LG-256638-OLS-24, Virginia Polytechnic Institute and State University (Libraries)

Harnessing ETDs: Pioneering AI-Driven Innovations in Library Service

William A. Ingram & Edward A. Fox

Project Justification

Virginia Tech's University Libraries, in collaboration with the Digital Library Research Laboratory in the Department of Computer Science, seeks IMLS support for an applied research grant that aims to expand the reach and impact of academic library services through integration with Large Language Models (LLMs). This project builds on the foundational work supported by a 2019–2023 IMLS grant (LG-37-19-0078-19), which pioneered the application of machine learning to enhance libraries' capacity to provide access to knowledge buried in electronic theses and dissertations (ETDs) and other long-form scholarship. This proposal is a substantial leap forward from our previous work, shifting from enhancing accessibility to actively integrating that important content with LLMs in innovative, service-oriented applications for libraries. Our aim is to take advantage of modern artificial intelligence (AI) techniques to enhance the depth and reach of library collections, support graduate students, and foster data-driven academic strategies for libraries and the universities they serve. This project aligns with the NLG Program's third goal: to improve the ability of libraries and archives to provide broad access to and use of information and collections—particularly under Objective 3.2, which supports the design and development of online library and archives services that meet user expectations for operating in an online environment.

ETDs have gained attention in academia and industry due to their potential to provide immediate access to valuable research findings. In 1997 Virginia Tech became the first institution to require that theses and dissertations be submitted as ETDs (Thompson, 2001). This early adoption and promotion of ETDs contributed significantly to the broader acceptance and implementation of electronic submissions at academic institutions worldwide. The development of ETD programs has led to increased discussions around the use of theses and dissertations to further scholarly research, indicating a growing recognition of the value of ETDs (Woods et al., 2020). This wealth of digital data, combined with advanced computational analysis techniques, presents unparalleled opportunities for digital scholarship. However, leveraging AI in academic research introduces hard challenges, notably ensuring factual accuracy and overcoming LLMs' tendency to produce plausible but potentially incorrect information. This situation highlights a gap between AI's potential for academic research and libraries' current capabilities to apply these technologies effectively.

We propose enhancing LLMs with contextually rich information from ETDs to improve the models' understanding and generation of content. Our goal is to make LLMs more reliable and useful for academic and scholarly work, enabling them to provide information and insights that are relevant and trustworthy. Our project will explore advanced information retrieval techniques combined with the latest in AI to create a symbiotic system where LLMs and ETDs enhance each other's value. ETDs offer rich, specialized, and diverse content that helps to inform and improve the LLMs' understanding and outputs. In return, LLMs provide the computational power and advanced language capabilities needed to analyze and make sense of the vast and complex content within ETDs. Our hypothesis is that integrating LLMs with a curated ETD corpus will enhance the accuracy and reliability of information discovery, synthesis, and use—enabling the creation of AI-driven tools for library services, offering novel insights into research trends, and fostering holistic support for graduate students.

Research Questions

Our proposal focuses on improving the content, coverage, and context of LLMs by supplementing their parametric memory with context-specific knowledge from ETDs, particularly for tasks requiring extensive knowledge and factual accuracy. The project is structured in three phases: ***Phase 1*** researches integrating LLMs with ETDs to improve in-context learning performance. ***Phase 2*** employs ETD-adapted LLMs for identifying research trends to aid in university assessment and planning. ***Phase 3*** investigates using ETD-adapted LLMs for targeted support of graduate students, using the rich information latent in ETDs to aid their research and academic endeavors.

We seek answers to the following research questions.

- **RQ1:** How can libraries integrate a curated ETD corpus with LLMs to enhance the accuracy and reliability of generated information?
- **RQ2:** How can the integration of ETDs with LLMs empower libraries to devise analytical tools that provide fresh perspectives on research trends and foster informed, data-driven assessment and planning?
- **RQ3:** How can integrating ETDs with LLMs facilitate the development of AI-driven tools to meet the comprehensive needs of graduate students?

These research questions, along with their corresponding project phases and list of major tasks, are summarized in Table 1 and detailed in the Project Work Plan.

Theoretical Background

While LLMs have gained prominence recently, their foundational concepts are more than 20 years old. The early models proposed by Bengio et al. (2000) laid the conceptual and architectural foundations for neural-based language processing. Modern LLMs extend those foundations to understand and generate human language by predicting the next word in a sequence, given the words that precede it. The predictive ability of modern LLMs such as GPT-3 (Brown et al., 2020) is achieved through training on extensive text corpora using billions of parameters, enabling the models to generate coherent and contextually appropriate text. LLMs leverage a deep learning architecture known as the Transformer (Vaswani et al., 2017), which improved upon the sequential processing paradigm in earlier recurrent neural network (RNN) models (Rumelhart et al., 1986).

The attention mechanism, introduced by Bahdanau et al. (2016) and refined in the Transformer model, allows the network to dynamically weigh the importance of different words in a sentence or passage, enabling the model to capture contextual relationships between words, irrespective of their positional distance in the text. This parallel processing capability improves both the efficiency and effectiveness of the model in handling sequence data. As a result, LLMs are highly effective for a range of language-based applications (Radford et al., 2019). These models can also be applied to specific tasks or domains that may not have been explicitly covered in their training data via downstream transfer to task-specific architectures. Task specification typically involves conditioning the model on additional information, either by including new data or a prompt in its input, or by updating some or all of the model parameters through fine-tuning (Bommasani et al., 2022).

An intriguing aspect of the extensive training of LLMs on vast datasets is their ability to internalize an implicit "knowledge base" (Petroni et al., 2019) in their parameters. Through the pre-training process, these models learn the structure and patterns of language. They also absorb a considerable amount of information and knowledge embedded in the training data. The parameters of LLMs, which can number in the billions, encode and represent a wide range of linguistic and factual knowledge. The emergence of in-context learning (Brown et al., 2020), in which a foundational model (Bommasani et al., 2022) can be adapted to a downstream task for which it was not specifically trained, is achieved simply by providing it with an appropriate prompt. Internalized knowledge gives LLMs the capability of responding to queries with information that is not explicitly stated in the input, but inferred from training. This capability for in-context learning resulting from scale allows them to answer questions based on their accumulated knowledge. However, it is important to note that this internalized knowledge is not always accurate or up-to-date, sometimes leading to the generation of incorrect or misleading information.

Although LLMs have been shown to offer unprecedented capabilities in information processing, contextual understanding, and user interaction, the use of LLMs for scholarly work presents challenges and limitations that libraries must consider. Perhaps most concerning is the propensity of LLMs to hallucinate facts or present inaccurate information with a degree of confidence that could mislead users (Marcus, 2020). Recall that LLMs are trained to maximize the likelihood of generating the next word in a sequence given the preceding context. The model's objective function (often a form of cross-entropy loss) rewards the model for correctly predicting the next word in its training sequences. This approach inherently focuses on local coherence and fluency, rather than global factual accuracy or truthfulness. Since the LLM is optimized to predict plausible continuations of text sequences, it does not inherently distinguish between truth and fiction. The model learns to generate text that is contextually and stylistically coherent with the input, leading to outputs that can be fluent and convincing yet factually incorrect or "hallucinated."

The observed variability in LLM responses and their sometimes overconfident propagation of inaccuracies underscore the need for mechanisms that can enhance their reliability, especially in dynamic fields like scholarly research and academic libraries. Continual training to keep a model updated with current events is costly and computationally demanding (Bommasani et al., 2022). Retrieval-augmented in-context learning emerges as a solution addressing these inherent limitations of LLMs, enhancing their capabilities for factual question answering by incorporating a retrieval mechanism, which, upon receiving a query or prompt, searches an external data source, such as a corpus of scientific literature or an academic database, to retrieve relevant information. In essence, the paradigm enriches the model's responses by augmenting its parametric knowledge (learned during training) with up-to-date information retrieved in real-time from external sources. By fetching relevant and accurate information to supplement the LLM's responses, retrieval augmentation reduces the likelihood of hallucinations (Shuster et al., 2021). Several approaches have been proposed (e.g., Guu et al. (2020); Lewis et al.

(2021); Shi et al. (2023)) in which documents or passages are retrieved from a large set of unstructured documents using a machine-learned matching function and the entire neural system is trained end-to-end. This is an active research area.

The origin of information retrieval (IR) systems can be traced back to fundamental concepts such as the term frequency-inverse document frequency (TF-IDF) (Salton and Yang, 1973). Early IR systems were based on Boolean logic, where users would construct queries using logical operators. These models lacked nuance and flexibility, especially for those without training. TF-IDF led to the development of vector space models, where documents and queries are represented in a multidimensional space (Salton et al., 1975). The relevance of a document to a query could then be determined by calculating the cosine similarity between their respective vectors.

Probabilistic retrieval models (Robertson and Jones, 1976) introduced probabilistic approaches to IR, considering the likelihood that a document is relevant to a query. BM25 (Robertson and Zaragoza, 2009) is based on the probabilistic retrieval framework; it is conceptually related to TF-IDF, sharing the fundamental principle of evaluating the importance of terms in documents relative to queries, but BM25 introduces additional mechanisms for term frequency saturation and document length normalization, making it a more refined and often more effective approach. With the advent of techniques like Latent Semantic Indexing (LSI) (Deerwester et al., 1990), the field began to shift towards a deeper semantic understanding of text, allowing for more nuanced and context-aware retrieval, beyond mere keyword matching.

Sparse retrievers like BM25 and LSI are grounded in lexical analysis, focusing on the presence or absence of words and their frequency-based statistics. “Sparse” here means that for any given document or query, most elements in the representation (e.g., words or terms) have zero (i.e., no) weight. This is in contrast to dense retrievers, which use representations where many elements of the representation have non-zero values. Dense embeddings excel in capturing the subtleties and complexities of language. This includes effectively handling polysemy (words with multiple meanings) and synonymy (different words with similar meanings), as explored in various studies on semantic representation (Pennington et al., 2014).

The introduction of Word2Vec (Mikolov et al., 2013) marked a pivotal moment in NLP by introducing an efficient method for learning and encoding the semantic relationships between words in a high-dimensional space. This breakthrough laid the foundation for subsequent developments in representation learning and the introduction of context-sensitive embeddings: ELMo (Embeddings from Language Models) (Peters et al., 2018), BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019), and GPT (Generative Pre-trained Transformer) (Radford et al., 2018). The concept of dense retrievers gained significant momentum with the introduction of these language models and neural network architectures, particularly those based on transformers. These models extended the idea of word embeddings to generate contextual embeddings for entire sentences or documents, enabling more effective matching between queries and documents based on semantic similarity. Indeed, within a year of the release of BERT, both Google and Microsoft announced that they were incorporating it into their core search technologies (Nayak, 2019; Zhu, 2019).

Just as NLP revolutionized IR with the introduction of transformer models, IR is now reciprocating by revolutionizing NLP through the incorporation of retrieval-augmented learning. This technique effectively bridges the gap between the general knowledge encoded in LLMs and the specialized requirements of specific tasks or queries. It allows training of dense retrievers based on feedback signals from the LLMs themselves (Lewis et al., 2021). When fine-tuned to the specific nuances of LLM responses, retrieval engines can fetch information that is contextually and semantically more appropriate. This method, which leverages the deep semantic understanding inherent in dense embeddings, has demonstrated effectiveness in complex NLP tasks such as semantic parsing and multitask retrieval (Karpukhin et al., 2020). By enabling LLMs to dynamically pull in relevant information from external databases or corpora, retrieval-augmented learning enhances the models’ ability to generate accurate and contextually appropriate responses and addresses the limitations associated with the fixed and implicit knowledge base of pre-trained models.

Relevance to Current Practice

Our applied research, integrating LLMs with ETDs, should facilitate innovative access to the rich collection of graduate scholarship. Underpinning our three RQs are fundamental, practical considerations essential for effective AI application in libraries, such as how to prepare and assess library collections for AI use, or how to engineer a prompt for few-shot learning (Brown et al., 2020). We will disseminate both our findings and the lessons learned during our research investigations, documenting and sharing our methodological journey of integrating AI technologies with library practices. We aim to empower libraries to navigate new technological landscapes, enhancing their services and support for the research community.

Our project aims to enhance library services by integrating advanced AI tools, thereby improving the quality and accessibility of information services in libraries. It will contribute to the professional development of library staff by equipping them with the skills to use AI tools. By integrating LLMs with ETDs, our project will provide libraries and their universities with advanced tools to understand research trends, thus contributing to better informed decision-making and resource allocation. By integrating advanced AI with library services, the project enhances the role of libraries in the national information infrastructure, supporting research, education, and innovation. By developing AI tools tailored for graduate students, particularly focusing on first-generation students, we aim to provide equitable access to advanced research tools and resources, thereby narrowing the gap in digital literacy and access.

Without prior experience in AI, libraries may face a steep learning curve in understanding the methodologies, tools, and best practices (e.g., machine learning algorithms, data pre-processing, model training, and evaluation). A goal of this proposal is to serve as a practical, hands-on example of how AI can be applied in a library setting by providing a tangible demonstration of its capabilities and requirements, helping libraries overcome the "cold start" problem so they can progress with confidence. By thoroughly documenting each step of the project—from model selection and data pre-processing to training and evaluation—and by sharing our source code, we hope our project can offer a step-by-step guide that other libraries can follow, and an adaptable base upon which to expand as they develop their own AI initiatives.

Project Work Plan

Table 1: Project Phases, Research Questions, and Major Tasks

Phase	Research Question	Major Tasks
Phase 1	**RQ1:** How can libraries integrate a curated ETD corpus with LLMs to enhance the accuracy and reliability of generated information?	◇ ETD Corpus Preparation ◇ Experimental Infrastructure ◇ Evaluation System ◇ Algorithmic Experimentation
Phase 2	**RQ2:** How can the integration of ETDs with LLMs empower libraries to devise analytical tools that provide fresh perspectives on research trends and foster informed, data-driven assessment and planning?	◇ Metric Selection ◇ Establishing the Ground Truth ◇ Model Development ◇ Prototyping and Evaluation
Phase 3	**RQ3:** How can integrating ETDs with LLMs facilitate the development of AI-driven tools to meet the comprehensive needs of graduate students?	◇ Needs Assessment ◇ Task Selection ◇ Model Development ◇ Prototyping and Evaluation

Phase 1 — Investigating Research Question 1

Phase 1 of our project, corresponding to **RQ1**, serves as the foundational stage, as it establishes the technical base necessary for the subsequent phases to build upon. The first task is to prepare and expand our ETD corpus, leveraging techniques from previous work. We will then establish a robust experimental infrastructure to explore various retrieval-augmented LLM configurations, focusing on open-source models for their transparency and suitability for academic research. We will design a rigorous evaluation system to test different architectures and configurations, employing challenge datasets and behavioral testing to simulate real-world scenarios. We will explore IR algorithms, comparing state-of-the-art models to find the most suitable ones for handling the ETD corpus. Figure 1 illustrates an overview of the *Phase 1* architecture and how it is fundamental for the downstream tasks that we explore in later phases.

ETD Corpus Preparation: We will prepare and preprocess an extensive corpus of ETDs, made up of full-text PDFs and Dublin Core metadata for ETDs sourced from US university libraries. We have already developed foundational methods and infrastructure for harvesting, and we have collected 500K ETDs during our previous project. This document set represents a broad spectrum of academic disciplines and includes a significant volume of documents ranging from 3,000 to over 50,000 per individual institution. We specifically target an oversampling of ETDs from Historically Black Colleges and Universities (HBCUs) and Hispanic Serving Institutions (HSIs). The overall size of the current collection is about four terabytes. The associated metadata is extensive and varied, encompassing information from over 2,000 distinct academic departments (before consolidating near-duplicate entries). This metadata offers valuable insights into academic collaborations and disciplinary trends. The composition of the ETD corpus is diverse: doctoral dissertations constitute 56%, master's theses 42%, and bachelor's theses about 2%. This rich collection is the foundation of our project, providing a deep and comprehensive dataset for our experiments. But for this project, as explained below, further dataset refinement is needed.

We acknowledge that the effectiveness and ethical integrity of AI systems strongly depend on the diversity of the data used to train them. If bias exists in the training data, the resulting models could perpetuate those biases in their output, leading to narrow or skewed representations of knowledge. This could manifest in several ways, such as prioritizing certain academic disciplines over others in response generation or having limited perspectives on topics that are diverse in nature. Essentially, the model might fail to accurately reflect the wide range of scholarly work across different disciplines, potentially overlooking important contributions from less represented areas. This affects not only the model's reliability and inclusiveness but also its utility across a diverse user base. We will address the following biases to ensure equitable representation. *Disciplinary bias*: Over representation of certain academic disciplines could skew the model's knowledge and output towards those fields. *Demographic bias*: A lack of diversity in authorship could result in models that do not adequately represent the perspectives or knowledge of all demographic groups. *Geographic bias*: Concentrating on institutions from specific regions could overlook valuable research and insights from other areas.

To do so, we will analyze the existing 500K corpus to identify gaps by categorizing ETDs by discipline, demographic indicators (if available), and institution characteristics, such as geographic location and size. Once gaps are identified, targeted harvesting can be aimed at institutions or disciplines that are underrepresented. By focusing efforts to enrich the corpus where they are most needed, we aim to enhance the diversity and comprehensiveness of the dataset for more balanced and inclusive AI model training.

Next, we will build on and refine the techniques we explored in our previous work (Ahuja et al., 2022) to segment and annotate ETDs to identify key sections (e.g., abstract, literature review, methodology, results) for our experiments. After extracting the textual content from PDF documents, we will clean the extracted text to remove any artifacts from the PDF conversion process, such as headers, footers, and page numbers. We will store the Dublin Core metadata in a database for easy access and association with the respective documents. A key lesson learned from our last project is that we underestimated the time and effort required to prepare the collection for machine learning tasks. We will explore techniques that leverage the capabilities of LLMs to guide the process (e.g., Wang et al. (2024)), potentially reducing the reliance on extensive hand-labeled data. The need for libraries to effectively assess the "readiness" of their collections for AI applications was a topic of concern at a recent workshop, "Leading the Future of AI and Public Archives" (Ingram et al., 2022). This proposal aims to directly address that concern.

Experimental Infrastructure: Our project will establish a robust experimental infrastructure capable of handling complex NLP tasks. Our exploration will focus on the Llama 2 (Touvron et al., 2023) suite of models and other open-source models like MPT (MosaicML NLP Team, 2023) or Falcon (Almazrouei et al., 2023) for comparative analysis. These models represent the cutting edge of what is achievable in the domain of open LLMs, but the field is advancing rapidly—we expect the state-of-the-art will have changed by the time our project launches. We prefer open models that offer transparency in their training data, algorithms, and internal processes, which is a plus for academic and research purposes, allowing for greater scrutiny, adaptability, and understanding of the model's behavior and decisions. In contrast, commercial and proprietary models provide limited insight into their workings, offering high performance but limited customization and interpretability. Conducting a comparative analysis of multiple models will enable us to identify the most suitable ones for encoding and synthesizing information from ETDs. Each model's unique training, size, and parameter count contribute to its ability to understand and process complex academic texts. When comparing these models, we can determine which offers the best balance between performance, computational efficiency, and adaptability to ETD content.

Evaluation System: Rigorous evaluation is essential to validate the effectiveness and applicability of LLMs in library settings, particularly when dealing with complex, long-form academic texts like ETDs. Automated testing pipelines are needed that can efficiently process large volumes of data and generate performance metrics, facilitating iterative improvements based on consistent and reliable testing conditions. We will focus on behavioral testing, which involves evaluating models based on their performance on tasks that simulate real-world scenarios. This form of testing often uses challenge datasets, which are curated collections of data specifically designed to test the limits of a model's understanding and capabilities. The following datasets are a selection of ones commonly used for NLP testing and benchmarking.

SQuAD (Stanford Question Answering Dataset) (Rajpurkar et al., 2016) is a well-known dataset for question-answering models, where the answer to every question is a passage of text from Wikipedia articles. Natural Questions (NQ) (Kwiatkowski et al., 2019) is a dataset that contains real user queries from Google search and the corresponding answers found in Wikipedia. OpenBookQA (Mihaylov et al., 2018) is focused on science questions requiring reasoning and understanding of basic science concepts. SciFact (Wadden et al., 2020, 2022) is a a dataset and framework designed for the verification of scientific claims by examining their concordance with a corpus of abstracts from scientific papers. Given the unique nature of ETD content, it could be necessary to create custom datasets by extracting queries, summaries, and factual questions directly from the ETDs themselves.

We will engage with experts in the humanities, social sciences, and cultural studies regularly throughout the project to guide us on how to address cultural competence and ethical soundness. The goal is to ensure that datasets used in training and testing the models reflect a wide array of cultural contexts and experiences. With their guidance, we will develop protocols that prioritize the inclusion of diverse perspectives in our challenge datasets. This may involve targeted collection efforts to ensure representation from a broad spectrum of disciplines and cultural backgrounds. It is important that our evaluation system can assess not only the accuracy and efficiency of LLMs, but also their ability to handle content from diverse cultural contexts, sensitively and appropriately. This motivates our creating custom challenge datasets reflecting the variety of ETD content and using them to test model performance on tasks requiring cultural competence.

Algorithmic Experimentation: Finally, to address **RQ1**, we will experiment with retrieval algorithms and techniques, and research how retrieved passages can best be used to enhance the LLMs' in-context learning capabilities, focusing on the refinement of prompts that effectively draw on the ETD corpus for generating contextually rich and accurate responses. We will experiment with state-of-the-art retrieval models, each with varied approaches to handling and structuring information. Systems such as DPR (Karpukhin et al., 2020), ColBERT (Khattab and Zaharia, 2020), RAG (Lewis et al., 2021), SPLADE (Formal et al., 2021), and Contriever (Izacard and Grave, 2021) represent the forefront of combining advanced retrieval techniques with deep learning to enhance the capabilities of LLMs in understanding and synthesizing complex information. These models leverage neural networks, particularly transformer-based models, to capture the semantic content of both queries and documents, going beyond simple keyword matching to a more nuanced understanding of text.

Training these retrieval components may involve fine-tuning on datasets relevant to the retrieval tasks, but typically does not require modifying the original pre-training of the base language models. By experimenting with a range of models, we will investigate which approach or combination thereof best aligns with the unique characteristics of ETDs. However, conducting experiments with multiple models requires significant computational resources and time. To assess the feasibility, we will start with preliminary experiments on a small scale with a few models to gauge their effectiveness. Based on these initial results, we will then focus on the most promising models for further in-depth experimentation. While our focus is on dense retrieval models, we will also consider a sparse retrieval model like BM25 for baseline comparison and combinations.

We will evaluate our experiments targeting **RQ1** iteratively, using outcomes to refine our approach and improve performance. Test failures may not necessarily implicate a deficiency in a model's fundamental capabilities. They are far more likely to reveal gaps in the retrieval corpus that prevent the model from achieving the desired learning target or problems with the structure or organization of the data that might necessitate optimization to align with the retrieval system's capabilities. These findings are intrinsically valuable, as they prompt further investigative questions and guide

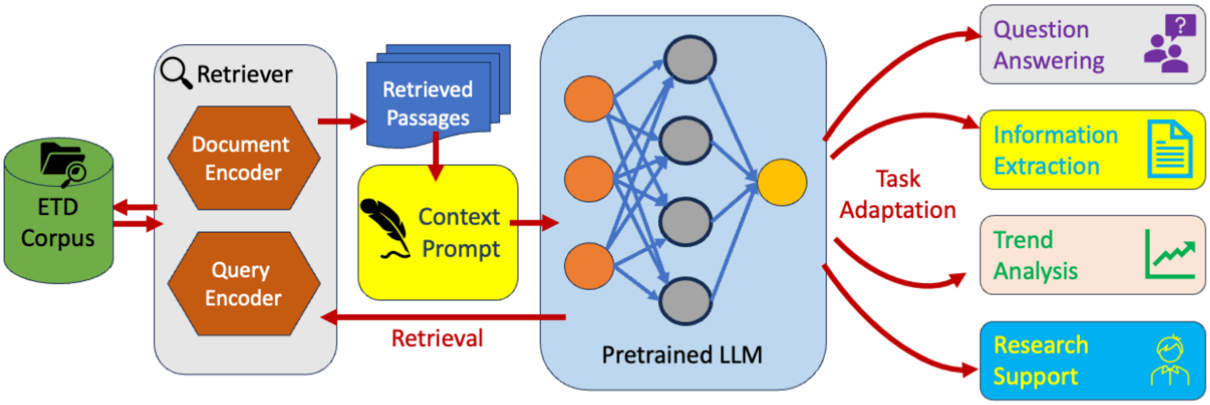

Figure 1: Retrieval-Augmented In-Context Learning from ETDs. Documents from the ETD corpus are encoded and queried to retrieve contextually relevant passages. These passages augment the LLM's context via an enhanced prompt, improving the model's ability to generate content tailored to specific library-related tasks such as question answering, information extraction, trend analysis, and research support.

the refinement of the system. Ultimately, these exploratory outcomes will aid us in configuring the system iteratively, enhancing our capacity to reach the specified learning targets effectively.

Phase 2 — Investigating Research Question 2

Libraries have become centers for data-driven insights into research trends and impacts, collaborating with research offices, academic departments, and administrative units to provide a cohesive picture of research impact and inform policy and strategy. By analyzing bibliometric and scientometric data, libraries provide valuable insights into publication trends, citation impacts, and the influence of research both within and beyond academia. This foray into research impact and intelligence signifies a transformative phase where libraries are not only custodians of information but also key players in shaping and understanding the research landscape. Librarians with expertise in competitive intelligence can help institutions understand their position relative to peer institutions. *Phase 2* of our project addresses **RQ2** in aiming to help libraries develop AI skills for data analysis, positioning them as innovative and technologically advanced units within the university.

Metric Selection: Working with the university's Analytics and Institutional Effectiveness division, we identified the 17 Sustainable Development Goals (SDGs) as a key metric to measure and assess academic research trends and contributions. The SDGs, established by the United Nations, provide a framework that offers a wide range of research impact areas (e.g., poverty, health, equality and climate change). Using these categories for research assessment can inform program development and resource allocation, ensuring that academic efforts align with global needs and the institution's strategic goals. Identifying research areas aligned with specific SDGs can facilitate collaborations and guide funding applications, especially for projects aimed at societal and environmental impact. The ability to report research contributions in the context of the SDGs enhances the institution's visibility and appeals to stakeholders, including prospective students, faculty, and funding bodies that are increasingly focused on sustainable development.

Establishing the Ground Truth: Since there is no existing dataset of ETDs labeled by the SDGs they support, we plan to use journal articles as a surrogate. Each year since 2018, Elsevier has released an updated set of specific Scopus search queries for each SDG, in an effort to identify research that support these goals. Using Elsevier's latest SDG search queries (Bedard-Vallee et al., 2023) we will build a ground-truth dataset by harvesting at least 1,000 article abstracts for each SDG query from Scopus, accumulating a dataset of potentially 17,000 abstracts. We will assign SDG labels to each abstract based on the search query it was retrieved with. It is not uncommon for a paper to support multiple SDGs, so abstracts may be given multiple labels. We will process the text so that it is structured and labeled appropriately. Although this data set contains article abstracts, rather than ETDs, our intuition is that a classifier trained on these data could be used to classify ETDs through transfer learning and suitable adaptation.

Model Development: Using our labeled SDG dataset, we will train a language model-based classifier to automatically classify papers by the SDGs they support. This is a multi-label classification problem, in that a single paper can contribute to more than one SDG. We will explore various model architectures to optimize classification accuracy and efficiency. Using transfer learning, we aim to extend the classifier's capabilities to ETDs, offering early insights into Virginia Tech's research landscape in the context of global sustainability goals, not yet present in the published literature. The next step is to use this trained classifier to aid in retrieval, to assess whether a retrieval-augmented LLM can effectively synthesize and summarize information about the university's research portfolio as reflected in its ETDs, highlighting their contributions to SDGs. This analysis aims to identify prominent and underrepresented research domains, facilitating data-driven institutional assessment and decision making. We will encourage applying our approach at other institutions.

Prototyping and Evaluation: To assess **RQ2**, we will develop the analytical tools described above and evaluate them within the context of Virginia Tech. We will develop a prototype dashboard capable of providing an overview summary at various organizational levels—institution, department, discipline, and individual faculty member—as these levels are described in the ETD metadata. The dashboard will dynamically present AI-generated summaries of research aligned with the SDGs, alongside direct links to access the ETDs themselves for further reading. During the prototyping and evaluation phase, we will work in close collaboration with Virginia Tech's Office of Sustainability and the Analytics and Institutional Effectiveness division of Academic Resource Management. These partnerships will help ensure that our analytical tools are aligned with the university's strategic goals, particularly in the context of sustainable development and institutional effectiveness. We will gather qualitative feedback from stakeholders within these units to gauge the usability and utility of these tools in academic and administrative contexts.

Phase 3 — Investigating Research Question 3

Needs Assessment for Graduate Students: The first task involves conducting a needs assessment to identify key areas where graduate students, especially those who are the first in their families to pursue graduate education, would like support. To do so, we will organize an advisory board comprised of graduate students at Virginia Tech, prioritizing the inclusion of first-generation students. We expect that this approach can provide invaluable insight and perspectives, particularly in understanding the unique challenges and needs of a diverse student body. Inclusion of first-generation graduate students on the advisory board is a strategic decision that aligns with our commitment to diversity and inclusivity, recognizing the distinct perspectives first-generation students bring to academia. We will ask them to offer first-hand perspectives on the types of research trends and analytical tools that would be most beneficial to their academic work. This focus can lead to the creation of tools that are more inclusive, supportive, and tailored to diverse student needs.

Task Selection: Using the graduate student needs assessment as a guide, we can prioritize and develop a set of specific tasks that are directly relevant and aligned to provide practical and effective support, ultimately improving their academic experience and future success. This could include assistance in effective writing, understanding research methodologies, or even career guidance, leveraging data-driven insights to help students navigate evolving opportunities in various fields, aligning their research with industry and academic trends. Recognizing the multifaceted nature of the "whole student," we will select tasks that demonstrate how AI tools can adapt and possibly anticipate the changing academic needs of graduate students, thus facilitating a more responsive and nuanced academic research ecosystem. These tasks will lead to concrete learning targets, helping define the success criteria and providing a clear direction for system development and evaluation, ensuring that AI-driven tools are tailored to improve student support in a meaningful way.

Model Development: We will investigate various model architectures for the task-specific goal of providing graduate student assistance. This includes feeding task-specific instructions and input-output demonstrations into LLMs for in-context learning. LLMs will be conditioned on these instructions to generate relevant and contextually appropriate answers and guidance, avoiding the need for extensive parameter fine-tuning. We will integrate feedback mechanisms to continuously gauge the effectiveness of our tools. This includes assessing the LLMs' ability to adapt to the evolving academic needs of graduate students, ensuring that the tools are responsive and relevant. A critical aspect of our project is to ensure that an LLM can efficiently retrieve and use information from ETDs despite their length and complexity. We will evaluate LLM retrieval-augmented capabilities, ensuring handling the task of synthesizing and summarizing content relevant to graduate student research and academic interests.

Prototyping and Evaluation: Based on the needs assessment, we will develop prototype AI tools that leverage ETD-augmented LLMs. The prototypes will aim to offer intuitive and effective support, utilizing the LLMs' capabilities to synthesize information from ETDs. We will test the prototype tools in real-world academic settings, focusing on their effectiveness in addressing identified needs of graduate students. This phase will involve iterative refinement, where feedback from advisory board members and other students is used to improve the tools. We will employ methodologies such as abstractive summarization to condense and present complex academic content in an accessible format. Throughout this work plan, we will maintain a focus on ethical AI practices, ensuring that the tools developed are culturally competent and sensitive to the diverse backgrounds and experiences of graduate students. Our goal is to create supportive and inclusive academic research aids.

All Phases — Reporting and Sharing

During every phase of our three-year applied research project, we will allocate time to synthesize our findings, disseminate knowledge, and ensure that the practical applications of our research are communicated to the library community. By sharing our progress, we aim to foster a broader understanding and contribute to the establishment of how AI can be used in libraries. Our evaluation will focus on how these tools and techniques improve access to information, support academic research, and enhance library services. We hope that this analysis will be useful in guiding future research directions and informing best practices in the application of AI in libraries.

Sharing with researchers: We plan to publish our findings in high-impact peer-reviewed journals and conferences. We will prioritize open-access journals to ensure wider accessibility of our research. Venues such as the Joint Conference on Digital Libraries (JCDL), Theory and Practice of Digital Libraries (TPDL), Special Interest Group on Information Retrieval (SIGIR), the International Symposium on Electronic Theses and Dissertations, and the Journal of the Association for Information Science and Technology (JASIST) will be targeted. Additionally, we will disseminate pre-publication preprints on arXiv and the Libraries' institutional repository. To foster discussions and collaborations within the research community, we plan to organize workshops in conjunction with top-tier conferences. These workshops will focus on themes relevant to our project, offering a platform for in-depth discussions and knowledge exchange. Our team, with a strong record of presenting at academic forums, will actively participate in these events to showcase our research progress and findings.

Sharing with library practitioners: In line with our commitment to advance knowledge and skills within the library community, we plan to organize an online workshop focused on professional development for librarians. This workshop will be designed to address the unique challenges and opportunities librarians face when working with large language models and digital collections. Key themes of the workshop will include: integrating AI and machine learning with library services, practical applications of LLMs in enhancing digital collections, strategies for managing and utilizing large datasets in libraries, and navigating ethical considerations in AI deployment within libraries. By focusing on these areas, we aim to empower librarians with the knowledge and tools needed to innovate in their roles.

Project Team

Project activities will be conducted at Virginia Tech's University Libraries in close collaboration with the Department of Computer Science. Responsibility for management, research, and dissemination will be shared between PI Ingram and Co-PI Fox. They will build upon the findings of their previous, IMLS-sponsored project, which pioneered the application of machine learning to enhance libraries' capacity to provide access to knowledge buried in ETDs. The grant will support one graduate research assistant (GRA) enrolled in the Ph.D. program in computer science at Virginia Tech.

William A. Ingram will serve as Principal Investigator and Project Director. He will manage resources, oversee progress, and ensure that project milestones are met within stipulated timelines. Ingram's project responsibilities will include ETD corpus preparation, model selection and training, and leading the development of specific AI-driven tools. As PI, Ingram will oversee the financial aspects of the project, including budget allocation. He will ensure that the project adheres to ethical guidelines, particularly in terms of data handling, privacy concerns, and AI biases, and will serve as project liaison with the IRB and other regulatory bodies, ensuring that the project complies with all required institutional, national, and international regulations. He will supervise the GRA, leveraging his expertise to foster a learning environment conducive to innovative research and development.

Edward A. Fox will serve as Co-PI. He will coordinate with PI Ingram to ensure cohesive project execution. Dr. Fox's project responsibilities include forming and managing the advisory board, incorporating its feedback into the project, and ensuring that the project remains aligned with user needs and expectations. He will foster collaborative relationships with academic, industry, and community stakeholders. He will lead efforts in disseminating project findings, publishing papers, and conducting workshops or webinars to share knowledge and progress with the broader community.

The project will benefit from synergistic collaboration between Ingram's technical expertise and Fox's extensive experience in digital libraries and information retrieval. Their combined skills will ensure a comprehensive approach to the project, encompassing both the development of innovative AI tools and effective dissemination of research findings. Both Ingram and Fox will engage with stakeholders, including the advisory board and broader academic community, ensuring the project aligns with user needs and academic standards. They will also jointly oversee risk management and contingency planning, ensuring adaptability and resilience.

A Graduate Research Assistant with research interests in NLP and IR will be hired from the Ph.D. program in Computer Science at Virginia Tech. They will undertake tasks such as organizing and preprocessing the ETD corpus, developing and fine-tuning AI models for specific tasks, and creating essential project code and algorithms, including data processing and model training. Responsible for maintaining detailed records of methodologies, experiment designs, and results, the GRA will also contribute to project reporting, participate in team meetings, and help disseminate project outcomes though article writing, conferencing, and workshops. They will benefit professionally from mentorship, training opportunities, and exposure to interdisciplinary research. By working in the library, the GRA will gain unique interdisciplinary experience, broadening their academic and professional horizons.

Diversity Plan

In line with Virginia Tech's Inclusive VT strategic initiatives, this project is committed to fostering diversity and inclusivity in all aspects of the project. Our dedication to these principles is more than a formal adherence—it is a recognition of the integral role diversity plays in creating equitable, humanistic technology that serves the public good and upholds democratic values. Reflecting on the lessons from Virginia Tech's Principles of Community, our project places a strong emphasis on valuing the inherent dignity and value of every person. We are particularly mindful of the need to create AI systems and research methodologies that are culturally competent and ethically sound, considering diverse cultural contexts and experiences. We are committed to creating models that actively promote inclusivity and fairness, acknowledging and addressing the diverse contexts and experiences of all users. For example, in the *ETD Corpus Preparation* task in Phase 1, we lay out concrete steps to avoid perpetuating biases and ensure equitable representation in the ETDs we collect. Our aim is to prevent the reinforcement of stereotypes or marginalization of any group and ensure that the benefits of AI advancements are accessible and meaningful to a broad spectrum of the community.

Our advisory board is focused on first-generation graduate students. We plan to involve students from diverse backgrounds on the board, providing them with opportunities to contribute to and learn from this initiative. This will not only support their academic and professional development, but also infuse our project with fresh, innovative ideas. We welcome students from diverse backgrounds, including BIPOC, LGBTQ, and other minoritized communities. Recognizing the importance of continuous learning and adaptation in our diversity and inclusivity efforts, we will regularly review and adjust our strategies to ensure that they remain effective and aligned with the evolving understanding of what it means to be truly inclusive in research and technology development. Through these concerted efforts, we aim to set a precedent for how AI integration in academic research can be conducted ethically, inclusively, and effectively.

Project Results

Our project and its deliverables are structured to contribute knowledge to researchers and practitioners alike. These deliverables include ***(1) A Comprehensive Report*** on the integration of LLMs with the ETD corpus. It will detail our methodology, processes, and the outcomes of our experiments, serving as a valuable resource for understanding the project's impact on library sciences and AI applications. ***(2) Publications and Workshops:*** Our team will produce research articles and organize workshops to disseminate findings within the academic and library communities. These activities will target influential conferences and journals, emphasizing the impact of our work in library and information science. The workshops will focus on professional development, sharing best practices, and fostering AI application skills in library settings. ***(3) Open Source Software and Datasets:*** All software developed, including prototypes and datasets created during the project, will be available as open source. This includes software prototypes that exemplify the practical application of our research and provide templates for similar future endeavors, as well as software for testing and measuring its effectiveness. ***(4) Experiential Learning Opportunities:*** A significant portion of our funding is allocated to support a full-time Graduate Research Assistant. This investment advances our research while benefiting the student's education and professional growth. The project will also serve as a practical framework for course projects (Chekuri et al., 2023), providing hands-on learning experiences that are directly applicable in real world settings. Our previous project successfully blended student participation in research with their educational advancement, demonstrating the dual benefit of enriching both the research scope and student learning. We plan to continue this approach by engaging students in practical tasks involving ETD datasets, LLMs, and AI-driven tools. These opportunities will allow students to apply their theoretical knowledge in practical scenarios, enhancing their educational experience and contributing significantly to the project's objectives.

Our project deliverables are designed to maximize impact, enrich interdisciplinary scholarship, and enhance professional development in librarianship. We are committed to responsible management and transparent reporting of our methods and progress, ensuring our contributions are accessible, reproducible, and valuable to the broader library and research community.

Harnessing ETDs: Pioneering AI-Driven Innovations in Library Service

William A. Ingram & Edward A. Fox

Schedule of Completion

	Year 1 (2024–2025)											
Activities and Milestones	**A**	**S**	**O**	**N**	**D**	**J**	**F**	**M**	**A**	**M**	**J**	**J**
Administrative Tasks												
Launch Project *Lead: Ingram and Fox*	■											
Hire graduate research assistant *Lead: Ingram*	■											
Review data management plan *Lead: Ingram*	■			■			■			■		
Develop and review project work plan *Lead: Ingram*	■	■			■	■			■			■
Review monthly budget reports *Lead: Ingram*	■	■	■	■	■	■	■	■	■	■	■	■
Create and update project website/blog *Lead: GRA*	■	■	■	■	■	■	■	■	■	■	■	■
Annual reporting *Lead: Ingram*											■	■
Phase 1 — Investigating RQ1												
ETD Corpus Preparation *Lead: Ingram*	■	■	■	■	■	■						
Experimental Infrastructure Setup *Lead: Ingram*		■	■	■	■	■						
Evaluation System *Lead: Fox and Ingram*						■	■	■	■			
Algorithmic Experimentation *Lead: Ingram*										■	■	■
Phase 2 — Investigating RQ2												
Metric Selection *Lead: Fox and Ingram*						■	■					
Establishing the Ground Truth *Lead: Ingram*							■	■				
Model Development *Lead: Ingram*									■	■	■	■
Reporting and Sharing												
Draft and Submit Conference Poster/Short Paper *Lead: Fox*					■	■						

	Year 2 (2025–2026)											
Activities and Milestones	**A**	**S**	**O**	**N**	**D**	**J**	**F**	**M**	**A**	**M**	**J**	**J**
Administrative Tasks												
Review data management plan *Lead: Ingram*	🟥			🟥			🟥			🟥		
Review and update project work plan *Lead: Ingram*	🟥				🟥	🟥			🟥			🟥
Review monthly budget reports *Lead: Ingram*	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥
Update project website/blog *Lead: GRA*	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥
Annual reporting *Lead: Ingram*											🟥	🟥
Phase 1 — Investigating RQ1												
Conclude Algorithmic Experimentation *Lead: Ingram*	🟦	🟦										
Phase 2 — Investigating RQ2												
Prototyping and Evaluation *Lead: Ingram and Fox*	🟧	🟧	🟧	🟧	🟧	🟧						
Phase 3 — Investigating RQ3												
Assemble Graduate Student Advisory Board *Lead: Fox*						🟩						
Needs Assessment with Advisory Board *Lead: Fox*						🟩	🟩	🟩				
Task Selection with Advisory Board *Lead: Fox*									🟩	🟩		
Model Development *Lead: Ingram*											🟩	🟩
Reporting and Sharing												
Draft and Submit Conference/Journal Paper *Lead: Fox*					🟪	🟪						
Draft and Submit Conference Workshop Proposal *Lead: Fox*					🟪	🟪						

	Year 3 (2026–2027)											
Activities and Milestones	**A**	**S**	**O**	**N**	**D**	**J**	**F**	**M**	**A**	**M**	**J**	**J**
Administrative Tasks												
Review data management plan *Lead: Ingram*	■			■			■			■		
Review and update project work plan *Lead: Ingram*	■				■	■			■			
Review monthly budget reports *Lead: Ingram*	■	■	■	■	■	■	■	■	■	■	■	■
Update project website/blog *Lead: GRA*	■	■	■	■	■	■	■	■	■	■	■	■
Final reporting *Lead: Ingram*											■	■
Project Wrap-up *Lead: Ingram*												■
Phase 3 — Investigating RQ3												
Re-convene Advisory Board *Lead: Fox*	■											
Elicit Iterative Feedback from Advisory Board *Lead: Fox*	■	■	■	■	■							
Model Development *Lead: Ingram*	■	■										
Prototyping and Evaluation *Lead: Ingram and Fox*			■	■	■	■	■					
Reporting and Sharing												
Draft and Submit Conference/Journal Paper *Lead: Fox*					■	■						
Draft and Submit Conference Workshop Proposal *Lead: Fox*					■	■						
Draft Library Workshop Plans & Documents *Lead: Ingram*							■	■				
Announce Library Workshop on Relevant Listservs *Lead: Ingram*								■				
Deliver Library Workshop *Lead: Ingram*									■			

Harnessing ETDs: Pioneering AI-Driven Innovations in Library Service

William A. Ingram & Edward A. Fox

Digital Products Plan

Type

Digital content, resources, or assets created or collected:

1. Enhanced ETD Corpus: A curated and processed dataset of electronic theses and dissertations, harvested from institutional repositories at universities across the country. Each ETD is made up of one or more full-text document files and metadata. This corpus will be enhanced with additional metadata, segmentation, and annotations to facilitate deep learning and NLP tasks.
   (a) Quantity: Extensive, covering over 500K ETDs previously collected, with additional targeted harvesting to fill identified gaps.
   (b) Format: Full text documents are usually in PDF format, with textual content extracted and stored as plain text or CSV format for ease of processing. Harvested metadata is usually encoded in a variant of simple Dublin Core and serialized as XML. Metadata elements are stored in a PostgreSQL database.
   (c) Standards: Metadata standards such as Dublin Core and ETD-ms (https://www.ndltd.org/standards/metadata/current.html) ensure compatibility with existing library systems. Text encoding and storage will follow best practices for NLP and AI research (e.g., plain text, UTF-8, CSV) ensuring interoperability and ease of use.
2. AI Models and Algorithms: Specifically developed or adapted LLMs and supporting algorithms for ETD analysis, including segmentation, passage extraction, vectorization, and encoding for information retrieval.
   (a) Programming Languages: Python, due to its extensive support for AI and machine learning libraries (e.g., pandas, numpy, matplotlib, scikit-learn).
   (b) Platforms/Frameworks: Open-source platforms like PyTorch for model development and training, chosen for their flexibility, extensive community support, and alignment with academic and research standards.
   (c) Pre-trained Language Models: Our project will leverage pre-trained models available on Hugging Face, such as Llama 2, BERT, and many others, tailored for specific NLP tasks relevant to our project goals.
   (d) Retrieval Models: Our project will explore various advanced retrieval models (e.g., DPR, ColBERT, SPLADE) for experiments on retrieval-augmented LLM architectures.
3. Software Tools and Digital Tools: Applications and tools designed to interface with the ETD corpus and AI models, facilitating access, analysis, and visualization of data and insights derived from ETDs.
   (a) Platforms/Frameworks: Web-based interfaces developed using HTML, CSS, and JavaScript for the frontend, with Python-based frameworks (e.g., Django, Flask) for backend operations. These choices are motivated by the need for cross-platform compatibility, ease of deployment, and the widespread familiarity within the developer community.
4. Documentation: Comprehensive documentation covering the use, maintenance, and potential customization of the digital tools and AI models, ensuring usability across various user groups.
   (a) Format: Online web pages or blog posts for easy access and updates. Software documentation in README files in plain text or markdown. Using literate programming with Jupyter notebooks to promote usability. Research papers and reports in PDF.

Availability

In alignment with IMLS guidelines, our project is committed to maximizing the availability of its outputs, including publications, computational models, software tools, and metadata.

1. Publications and Research Outputs: All scholarly articles resulting from this research will be submitted to open-access journals or repositories. Preprints will be shared on platforms like arXiv or in Virginia Tech's institutional repository ahead of journal publication.

2. Software and Digital Tools: Software developed, including source code for computational models and proof-of-concept tools, will be hosted on GitHub, offering transparent access to both source code and documentation for installation, usage, and adaptation by the community.

3. Datasets: For publicly distributable data derived from our research, we will use the Virginia Tech Data Repository, hosted by Figshare.

Access

Open-Source Licensing: Consistent with IMLS expectations, our project will release digital products, including software and datasets, under open-source licenses. Specifically, we plan to use the BSD 3-Clause License for software, recognizing its compatibility with academic and commercial use. For datasets and documentation, we will employ Creative Commons licenses, likely CC BY 4.0, which allows others to share, use, and build upon our work, provided they attribute the original creation to us.

Limitations: While our intent is to minimize restrictions on the use of our digital products, certain datasets, particularly those derived from ETDs, may be bound by licensing agreements or privacy laws that restrict redistribution. In such cases, we will provide a link to the ETD repository of origin.

Privacy Concerns: Given our use of publicly available ETDs, we are mindful of privacy concerns, especially regarding personally identifiable information (PII). We will rigorously review datasets to anonymize or redact PII where necessary, ensuring compliance with privacy laws and ethical standards.

Sustainability

Permanent Preservation: Digital products intended for long-term use, such as final datasets, software tools, and finalized research papers, will be preserved in institutional repositories and platforms designed for permanence:

1. Virginia Tech Data Repository (Figshare): Hosts datasets and documentation.

2. VTechWorks—Virginia Tech's Institutional Repository (DSpace): Provides a secure environment for preserving and sharing academic outputs across a range of content types.

3. Code Ocean: This cloud-based computational reproducibility platform has tools to enhance the accessibility and reproducibility of research code and data, which can be packaged and stored in a long-term preservation repository.

Medium- and Short-term Storage: Digital products like draft or working documents, code under development, and any ephemeral products will be stored in cloud-based services with strong backup and recovery protocols, ensuring they are available for the medium or short term as required. The duration of retention for these materials will be determined based on their ongoing utility and relevance to the project and broader research community.

Technical documentation: Comprehensive documentation will accompany all digital products, detailing their use, technical specifications, and maintenance requirements. This ensures that future users and project staff can understand and manage these resources effectively.

Migration planning and commitment of organizational funding: Virginia Tech University Libraries is an institutional member in the APTrust. This partnership provides a robust infrastructure for migrating digital assets to current formats and platforms, safeguarding against data loss and ensuring ongoing accessibility. Membership in the APTrust is part of Virginia Tech University Libraries' broader strategy for digital sustainability, offering access to shared resources, best practices, and a community of institutions dedicated to the preservation of digital scholarship. This collaborative model enhances our ability to sustain digital products over the long term, benefiting from shared expertise and economies of scale.

Harnessing ETDs: Pioneering AI-Driven Innovations in Library Service

William A. Ingram & Edward A. Fox

Data Management Plan

This project will engage in extensive data collection and generation activities across its multiple phases, with the primary aim of integrating and enhancing library services through advanced AI methodologies. Below is a detailed overview of the types, volumes, purposes, collection methods, scope, and timelines associated with the data involved in this initiative.

Data Collection and Generation

ETD Corpus: The foundational dataset consists of a curated corpus of previously amassed Electronic Theses and Dissertations (ETDs), currently exceeding 500,000 documents and approximately four terabytes in size. This corpus spans a diverse range of academic disciplines and includes both full-text PDF documents and Dublin Core metadata. Using the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) and bespoke web crawling/scraping techniques, we will continually augment our ETD corpus to ensure its diversity and comprehensiveness. The specific volume of data added will depend on the discovery and accessibility of new ETDs, but we expect this to be in the tens (if not hundreds) of thousands. We will segment the ETD corpus, preprocess the text (cleaning, tokenizing), and create task-specific training datasets and challenge datasets for evaluation.

AI Models and Algorithms: We will adapt and fine-tune a range of AI models, specifically large language models (LLMs), to process and analyze the ETD corpus. We will use open source models such as Llama 2, BERT, and others relevant to our NLP tasks, downloaded from platforms like Hugging Face, which hosts a wide array of pre-trained LLMs, or from GitHub repositories. In addition, we will explore various dense retrieval models, most of which are available on GitHub. For models we train and/or fine-tune, we will serialize and store the model weights in a widely compatible format, such as ONNX or the native format of the training framework, to ensure broad usability and to facilitate sharing.

Software and Algorithms: We will create software applications and Jupyter notebooks to assess the performance of different LLM and IR model architectures, focusing on their effectiveness in processing and retrieving information from the ETD corpus. We primarily develop in Python, due to its extensive support for AI libraries and frameworks necessary for working with LLMs and developing IR systems. In addition to model and architecture development, we will create web-based applications for data interaction, applying modern web technologies for frontend and Python for backend, ensuring the tools are accessible, intuitive, and scalable. Software development will commence early in the project and will scale with project growth, with initial prototypes expected within the first year, and refinements following evaluation and testing.

Sensitive Information, PII, and Intellectual Property

We do not plan to collect any personally identifiable information (PII), confidential information, or proprietary information. Our primary data source, the corpus of ETDs, consists of publicly available academic works submitted to and hosted by university libraries. However, we acknowledge the importance of ethical considerations and data privacy, especially in handling any incidental personal data that might appear within the ETDs. We will rigorously review datasets to anonymize or redact PII when necessary, ensuring compliance with privacy laws and ethical standards. Should any potentially sensitive information be identified during our data processing activities, we will implement appropriate measures to ensure its protection, including data aggregation, which is common in statistical analyses.

Working with ETDs, we will ensure that our research practices respect the authors' intellectual contributions. In most cases, copyright for the full text of ETDs, including the abstracts, resides with the authors. When using ETDs for our research, we adhere to the principles of fair use, which allow limited use of copyrighted material without permission for purposes such as criticism, comment, news reporting, teaching, scholarship, or research. Our use of ETDs aims to advance scholarly research and education by providing insights into the academic discourse encapsulated in these summaries of academic work. Our intention is not to replicate the full content of ETDs but to study and analyze the thematic and structural elements of academic writing. This analysis contributes to the fields of information retrieval and digital libraries by improving access to and understanding of scholarly work.

Technical Requirements

For the effective retrieval, display, processing, and reuse of data in our project, several technical requirements and dependencies are requried. These are outlined below, along with how these tools can be accessed.

Hardware requirements: High-performance computing resources are essential for processing the large ETD corpus and running complex AI models. This includes multi-core CPUs, high RAM capacity, and, importantly, GPUs for deep learning tasks. Adequate storage solutions are needed to host the extensive ETD corpus, intermediate processing data, and the resulting models. Both fast-access storage (SSD) for active processing and long-term storage solutions are required.

Software requirements: Python is our primary language due to its extensive libraries for data science (pandas, NumPy, Matplotlib), machine learning (scikit-learn), deep learning (PyTorch), and web development (Django, Flask). We use PyTorch for developing and training AI models. NLP libraries such as NLTK and spaCy are required for tasks like tokenization, stemming, lemmatization, and other methods of preparing text for machine learning tasks. The project will also rely on various open-source Python libraries, specified in a requirements.txt file or a Pipenv/Poetry setup for easy installation. For LLMs and IR models, dependencies include pre-trained models and frameworks like Hugging Face's transformers library, freely available for research purposes. Web development frameworks are required for creating user interfaces to interact with the data and models; frameworks include Django (Python) or React (JavaScript). Database management systems like PostgreSQL are required to manage structured and unstructured data, respectively.

Documentation

For the software and algorithms we develop, we will produce extensive documentation, including installation guides, user manuals, and API documentation. All software will include a README file, offering a direct reference to the source code to which it refers. For each AI model developed, we will create detailed codebooks that describe the variables, algorithms used, and any modifications or training parameters. Documentation on analytical methods and procedural information will include detailed descriptions of the NLP and IR techniques employed, model training and evaluation procedures, and any software development practices. All documentation will be stored in open, non-proprietary formats such as Markdown, PDF, and plain text to ensure long-term accessibility and ease of use. We will apply clear, open licenses to documentation (e.g., Creative Commons) to clarify usage rights and access conditions.

Preservation

Virginia Tech has a strong commitment to the long-term preservation of research data through a comprehensive approach that includes institutional infrastructure, policies, and partnerships. The University Libraries maintains two preservation repositories, VTechWorks and the Virginia Tech Data Repository, which are central to its strategy for preserving and providing access to digital research outputs. VTechWorks, built on DSpace, offers a secure and accessible platform for the long-term preservation of a wide range of scholarly works, including articles and datasets. The Virginia Tech Data Repository, hosted on Figshare, is specifically tailored for the storage, sharing, and preservation of research data, facilitating compliance with open data mandates and enhancing the visibility of Virginia Tech's research. We will make use of these services for long term preservation of research outputs.

Review Frequency

This Data Management Plan (DMP) will undergo regular reviews and updates to ensure its relevance and effectiveness throughout the project lifecycle. The implementation of the DMP will be closely monitored by the PI and Project Director, with adjustments made as necessary to reflect new insights, technological advancements, and any changes in best practices or compliance requirements in data management. The DMP will be reviewed ***(1) initially:*** immediately after the project kickoff, to ensure all team members understand the DMP's requirements and their responsibilities; ***(2) quarterly:*** quarterly reviews will assess the plan's effectiveness, identify any issues in data management practices, and incorporate any necessary adjustments; ***(3) after major milestones:*** the plan will be reviewed following the completion of significant project phases or after achieving major milestones, ensuring that the DMP remains aligned with the project's current status and future direction; and ***(4) annually:*** an in-depth annual review will evaluate the DMP's comprehensive adherence to ethical guidelines, privacy concerns, and regulatory compliance. This will also be an opportunity to integrate any significant changes in data management technologies or standards.