforts in the development and deployment of AI models to assist cataloging work. The study plans to survey about 200 cataloging librarians from libraries of different types and sizes to ensure that insights are inclusive. Participants will be recruited through the advisory board members' connections as well as Listservs, such as *AUTOCAT* and *Code4lib*.

The survey questions will evaluate librarians' intention to use AI-assisted cataloging systems and determinant factors, including "performance expectancy," "facilitating conditions," "social influence," and "perceived risks." Note that we plan to exclude "effort expectancy" – since the system is assumed to allow anyone who can type and search to use its basic functions, so it will require little additional training. Like many other studies (Velverthi et al., 2023; Widyanto et al., 2022), we are adding "perceived risks" to the traditional UTAUT model, according to our preliminary interviews with UNT cataloging librarians. This factor measures librarians' perceptions of risks associated with the adoption of AI-assisted tools. Figure 4 presents our conceptual model for the survey design. It will inform our survey question design, with questions measuring the four determinant factors and librarians' intention to use AI-assisted tools. (See Attachment for example survey questions.)

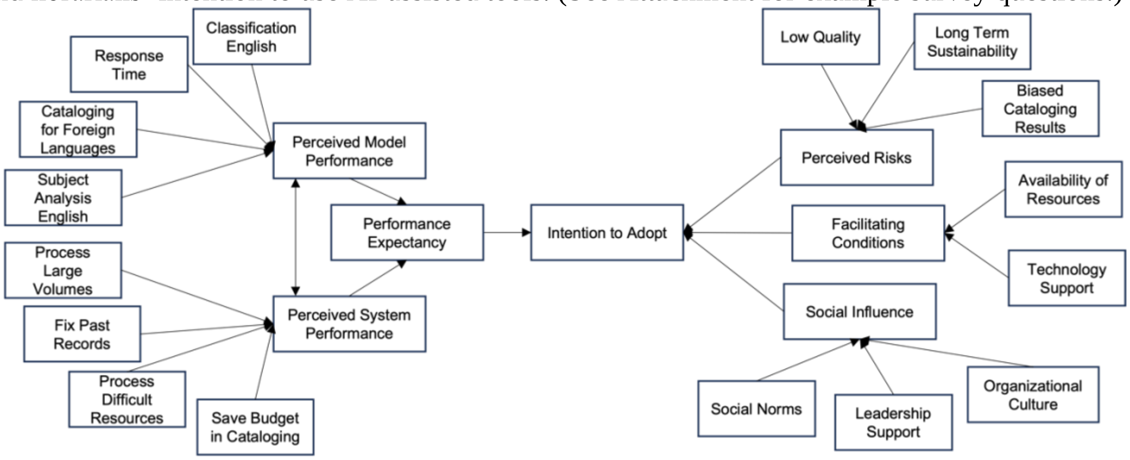

Figure 4. Conceptual model to assess factors related to the intention to adopt AI-assisted cataloging systems.

"Performance expectancy" measures individuals' perceptions about whether and how a system might help improve cataloging. We plan to investigate from two perspectives. (i) The perceived performance of our AI-based models in the earlier classification and subject analysis tasks. (ii) The perceived utility of possible AI-assisted systems for library cataloging work. To aid in evaluating perceived performance, we will provide survey participants the computing-base evaluation metrics, human-evaluation results, sample outputs, and response times from Tasks 1.1, 1.2, and 1.3. Perceived performance will help contextualize librarians' assessment about the utility of hypothetical AI-assisted systems. We will investigate the following aspects of perceived utility. (i) The volume of new resources entering library collections creates pressure on cataloging departments. Might AI-assisted systems help ease this pressure? (ii) Library catalogs commonly have backlogs of legacy records that are not optimally discoverable because they do not take advantage of newer standards and controlled vocabularies. Might AI-assisted systems help tackle these backlogs to enhance legacy records and make them more discoverable? (iii) Cataloging librarians often encounter specialized subject matter, technical terminology, lengthy descriptions, and unfamiliar languages. Might AI-assisted systems facilitate their understanding of these resources to aid in cataloging them more effectively? (iv) Libraries often face resource-allocation and budgetary constraints. To what extent might AI-assisted systems help libraries and catalogers cope with these constraints?

"Perceived risks" are concerns about the potential negative impact of AI-assisted cataloging systems that would inhibit adoption. We plan to investigate three potential risks. (i) AI output may be inherently biased. (ii) AI output may be perceived as so low quality that it would take more effort to make it suitable to use than simply cataloging without AI assistance. (iii) AI-assisted cataloging may pose various short-term and long-term threats to the cataloging profession – e.g., in the short term it may contribute to the view that catalogers are expendable and fuel further cuts to libraries' cataloging budgets, and in the long term it may be used to replace catalogers.

"Facilitating conditions" include the perceived availability of infrastructure and resources to support the effective implementation and use of the technology. We plan to investigate the levels of physical infrastructure and expert technology support in librarians' workplaces. "Social influence" includes the social norms, leadership support, and organizational cultures that shape librarians' perceptions about how useful AI-assisted cataloging systems might be.

The survey will include open-ended questions for participants to suggest additional factors that might encourage or inhibit their intention to adopt AI-assisted cataloging. It will also collect librarians' demographic and work experience-related characteristics, as well as the size and type of the library in which they work. We will attempt to determine what (if any) differences in perceptions occur along these dimensions. There will be an option for participants to leave their emails for the semi-structured study that follows in Task 2.2.

After collecting the data, we will generate descriptive statistics to produce an overall assessment of librarians' perceptions of determinant factors. We will then employ factor analysis and structural equation modeling to gain a quantitative understanding of the determinant factors and their relations to librarians' intention to use hypothetical AI-assisted cataloging systems.

4.2.2 Investigate Librarians' Perceptions about How to Integrate AI models with Cataloging Workflows (Task 2.2)

In Task 2.2, we will design a user study in which we attempt to obtain cataloging librarians' perceptions about how AI might be integrated into their workflows based on the output of our LLM-based models. We will sample 20-25 cataloging librarians from the 2.1 survey participants who expressed interest in being interviewed, and we will conduct a one-on-one, semi-structured interview with each of them. Interview participants will be provided with all relevant information about our best-performing models from Tasks 1.1 to 1.3. We will guide them in developing interface mock-ups and paper prototypes that illustrate how they would envision using this output to assist with the cataloging process, asking them follow-up questions as needed to understand their ideas. We will also present them with different realistic scenarios in which our AI models might be used: batch generation of classification and subject analysis results, an interactive chatbot (like ChatGPT), and autocomplete-style suggestions for metadata fields. Users will then assess the relative strengths and weaknesses of each usage scenario. Participants will be invited to discuss their opinions surrounding the following questions for more insights about the applicability of AI models:

- How could AI-generated outputs (such as what we obtained from Tasks 1.1, 1.2, and 1.3) contribute to the efficiency and accuracy of librarian's cataloging operations?
- How could librarians leverage these outputs to enhance the quality and consistency of metadata?
- How might AI-assisted cataloging contribute to the development of metadata management strategies, policies, and procedures?

Interviews will be recorded and transcribed for qualitative analysis to identify recurring themes and patterns. Through this process, we hope to better understand user needs, expectations, and preferences that could either inform the development of potential AI-assisted cataloging tools or help us understand what further study or development is needed to make *practical* AI-assisted tools.

5 Project Team and Resources

5.1 Project Team

Key Staff:

- The PI, Dr. Lingzi Hong, is an Assistant Professor of Data Science at UNT. Her research focuses on computational linguistics, user behavior modeling, and data literacy education. Dr. Hong will focus on the development of programs that utilize LLMs for classification and subject analysis of library resources, and the design of questions and protocols for surveys and interviews.
- The Co-PI, Jason Thomale, is an Associate Librarian who develops and maintains resource discovery systems at the UNT Libraries. Mr. Thomale will help extract and harvest the data from the UNT Libraries' systems to be used throughout the project, and he will assist in evaluating LLMs' outputs, recruiting participants, and developing and conducting the survey and interviews.

Advisory Board Members:

- Kevin Yanowski, Department Head of the Cataloging and Metadata Services at the University of North Texas Libraries. He will provide necessary support to access the UNT library data, provide feedback about project progress, and help connect with catalogers for user studies.
- Stacey Wolf, Associate Librarian at the Cataloging and Metadata Services at the University of North Texas Libraries. She will assist with data access and the dissemination of research findings.
- Casey Mullin: Head of Cataloging and Metadata Services at Western Washington University Libraries. He will provide feedback about project progress and help connect with catalogers for user studies.
- Charlene Chou: Head of Knowledge Access at the New York University Libraries. She will provide feedback about project progress and help with the dissemination or research outcomes.

- Sarah Hovde: Monographs & Media Cataloging Librarian (Librarian II) at the University of Maryland Libraries. She serves on the OLAC cataloging policy committee and the ALA core subject analysis committee. She will assist with the recruitment of research participants and the dissemination of research outcomes.
- Dr. Jian Wu: Assistant Professor of Computer Science at Old Dominion University. He has been Tech Director of CiteSeerX for five years. He will provide feedback about the development of AI models for cataloging tasks.
- Dr. C. Lee Giles: David Reese Professor, College of Information Sciences and Technology at the Pennsylvania State University. He will provide feedback about project progress and help with dissemination of project outcomes.

5.2 Data and Computing Resources

5.2.1 Data

The UNT Libraries can provide a wide variety of metadata records and digitized full-text resources for project use. The catalog contains >3.7 million MARC records. Of those, >659,000 contain tables of contents in the MARC 505 field, >413,000 contain summaries or abstracts in the MARC 520 field, and >139,000 contain both a table of contents and a summary or abstract. The MARC records are readily available to extract and use. Likewise, the UNT Digital Library contains >1 million objects that include qualified Dublin Core metadata. We have identified a few collections containing full-text materials that might work: >20,000 public Electronic Theses and Dissertations (ETDs), >43,000 Congressional Research Service Reports, and >1,900 End of Term publications (from various United States government agencies). We have developed scripts to harvest the metadata and full text from the UNT Digital Library via their OAI-PMH 2.0 interface.

5.2.2 Computing Resources

The PI Dr. Lingzi Hong has a secured server located in her office. It has an Intel Core i7 processor, 32GB memory, 512 GB hard drive, and 2 Nvidia graphic cards: an RTX 2080 and a GTX Titan XP. She also owns a separate server managed by Computing for Arts and Sciences of UNT. It has an Intel Xeon Gold 6226R processor, 128 GB memory, and 3 Nvidia RTX 8000 graphic cards.

6 Project Results

Introducing automated or semi-automated tools may help improve how the library operates and the services it provides its users. For individual cataloging librarians, these tools may facilitate tasks that were once labor-intensive, rote, or boring, freeing up valuable time for more interesting and intellectually demanding activities: cataloging more challenging resources, curating collections, contributing to the improvement or development of standards, and conducting research. Automated tools may help enhance legacy records and improve the handling of non-English-language materials, thereby improving discovery and ensuring the accessibility of these resources for a broader audience. This, in turn, fosters a more inclusive and multicultural library environment.

We are committed to making all project outputs identified in Table 1 widely available to researchers and practitioners. This includes methodologies and code for our LLM-based models, input and output data, our survey instruments, and reports we create based on the survey and interview results. These will be licensed under the Creative Commons Attribution 4.0 International License (CC BY 4.0) for extensive reuse and distribution. We will deposit the materials to the UNT institutional repository – UNT Scholarly Works – to maintain long-term accessibility. Additionally, we will build a project website to share project updates and outcomes. Publications from this research are aimed at venues with a focus on computational linguistics, digital libraries, library and information science, or human-computer interactions, such as *JCDL, ASIS&T, JASIS&T, Information Processing and Management (IP&M)*, *Library Resources & Technical Services (LRTS)*, *Information Technology and Libraries (ITAL)*, and *Code4lib Journal*. We will also share our results with Listservs frequented by cataloging and systems librarians, such as *AUTOCAT* and *Code4lib*, for further dissemination and solicitation of feedback from pertinent groups.

Schedule of Completion

The proposed project will begin August 1, 2024, and conclude July 31, 2026.

Activity	**Aug 2024**	**Sep 2024**	**Oct 2024**	**Nov 2024**	**Dec 2024**	**Jan 2025**	**Feb 2025**	**Mar 2025**	**Apr 2025**	**May 2025**	**Jun 2025**	**Jul 2025**
Project kickoff												
Project advisory board meeting 1												
Task 1.1-1 Develop classification models												
Task 1.1-2 Evaluate classification models												
Task 1.2-1 Develop subject analysis models												
Task 1.2-2 Evaluate subject analysis models												
Task 1.3-1 Develop models for foreign languages												
Task 1.3-1 Evaluate models for foreign languages												
Project advisory board meeting 2												
Activity	**Aug 2025**	**Sep 2025**	**Oct 2025**	**Nov 2025**	**Dec 2025**	**Jan 2026**	**Feb 2026**	**Mar 2026**	**Apr 2026**	**May 2026**	**Jun 2026**	**Jul 2026**
Task 2.1-1 Design survey questions												
Task 2.2-1 Design interview protocols												
Task 2.1-2 Survey data collection												
Task 2.1-3 Survey data analysis												
Task 2.2-2 Conduct interviews												
Task 2.2-3 Interview data analysis												
Project advisory board meeting 3												
Dissemination												

Digital Products Plan

Types of Digital Products

The digital products generated by this project will include data extracted from MARC records in the UNT Libraries' Integrated Library System and qualified Dublin Core records from the UNT Digital Library. The data collection and usage will adhere strictly to data policies set forth by the UNT Library systems and comply with data privacy and copyright laws and regulations. The project will train LLM-based models for automatic generation of cataloging results. The code created for statistical analysis, visualizations, and deep learning models will be kept and shared to the extent possible. The project will develop prototype tools, which will generate source code, scripts, configuration, data, and documentation files. These files will all be kept.

The project will conduct surveys and interview studies. The original data collected through surveys and interviews will not be shared. Only data at the aggregate level will be shared. For example, reports generated to summarize the survey and interview results will be promoted and shared publicly, particularly within the cataloging librarian communities. This is to ensure the research outcomes will make an impact to the target group. We will undergo a thorough deidentification process to ensure the survey and interview participants will not be identified.

The project will also create digital content, such as social media posts, project team achievements, and project posters for a general audience on the project website. The content shared will strictly comply with ethical guidelines in privacy and security.

Data and Metadata Standards

The project team, in collaboration with the University of North Texas Digital Libraries (UNT DL), will choose file formats that will enable the most effective and secure management over time. Best practices for data curation and digital preservation will govern the choices of file formats. Metadata records will be created to describe each of the project’s digital resources. The UNT DL utilizes a system of metadata based on qualified Dublin Core metadata elements that ensures long-term preservation and viability of the digital objects housed in their system. Metadata records will be retained and available for open access through the UNT DL.

Policies for Access, Sharing, and Privacy

Sharable data will be saved in the UNT DL, which is open access and free. Access restrictions can be placed on all items in a deposit. We adhere to the ethical guidelines on access restrictions, confidentiality, and community and individual wishes on data sharing. Requests can be made to the Principal Investigator for limited access to the data generated by the project team. Data and files not to be shared will be saved on two secure servers, which the PI has full control, and an external hard drive in a locked location.

Policies for Re-use, Re-distribution, and Derivatives

To the extent possible, digital products made by this project will be available to the target group, other researchers, or the public. The reuse and redistribution will strictly adhere to the data policies set forth by the platforms where data is collected. Metadata describing all data and derived datasets will include information about requesting access to and use of the data and datasets. Others using the project’s collected, created, and/or derived data products will be requested to provide a standard citation (proper attribution provided by the project) to credit both the IMLS and this project for access and use of the data.

Plans for Archiving and Preservation
The UNT DL is committed to long-term access and stewardship of publicly available research outputs. Recently, the UNT Libraries completed a self-audit using the criteria of the Trusted Repositories Audit & Certification: Criteria and Checklist (TRAC). Utilizing the facilities provided by the UNT Data Repository, all data generated by this project will be backed up frequently to protect from loss of data from hardware failures, fire, theft, etc. All research products and project-related materials, such as technical reports, presentations, and publications, will be made long-term accessible through the UNT Scholarly Works open access repository, hosted in the UNT DL.

Data Management Plan

Types of Data

Research data that will be used and generated in this project will include bibliographic data in XML and MARC21 formats, Library of Congress Classification and Library of Congress Subject Headings in plain text, survey data in Excel and plain text, and interview data in plain text and images. The data collection and usage will adhere strictly to the data policies set forth by the appropriate platforms and comply with federal and state privacy laws and regulations.

In the process of conducting the project, both quantitative and qualitative data will be generated. Part of the data will be the output of data engineering, analysis, and visualizations. All the code created for the statistical analysis, visualizations, machine learning, and deep learning models will be kept and shared. To the extent possible, data will be shared using standard formats such as JSON and CSV (as described in IETF RFC 4180). Code including R files, python files, textual documents, as well as images of visualizations will be shared. The project will develop prototype tools, which will generate source code, scripts, configuration, data, and documentation files. These files will all be kept.

In terms of project outcomes, digital content on research findings will be shared in reports, posters, video presentations, and publications. The content shared will strictly comply with ethical guidelines in privacy and security.

Data and Metadata Standards

The project team, in collaboration with the University of North Texas Digital Libraries (UNT DL), will choose file formats that will enable the most effective and secure management over time. Best practices for data curation and digital preservation will govern the choices of file formats for text and images. Where proprietary software is used, every effort will be made to provide the output of the software in a format as open as possible. Metadata records will be created to describe each of the project's digital resources. The UNT DL utilizes a system of metadata based on qualified Dublin Core metadata elements that ensures long-term preservation and viability of the digital objects housed in their system. Metadata records will be retained and available for open access through the UNT DL.

Policies for Access, Sharing, and Privacy

Sharable data will be saved in the UNT DL, which is open access and free. Access restrictions can be placed on all items in a deposit. We adhere to the ethical guidelines on access restrictions, confidentiality, and community and individual wishes on data sharing. To store and secure the data, we will store the data in at least three secure locations that are locked, pass-coded, or encrypted, which include a secured server, an external hard-drive in a locked location, and a secure cloud-based server. Requests can be made to the Principal Investigator for limited access to the data generated by the project team. Data and files not to be shared will be saved on two secure servers, which the PI has full control, and an external hard drive in a locked location.

Policies for Re-use, Re-distribution, and Derivatives

To the extent possible, data and derived datasets collected and/or created by the project will be available to other researchers for reuse. The reuse and redistribution will strictly adhere to the data policies set forth by the platforms where data are collected. Metadata describing all data and derived datasets will include information about requesting access to and use of the data and datasets. Others using the project's collected, created, and/or derived data products will be requested to provide a standard citation (proper attribution provided by the project) to credit both the IMLS and this project for access and use of the data.

Plans for Archiving and Preservation

The UNT DL is committed to long-term access and stewardship of publicly available research outputs. Recently, the UNT Libraries completed a self-audit using the criteria of the Trusted Repositories Audit & Certification: Criteria and Checklist (TRAC). Utilizing the facilities provided by the UNT Data Repository, all data generated by this project will be backed up frequently to protect from loss of data from hardware failures, fire, theft, etc. All research products and project-related materials, such as technical reports, presentations, and publications, will be made long-term accessible through the UNT Scholarly Works open access repository, hosted in the UNT DL.

Roles and Responsibilities

The PI will have overall responsibility for the research data management. The PI will regularly check data for accuracy and completeness and review data analysis results for consistency. Disseminating the results generated by this project will be carried out by the PI and graduate research assistants in a multitude of ways. We are asking to support travel by the PI and the graduate research assistant. Major conferences in this area include: *the Joint Conference on Digital Libraries (JCDL)*, *the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, *the Annual Meeting of the Association for Information Science and Technology (ASIS&T)*. Research findings will also be published in peer-reviewed journals. A small selection of these journals includes *JASIS&T, Information Processing and Management (IP&M), Library Resources & Technical Services (LRTS), Information Technology and Libraries (ITAL)*, and *Code4lib Journal*. We will also share our results with Listservs frequented by cataloging and systems librarians, such as *AUTOCAT* and *Code4lib*, for further dissemination and solicitation of feedback from pertinent groups.