Topic Modeling for Heterogeneous Digital Libraries: Tailored Approaches Using Large Language Models

Pradyumna Upendra Dasu

Thesis submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Master of Science
in
Computer Science and Applications

Edward A. Fox, Chair
Xuan Wang
Yinlin Chen

December 13, 2024
Blacksburg, Virginia

Keywords: Topic Modeling, Natural Language Processing, Large Language Models, Electronic Theses and Dissertations, Digital Libraries, Information Storage and Retrieval

Topic Modeling for Heterogeneous Digital Libraries: Tailored Approaches Using Large Language Models

Pradyumna Upendra Dasu

ABSTRACT

Digital libraries hold vast and diverse content, with electronic theses and dissertations (ETDs) being among the most diverse. ETDs span multiple disciplines and include unique terminology, making achieving clear and coherent topic representations challenging. Existing topic modeling techniques often struggle with such heterogeneous collections, leaving a gap in providing interpretable and meaningful topic labels. This thesis addresses these challenges through a three-step framework designed to improve topic modeling outcomes for ETD metadata. First, we developed a custom preprocessing pipeline to enhance data quality and ensure consistency in text analysis. Second, we applied and optimized multiple topic modeling techniques to uncover latent themes, including LDA, ProdLDA, NeuralLDA, Contextualized Topic Models, and BERTopic. Finally, we integrated Large Language Models (LLMs), such as GPT-4, using prompt engineering to augment traditional topic models, refining and interpreting their outputs without replacing them. The framework was tested on a large corpus of ETD metadata, including through preliminary testing on a small subset. Quantitative metrics and user studies were used to evaluate performance, focusing on the clarity, accuracy, and relevance of the generated topics. The results demonstrated significant improvements in topic coherence and interpretability, with user study participants highlighting the value of the enhanced representations. These findings underscore the potential of combining customized preprocessing, advanced topic modeling, and LLM-driven refinements to better represent themes in complex collections like ETDs, providing a foundation for downstream tasks such as searching, browsing, and recommendation.

Topic Modeling for Heterogeneous Digital Libraries: Tailored Approaches Using Large Language Models

Pradyumna Upendra Dasu

GENERAL AUDIENCE ABSTRACT

Digital libraries store vast information, including books, research papers, and electronic theses and dissertations (ETDs). ETDs are incredibly diverse, covering most academic fields and using highly specialized language. This diversity makes it challenging to create clear and meaningful summaries of the main themes within these collections. Our study addresses this challenge by developing a three-step framework and applying it to ETDs. First, we cleaned and standardized the data to make it easier to analyze. Second, we used advanced techniques to uncover patterns and group similar topics together. Finally, we improved these topics using powerful tools like GPT-4, which helped make the themes more precise, more accurate, and easier to interpret. We tested this framework on both a small and a large collection of ETDs. Combining quantitative evaluations and user feedback showed that our methods significantly improved how the topics represented the content. This work lays the foundation for more effective future tools to help people search, explore, and navigate large collections of academic works.

Dedication

Dedicated to my beloved parents, Mehar Vani Dasu and Aniruddha Suryanarayana Dasu, and to my lovely brother Sarat Dasu.

Acknowledgments

I would like to express my deepest gratitude to my advisor, Dr. Edward A. Fox, for his invaluable guidance, mentorship, and encouragement throughout my research journey. His insights and support have been crucial to the completion of this thesis. I am also immensely grateful to my committee members, Dr. Yinlin Chen and Dr. Xuan Wang, for their thoughtful feedback and expertise. Their perspectives and advice have significantly enriched my research, and I am fortunate to have had their support. I would like to thank the Institute of Museum and Library Services (IMLS) for funding our research through grants LG-37-19-0078-19, LG-256638-OLS-24, and RE-256655-OLS-24. Their support has made this work possible, and I sincerely appreciate their contribution. My labmates Satvik Chekuri, Dr. Bipasha Banerjee, Sareh Ahmadi, and Chenyu Mao have also been invaluable sources of support. They offered insightful discussions and technical help, making the research challenges more manageable. I am grateful for their presence and encouragement. Additionally, I am thankful to the Finance Information Technology team at Virginia Tech, where I had the opportunity to work as a graduate assistant. The support and encouragement I received from the team have been invaluable in balancing my responsibilities and advancing my academic goals. Finally, I extend my heartfelt appreciation to my family and friends for their unwavering encouragement and support. Their presence and motivation have been my anchor throughout this journey.

Contents

List of Figures

List of Tables

List of Abbreviations

AI Artificial Intelligence

API Application Programming Interface

BERT Bidirectional Encoder Representations from Transformers

c-TF-IDF Class-based Term Frequency-Inverse Document Frequency

CNN Convolutional Neural Network

CSV Comma-Separated Values

CTM Contextualized Topic Model

ETDs Electronic Theses and Dissertations

GMM Gaussian Mixture Model

GPT Generative Pre-trained Transformer

HDBSCAN Hierarchical Density-Based Spatial Clustering of Applications with Noise

IRB Institutional Review Board

KNN K-Nearest Neighbors

LDA Latent Dirichlet Allocation

LLM Large Language Model

ML Machine Learning

NER Named Entity Recognition

NeuralLDA Neural Latent Dirichlet Allocation

NLP Natural Language Processing

PCA Principal Component Analysis

PoS Part of Speech

ProdLDA Product of Experts Latent Dirichlet Allocation

RNN Recurrent Neural Network

SVD Singular Value Decomposition

TF-IDF Term Frequency-Inverse Document Frequency

UMAP Uniform Manifold Approximation and Projection

Chapter 1

Introduction

In an era where the volume of scholarly content—such as research papers, articles, and Electronic Theses and Dissertations (ETDs)—grows every decade, academic institutions face an unprecedented challenge: how can we ensure that the vast collections of knowledge stored in digital repositories are easily accessible and well-organized? Addressing this challenge aligns with priorities set forth by the Institute of Museum and Library Services (IMLS), which advocates for improved digital access and information management in libraries and archives across the United States of America [1].

Among the most valuable of these digital collections are theses and dissertations—scholarly documents that represent original research conducted by graduate students. Traditionally, these works were stored in physical formats, limiting their accessibility to the broader academic community. In 1997, Virginia Tech became the first university to mandate the electronic storage of these artifacts, significantly improving their global accessibility [2]. However, with the rapid expansion of digital content, managing and categorizing extensive text collections like ETDs has become increasingly challenging.

ETDs, which constitute a significant portion of academic repositories, offer rich sources of knowledge across diverse disciplines. Their interdisciplinary and often complex nature makes efficient content discovery within these collections critical yet difficult [3]. To address these challenges, scalable methods for thematic exploration are essential. This research emphasizes enhancing traditional topic modeling methods with Large Language Models (LLMs) to

improve semantic coherence and interpretability.

Topic modeling, an unsupervised machine learning technique, has emerged as a promising tool for uncovering latent themes in large text corpora [4, 5]. In the context of ETDs, topic modeling enhances the organization of academic works, enriches metadata, and offers new ways to explore scholarly themes. Furthermore, recent advances in language models—specifically LLMs such as GPT-4 [6], Llama-3.2 [7, 8], and Mixtral-8x7b [9]—have introduced advanced capabilities in semantic representation, coherence, and contextual understanding, providing new opportunities for improving topic modeling.

1.1 Motivation

Digital libraries (DLs) with heterogeneous content face significant challenges in providing effective browsing, searching, and discovery mechanisms. Topic modeling offers a data-driven approach to exploring thematic structures, potentially improving how users interact with these collections. This approach would particularly benefit researchers and students in identifying relevant works.

This research focuses on ETDs, which present unique challenges due to their interdisciplinary nature and specialized vocabularies. Since Virginia Tech's pioneering ETD mandate in 1997, many thousands of these documents have been collected across U.S. institutions [2]. Our study examines a representative subset of this corpus [10].

While effective at uncovering latent themes, traditional topic models often struggle with semantic richness and interpretability. This research aims to augment these models with LLMs to enhance topic representations while preserving the foundational framework of topic modeling, ultimately supporting interdisciplinary research and knowledge sharing.

1.2 Problem Statement

Despite the development of various topic modeling approaches, a key challenge remains: applying them effectively to large, diverse corpora like that of ETDs, to enhance thematic exploration and content comprehension. Improved topic representations could significantly benefit existing browsing, searching, and discovery methods, but implementing these improvements within digital libraries is complex.

Traditional keyword-based representations often fail to capture the semantic complexity of academic documents, frequently producing lists of terms that lack context and meaningful relationships. These representations can be particularly challenging to interpret in academic contexts where specialized terminology and interdisciplinary concepts are common. Moreover, traditional approaches may miss meaningful thematic connections beyond simple word co-occurrences.

Traditional topic models often struggle to generate coherent and meaningful topics in heterogeneous and complex datasets, such as ETDs. Additionally, many existing preprocessing techniques are not fully optimized for the complexity of academic text, which often involves specialized vocabularies and lengthy documents. LLMs present an opportunity to generate more contextually rich and semantically coherent topics by incorporating advanced semantic representations. However, LLMs are computationally intensive, and their scalability for use in large digital libraries has not been fully explored, posing challenges for practical implementation.

This research explores how tailored topic modeling methods, academic-specific preprocessing, and LLM integration can enhance topic coherence and representation in ETDs. Rather than replacing topic models, LLMs augment their outputs, addressing their limitations in handling semantic richness and interpretability.

1.3 Hypothesis

The following hypotheses guide this research:

- **H1:** Custom preprocessing techniques can improve state-of-the-art topic modeling performance applied to ETD corpora.
- **H2:** Transformer-based topic modeling approaches (like BERTopic) can achieve higher coherence scores than traditional probabilistic and neural approaches (i.e., LDA, ProdLDA, NeuralLDA, and CTM) when applied to ETD corpora.
- **H3:** Integration of Large Language Models can enhance topic interpretability while maintaining semantic alignment with original topics, as measured through embedding-based similarity metrics and user evaluations.

These hypotheses aim to test whether a combination of customized preprocessing, advanced topic modeling techniques, and LLM-based refinements can significantly enhance the thematic representation of ETDs. Validation is performed through quantitative metrics and user studies, ensuring the findings are robust and practically relevant.

1.4 Research Questions

This research addresses several key challenges in topic modeling, particularly in complex and diverse datasets like ETDs. The following research questions (RQs) guide the study:

- **RQ1:** How do custom preprocessing techniques influence the performance of various topic modeling architectures when applied to ETD corpora?

- **RQ2:** What are the comparative advantages and limitations of traditional, neural, and transformer-based topic modeling approaches for ETD collections?
- **RQ3:** How effectively can different LLMs enhance topic representations for ETD collections while maintaining semantic alignment with the original topics?

These research questions focus on understanding how preprocessing, topic modeling performance, and LLM integration can improve academic content representation and thematic exploration in digital libraries.

1.5 Contributions

This research offers the following contributions:

- **Preprocessing Framework:** Development and validation of domain-specific preprocessing techniques that improve topic modeling performance, demonstrated by significant gains in coherence scores across different architectures.
- **Comparative Model Analysis:** Systematic evaluation of topic modeling approaches revealing that BERTopic achieves superior coherence (CV: 0.746-0.778). At the same time, traditional LDA maintains better topic diversity (>0.9) but lower coherence, showing clear differences in model architectures.
- **LLM Integration Framework:** Novel approach combining topic models with LLMs, demonstrating improved topic interpretability without sacrificing computational efficiency.
- **Scalability Analysis:** Application of our framework on a large-scale dataset (333,867

documents), demonstrating high topic coherence (CV: 0.7417) and optimal topic diversity when the number of topics is 75.

- **Model Performance Analysis:** Comparative evaluation of topic representation capabilities across GPT-4, Llama 3.1, two variants of Llama 3.2, and Mixtral8x7b achieving semantic similarity scores between 0.758-0.829 for different representation tasks.
- **User-Centered Validation:** Preliminary user study findings (n=10) demonstrating a preference for LLM-enhanced topic representations over traditional keywords, providing initial validation of the framework's practical utility while highlighting the need for expanded user evaluation.

These contributions explore the potential for improved topic representations, which could lead to better thematic exploration and understanding of academic content in digital libraries. This research lays the groundwork for future developments.

1.6 Thesis Outline

The remainder of this thesis is structured as follows:

- **Chapter 2: Literature Review** examines the evolution of topic modeling approaches and LLMs in digital libraries, identifying current limitations and opportunities for enhancement.
- **Chapter 3: Data** describes the ETD corpus characteristics, emphasizing the complexities of academic text and the necessity for specialized preprocessing.
- **Chapter 4: Preliminary Analysis** investigates the limitations of traditional topic

representations in ETD collections, establishing baseline performance and identifying specific areas for improvement.

- **Chapter 5: Enhancement of Text for Topic Modeling** introduces our specialized preprocessing pipeline for academic content, including keyword extraction, metadata integration, and LLM-assisted enhancements.
- **Chapter 6: Development and Enhancement of Topic Modeling Approach** details implementing and evaluating our integrated topic modeling and LLM framework on the curated ETD metadata dataset.
- **Chapter 7: Results** presents comprehensive experimental findings with a discussion of implications, model performance analysis, and the potential impact on digital library applications.
- **Chapter 8: Conclusion and Summary** synthesizes key findings and contributions and proposes future research directions for topic modeling in digital libraries.

Chapter 2

Literature Review

2.1 Introduction

The rapid growth of digital academic content, particularly in repositories and Digital Libraries (DLs), has intensified the need for sophisticated information organization and retrieval techniques [11]. Topic modeling has emerged as a fundamental methodology for uncovering hidden thematic structures in extensive text collections, transforming how users interact with and discover digital content [12, 13]. This approach is particularly valuable for analyzing Electronic Theses and Dissertations (ETDs), where understanding research trends and thematic clusters across disciplines can provide invaluable insights into academic contributions [14].

In digital libraries, topic modeling enables enhanced access to poorly described texts through dynamic subject tagging and virtual collection creation [4, 5]. Combined with recent advances in Large Language Models (LLMs), these capabilities have substantially improved the generation of coherent, interpretable topics [15, 16]. Such enhancements support critical tasks like document recommendation and thematic cluster retrieval, particularly in analyzing ETDs and their metadata [17, 18].

This literature survey examines four critical aspects of topic modeling: (1) its evolution from classical approaches to modern transformer-based models, (2) the critical role of pre-

processing in academic texts, (3) evaluation methodologies spanning traditional metrics and emerging LLM-based approaches, and (4) the impact and integration of Large Language Models in topic modeling. This review identifies key research gaps and establishes the foundation for our proposed methodological advances that are discussed in subsequent chapters.

2.2 Evolution of Topic Modeling

Topic modeling has evolved substantially over three decades, progressing from fundamental statistical approaches to sophisticated neural architectures. This evolution reflects the increasing complexity of information retrieval challenges and the advancement of computational capabilities. Figure 2.1 illustrates the progression from classical foundations to modern transformer-based approaches, highlighting key methodological advances that have shaped the field.

2.2.1 Classical Foundations

Latent Semantic Indexing (LSI) [19] established the initial framework for topic modeling in 1990, introducing singular value decomposition to uncover latent semantic structures in text corpora. This approach provided the first systematic method for identifying hidden themes in document collections, though its deterministic nature limited flexibility across diverse text types. The introduction of Probabilistic LSI (PLSI) [20] enhanced this framework through probabilistic modeling, offering improved adaptability to different document types. However, PLSI encountered significant scalability challenges with extensive document collections, particularly in its parameter estimation process.

Latent Dirichlet Allocation (LDA) [21] emerged in 2003 as a fundamental advancement that

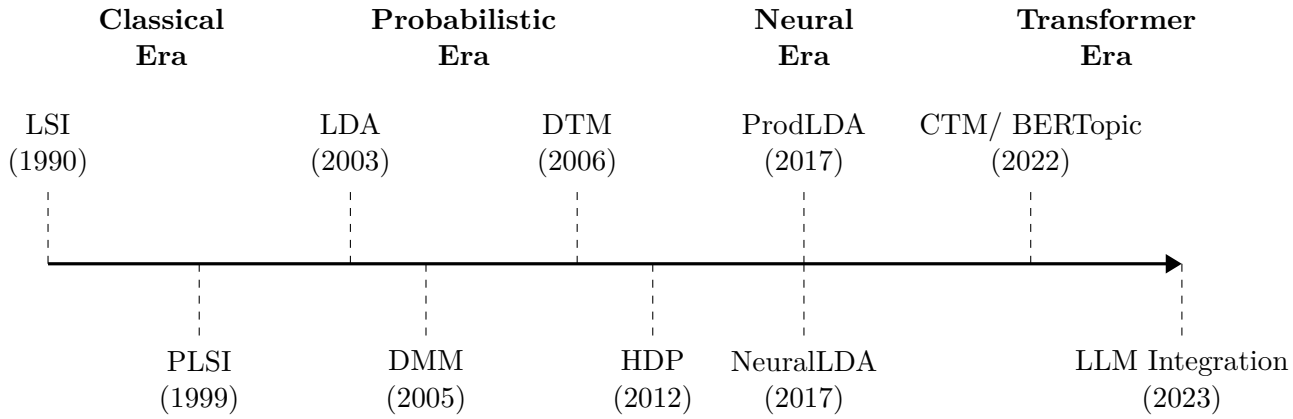

Figure 2.1: Evolution of Topic Modeling Approaches (1990–2023). This timeline highlights key developments in topic modeling, spanning four eras: Classical (e.g., LSI, 1990), Probabilistic (e.g., LDA, 2003), Neural (e.g., ProdLDA/NeuralLDA, 2017), and Transformer-based (e.g., BERTopic, 2022). It showcases the progression from foundational statistical methods to modern transformer-based techniques.

established topic modeling as a distinct methodology. LDA addressed PLSI's limitations through an innovative representation of documents as mixtures of topics via Dirichlet distributions. This probabilistic framework provided both theoretical rigor and practical flexibility, enabling more nuanced thematic analysis and serving as the foundation for numerous subsequent developments in the field.

2.2.2 Probabilistic Extensions

The success of LDA led to the development of increasingly sophisticated probabilistic models. The Dirichlet Multinomial Mixture (DMM) [22] enhanced text classification capabilities through refined probabilistic modeling, providing more accurate document categorization. The Hierarchical Dirichlet Process (HDP) [23] addressed a critical limitation in existing

models by introducing automatic topic number inference, eliminating the need for manual specification of topic quantities.

Dynamic Topic Models (DTM) [24] expanded analytical capabilities by enabling temporal topic evolution analysis. This advancement proved valuable for analyzing long-term text collections, allowing researchers to track thematic changes over time. Correlated Topic Models (CTM) [25] further enhanced the field by capturing meaningful relationships between topics, acknowledging that thematic structures in text rarely exist in isolation.

2.2.3 Neural and Transformer Approaches

The integration of deep learning methodologies, particularly Variational Autoencoders (VAEs), marked a significant advancement in topic modeling capabilities. The pioneering work by Srivastava and Sutton [26] introduced both NeuralLDA and ProdLDA, which leverage neural variational inference to exceed the capabilities of traditional probabilistic models. While NeuralLDA demonstrated how to neuralize traditional LDA using VAEs, ProdLDA advanced this framework by replacing the Dirichlet prior with a "Product of Experts" model. This modification substantially improved topic coherence and computational efficiency. Both models utilize the reparameterization trick in their VAE architecture to enable effective backpropagation through the sampling process, allowing for end-to-end training of the neural topic models.

The emergence of transformer architectures has fundamentally altered topic modeling approaches. Contextualized Topic Models [25] leverage BERT embeddings to capture nuanced contextual relationships between words, moving beyond the limitations of traditional bag-of-words approaches. BERTopic [27] further advances this direction by combining transformer embeddings with sophisticated clustering techniques. This combination generates highly co-

herent, context-aware topics that align more closely with human understanding of thematic structures.

2.2.4 Clustering Techniques for Topic Discovery

Clustering is pivotal in modern topic modeling, particularly in transformer-based methods like BERTopic [27]. These models combine embeddings with clustering techniques to group semantically similar documents, facilitating topic discovery. Several clustering approaches have demonstrated effectiveness in this domain:

- **K-means Clustering:** A centroid-based approach, k-means is known for its efficiency and scalability. It works effectively with dense embeddings generated by models like BERT and Sentence Transformers. However, it requires the number of clusters to be predefined, which can be limiting when the ideal topic count is unknown [28, 29].
- **HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise):** This density-based method excels in identifying clusters of varying shapes and densities, making it suitable for noisy datasets. Unlike k-means, HDBSCAN does not require specifying the number of clusters, as it dynamically determines cluster structures based on density connectivity [30].
- **Gaussian Mixture Models (GMM):** These probabilistic models assume data points are generated from a mixture of several Gaussian distributions. GMMs offer soft clustering capabilities, providing membership probabilities for each document across topics [31].
- **Agglomerative Hierarchical Clustering:** This bottom-up approach begins with individual documents as clusters and iteratively merges similar clusters. It produces a

hierarchical structure of topics, allowing for multiple granularity levels of analysis [32].

- **DBSCAN (Density-Based Spatial Clustering of Applications with Noise):** The predecessor to HDBSCAN, this method groups points that are closely packed together while marking points in low-density regions as noise. It's particularly effective for discovering clusters of arbitrary shapes [33].

Hybrid Approaches: Modern topic modeling frameworks often combine multiple clustering techniques to leverage their complementary strengths. For instance, some systems employ HDBSCAN for initial clustering followed by k-means for refinement or use hierarchical clustering with density-based methods. These hybrid approaches enhance topic coherence and diversity, particularly in high-dimensional embedding spaces [27].

2.2.5 Current Challenges and Opportunities

Contemporary topic modeling presents a trade-off between semantic coherence and computational efficiency. Modern approaches have substantially improved topic interpretability but often require considerable computational resources. Traditional models like LDA maintain their utility for processing large-scale collections, while neural approaches offer enhanced semantic understanding at higher computational costs.

This dichotomy presents opportunities for methodological innovation through the strategic combination of traditional efficiency with modern semantic richness. Hybrid approaches could leverage efficient traditional models for initial topic discovery while enhancing their interpretability through targeted post-processing. The field's ongoing development focuses on balancing computational feasibility with semantic sophistication, reflecting broader trends in Natural Language Processing [34, 35, 36, 37, 38].

The effectiveness of topic modeling approaches, regardless of their theoretical foundation, depends fundamentally on the quality of the input text. This dependence underscores the critical role of preprocessing in preparing text data for analysis [39, 40, 41], which the following section examines in detail.

2.3 Preprocessing in Topic Modeling

Text preprocessing is fundamental to the effectiveness of topic modeling, mainly when dealing with academic documents characterized by specialized vocabulary and complex structure. Research shows that preprocessing often demands more effort than the analysis itself [42], as the quality of generated topics relies heavily on preparing the input text.

2.3.1 Core Preprocessing Steps

Preprocessing begins with tokenization, decomposing text into discrete units or tokens that serve as the foundation for analysis [43]. Tokens can represent individual words or parts of words or multi-word expressions (n-grams) [44, 45, 46], with the choice affecting the granularity of discovered topics. This step is critical for ensuring the input text is suitable for subsequent modeling.

Standardization follows, ensuring consistency in text representation. This includes converting text to lowercase and removing punctuation, numbers, and special characters [43]. While standardization reduces noise and ensures compatibility across documents, special care must be taken with academic texts, where such elements may carry semantic importance.

Stopword removal is another key step in preprocessing, aimed at eliminating high-frequency terms that contribute little to thematic meaning. Generic stopword lists are often insufficient

for academic texts, as domain-specific filler words must be excluded without removing critical technical terms [47, 48]. Thus, tailored stopword lists are essential for academic contexts.

2.3.2 Advanced Processing Techniques

Normalization techniques like stemming and lemmatization play a significant role in reducing word forms to their roots or base forms, enhancing term consistency [49]. Stemming uses rule-based suffix removal, while lemmatization relies on morphological analysis to preserve meaning. In academic texts, careful selection between these approaches is crucial to maintaining the semantic integrity of technical terms.

Academic documents pose unique challenges, including the presence of citations, equations, and domain-specific terminology. Standard preprocessing pipelines may inadequately address these complexities, necessitating customized approaches that remove non-relevant elements while preserving critical content. Advanced text normalization techniques adapted to academic vocabulary and structure are essential for effective preprocessing in this domain [50, 51].

2.3.3 Influence on Topic Model Performance

Preprocessing decisions directly and profoundly impact topic model effectiveness [52]. Properly handled preprocessing enhances both topic coherence and diversity, while inadequate preprocessing can lead to noisy or incoherent topics. This is particularly evident in academic texts, where conventional techniques may fail to account for specialized vocabulary and structural nuances.

Empirical studies underscore the importance of preprocessing in reducing data dimensional-

ity, improving computational efficiency, and preserving semantic richness. However, overly aggressive preprocessing risks discarding meaningful information, especially in technical texts where specific terms and phrases carry significant weight. Thus, the balance between noise reduction and semantic preservation is critical [53, 54].

Evaluating the effectiveness of preprocessing strategies requires rigorous assessment of the resulting topics. The following section discusses various approaches to topic model evaluation, including traditional coherence metrics and modern techniques leveraging transformer-based word embeddings.

2.3.4 PDF-to-Text Tools

Extracting text from PDFs is a crucial preprocessing step, particularly for Electronic Theses and Dissertations (ETDs). This process involves converting complex document formats into plain text suitable for analysis. Several tools have demonstrated effectiveness in this domain:

- **PyPDF:** A lightweight Python library for extracting text from simple PDFs. While effective for well-structured documents, it may encounter limitations when processing PDFs containing images or complex layouts. The library offers improved performance and maintenance over its predecessor PyPDF2 [55].
- **Apache Tika:** A robust text extraction framework capable of handling various file formats, including PDFs with embedded multimedia content. Its content analysis toolkit provides comprehensive metadata extraction capabilities [56].
- **Tesseract OCR:** An open-source optical character recognition engine, particularly valuable when processing scanned PDFs or image-based text. It supports multiple languages and can handle complex document layouts [57].

The selection of appropriate extraction tools depends significantly on document structure and quality. In the context of ETDs, specific challenges include handling academic citations, mathematical equations, and document metadata. These tools, when properly configured, ensure accurate extraction of textual content, thereby providing a reliable foundation for subsequent preprocessing and topic modeling workflows.

2.4 Evaluation of Topic Models

Topic model evaluation presents unique challenges due to the unsupervised nature of the task and the complexity of assessing thematic coherence. Unlike supervised learning models with clear performance metrics, topic models require sophisticated evaluation approaches that consider both statistical validity and human interpretability [58, 59, 60]. The foundational work of Chang et al. [61] established that statistical measures alone may not correlate with human judgments of topic quality, necessitating more comprehensive evaluation frameworks.

2.4.1 Topic Coherence Metrics

Topic coherence measures have emerged as primary evaluation tools, assessing how semantically related the terms within each topic are to each other. Newman et al. [62] demonstrated that Pointwise Mutual Information (PMI) correlates well with human judgment when evaluating word pairs within topics. This approach was refined by introducing Normalized PMI (NPMI), which showed improved correlation with human ratings [63, 64].

Röder et al. [58] conducted comprehensive evaluations of coherence measures, developing new methods that outperform previous metrics. However, these measures are not without limitations. Research by Lau and Baldwin [65] revealed that coherence scores exhibit sen-

sitivity to the number of terms per topic, with correlation to human judgment typically decreasing as the number of terms increases.

2.4.2 Advanced Evaluation Approaches

Beyond basic coherence, modern evaluation frameworks incorporate multiple complementary metrics. Topic diversity measures assess how effectively models generate distinct topics across a corpus. Morstatter and Liu [66] introduced topic consensus metrics, evaluating the degree of agreement among human annotators regarding topic interpretations. Dieng et al. [16] proposed evaluating topics through intra-topic coherence and inter-topic separability, providing a more comprehensive assessment of topic quality.

Recent advances have leveraged automation to enhance evaluation processes. Bhatia et al. [67] developed automated topic intrusion detection, reducing reliance on manual evaluation tasks. Integrating Large Language Models (LLMs) has further advanced automated evaluation, with Stammbach et al. [68] demonstrating their effectiveness in intrusion detection tasks.

2.4.3 Emerging Evaluation Challenges

The evolution of topic modeling techniques, particularly those incorporating LLMs, presents new evaluation challenges. Traditional coherence metrics, designed for keyword-based topics, may not adequately assess topics generated through advanced language models. Ding et al. [69] demonstrated that word-embedding-based coherence metrics can effectively complement traditional approaches, suggesting promising directions for evaluating semantically enhanced topics.

The evaluation landscape continues to evolve, particularly in specialized domains like academic content, where term relationships exhibit complex hierarchical and semantic structures. While coherence measures remain fundamental, integrating LLMs necessitates new evaluation approaches to assess the mathematical soundness and semantic meaningfulness of enhanced topic representations [68, 70, 71]. The following section examines how large language models are transforming topic modeling while creating new opportunities and challenges for evaluation.

2.5 Large Language Models in Topic Modeling

Large Language Models (LLMs) have fundamentally transformed Natural Language Processing (NLP), introducing unprecedented capabilities for understanding and generating text. This transformation extends to topic modeling, where LLMs offer enhanced semantic understanding and improved topic coherence [72]. The development of these models represents a significant advancement from traditional statistical approaches to sophisticated neural architectures capable of capturing complex linguistic patterns.

2.5.1 Architectural Evolution

The transformer architecture [73] marks a pivotal development in LLM evolution, enabling efficient parallel computation and improved handling of long-range dependencies in text. This architecture underpins models like BERT [74], revolutionizing natural language understanding through contextual word representations. Subsequent developments, including OpenAI's GPT series [6, 75, 76] and Google's PaLM [77], have further refined these capabilities, surpassing earlier approaches based on Recurrent Neural Networks [78], LSTM [79],

and GRU [80].

Recent models such as GPT-4, LLaMA-3, and Mistral-8x7B demonstrate remarkable versatility across diverse NLP tasks [6, 8, 81]. Their effectiveness correlates strongly with model scale, following empirically observed scaling laws [82, 83, 84]. This scaling enables increasingly sophisticated language understanding, though it also introduces significant computational demands [85, 86].

2.5.2 Integration with Topic Modeling

Integrating LLMs into topic modeling frameworks has substantially improved topic coherence and interpretability [87]. These models enhance various aspects of the topic modeling pipeline, from document representation to topic interpretation. LLM embeddings capture intricate relationships between words and phrases, improving the clustering and topic discovery processes [88]. LLMs generate human-readable descriptions in topic labeling that clarify the meaning of topic clusters [71, 89].

However, this integration presents significant challenges. Privacy concerns arise from LLMs' potential to memorize training data [90], and ethical considerations emerge regarding biases in model outputs [91, 92]. These challenges necessitate careful consideration in academic and research applications [93].

2.5.3 Infrastructure and Accessibility

Recent developments in LLM infrastructure have enhanced model accessibility and efficiency. The Hugging Face platform provides extensive pre-trained model access and integration tools [94], while Groq's innovative Tensor Streaming Processor architecture [95] and inference

platform [96] demonstrate significant advances in processing efficiency.

Table 2.1 provides a comparison of several state-of-the-art LLMs used in this study, highlighting their characteristics, strengths, and capabilities. These models include GPT-4o, a proprietary model optimized for complex reasoning tasks, and open-source alternatives such as the Llama-3 variants and Mixtral-8x7B, which balance performance and efficiency. The details emphasize critical attributes like context length, training data size, and specific strengths, offering insights into their suitability for diverse NLP tasks, including topic modeling.

These infrastructural improvements benefit topic modeling applications, enabling more sophisticated topic analysis while managing computational constraints. Table 2.1 highlights the trade-offs between proprietary and open-source models, illustrating how efficient architectures like Mixtral-8x7B and resource-scaled options like Llama-3.2-3B democratize access to advanced NLP capabilities. Careful consideration of computational resources and model selection remains crucial [97, 98].

2.5.4 Future Directions

The continuing evolution of LLMs suggests several promising directions for topic modeling. Developing more efficient architectures and improved pre-training approaches may address current computational limitations. Additionally, emerging techniques for reducing model bias and ensuring privacy compliance will likely influence how LLMs are integrated into topic modeling applications, particularly in academic and research contexts.

The intersection of LLMs and topic modeling represents a dynamic area of research, balancing enhanced semantic understanding with practical implementation challenges. Future developments will likely focus on optimizing this balance, potentially through hybrid ap-

Table 2.1: Comparison of Large Language Models

Model	Characteristics
GPT-4o [6]	• Developer: OpenAI • Version: GPT-4o (2024-08-06) • Architecture: Advanced transformer-based model with multimodal capabilities • Context Length: 128K tokens • Output Limit: 16,384 tokens • Training Data: Up to October 2023 • Key Strengths: Complex multi-step tasks, reasoning • Notable Features: Optimized for better speed and cost efficiency compared to GPT-4 Turbo • Availability: Proprietary
Llama 3.1-70B [8]	• Developer: Meta • Architecture: Transformer with GQA • Parameters: 70B • Context Length: 128k tokens • Key Strengths: General performance, code generation • Training Data: 15T+ tokens up to December 2023 • Availability: Open source
Llama 3.2-90B-vision [7]	• Developer: Meta • Architecture: Transformer with GQA • Parameters: 90B • Context Length: 128K tokens • Key Strengths: Dialogue tasks, general tasks, vision tasks • Training Data: 6B image and text pairs, up to December 2023 • Availability: Open source
Llama 3.2-3B [7]	• Developer: Meta • Architecture: Standard Transformer • Parameters: 3B • Context Length: 128K tokens • Key Strengths: Resource-efficient, edge deployment • Training Data: Up to 9 trillion tokens of data from publicly available sources. • Availability: Open source
Mixtral-8x7B [81]	• Developer: Mistral AI • Architecture: Sparse MoE • Parameters: Each input activates two of eight experts, combining outputs to use 47B parameters efficiently with only 13B during inference. • Context Length: 32K tokens • Key Strengths: Technical tasks, fast inference • Training Data: Not disclosed • Availability: Open source

proaches that combine the semantic richness of LLMs with the computational efficiency of traditional topic modeling methods.

2.6 Tools and Libraries

This section highlights the Python libraries, tools, and frameworks utilized for implementing topic modeling, embedding generation, keyword extraction, preprocessing, and evaluation. These tools are integral to modern Natural Language Processing (NLP) pipelines, ensuring state-of-the-art performance and reproducibility.

2.6.1 Text Preprocessing Libraries

Preprocessing academic documents, such as ETDs, is essential for ensuring clean and structured input for topic modeling. The following libraries were employed:

- **NLTK:** Used for tokenization, stopword removal, and text normalization tasks. It provided a robust set of tools for linguistic preprocessing [99].
- **regex:** Enabled pattern matching and cleaning of text, such as removing unwanted characters, citations, and other non-linguistic elements from the input data [100].

2.6.2 Embedding Models and Frameworks

Embedding models were central to this research, transforming text into dense vector representations for topic modeling, keyword extraction, and evaluation. The following frameworks and models were employed:

- **`sentence-transformers`:** Used to generate document embeddings. Models such as `all-mpnet-base-v2` [101] were employed for keyword extraction, while `cde-small-v1` [102] was selected for enhancing topic coherence in BERTopic. Additionally, `gte-large-en-v1.5` [103] was accessed through `sentence-transformers` for evaluating LLM-based topic representations.
- **`Transformers`:** Enabled the use of transformer-based embeddings, including models like BERT and custom variants, for contextual text representation [104].

The embedding models were selected based on their performance on the MTEB (Massive Text Embedding Benchmark) leaderboard [105]. MTEB ranks embedding models on tasks like clustering, retrieval, and semantic textual similarity. Models like `all-mpnet-base-v2` and `cde-small-v1` demonstrated competitive performance across diverse datasets, making them suitable for this research.

2.6.3 Topic Modeling Libraries

The topic modeling frameworks employed in this research spanned classical, neural, and transformer-based approaches:

- **Gensim:** Implemented Latent Dirichlet Allocation (LDA) for baseline topic modeling and was used for evaluation metrics like perplexity [106].
- **OCTIS:** Supported neural topic modeling, including NeuralLDA, ProdLDA, and Contextualized Topic Models (CTM), and provided an environment for evaluating coherence and diversity [107].
- **BERTopic:** Combined transformer-based embeddings with clustering techniques like k-means and HDBSCAN [108] to generate coherent and interpretable topics [27].

2.6.4 Keyword Extraction and Evaluation Libraries

Libraries for keyword extraction and evaluation were selected to ensure the generation of interpretable and semantically meaningful topics:

- **KeyBERT:** Utilized embeddings from `sentence-transformers` to extract top keywords from each document's abstract, providing concise and contextually relevant summaries [109].
- **scikit-learn:** Provided essential tools for clustering, preprocessing, and evaluation metrics such as topic coherence and diversity [110].
- **Gensim and OCTIS:** Both libraries were employed to evaluate topic models using traditional and advanced metrics, including coherence, diversity, and perplexity [106, 107].

2.6.5 LLM Integration Tools

Large Language Models (LLMs) were integrated to enhance topic interpretability and improve topic representations:

- **OpenAI GPT-4:** Accessed via API to generate topic labels, keyphrases, and descriptions, significantly improving the interpretability of topic modeling outputs [6].
- **Groq Playground:** Used for executing inference on open-source models, including Llama-3 and Mistral, providing a robust and efficient platform for experimentation with large models [96].

This diverse set of tools and libraries ensured the efficient implementation of preprocessing, modeling, and evaluation workflows. By leveraging state-of-the-art frameworks like Gen-

sim, OCTIS, and BERTopic, alongside advanced embedding models ranked on the MTEB leaderboard, this research achieved high-quality and reproducible results in topic modeling and keyword extraction.

2.7 Summary

This chapter has traced the evolution of topic modeling from its foundational methods to current approaches leveraging neural and transformer-based architectures. Through this examination, several key challenges and opportunities have emerged in applying topic modeling to academic texts, particularly in the context of digital libraries.

The field presents a notable dichotomy between traditional and modern approaches. While conventional topic modeling methods like LDA have proven valuable for processing extensive document collections, they often produce keyword-based representations that lack interpretability, particularly in academic contexts. Conversely, newer approaches leveraging neural architectures and LLMs offer enhanced semantic understanding but face computational challenges when applied to extensive document collections. This creates an opportunity for hybrid approaches that strategically combine traditional methods' efficiency with modern techniques' semantic richness.

The preprocessing of academic texts, especially ETDs, presents unique challenges that general-purpose text processing methods do not adequately address. Academic documents' specialized vocabulary and complex writing patterns require more sophisticated preprocessing approaches. Current methods often fail to properly handle domain-specific terminology and multi-word technical phrases properly, suggesting the need for specialized preprocessing pipelines tailored to academic content [50, 51].

The evaluation of topic model outputs becomes increasingly complex when dealing with enhanced representations beyond simple keyword lists. While traditional coherence metrics provide valuable insights, they may not fully capture the quality of topics enhanced by LLMs or other advanced techniques. This indicates the need for new evaluation frameworks to assess semantic coherence and practical utility in academic contexts.

While promising, integrating large language models with topic modeling requires careful consideration of computational feasibility. Simply applying LLMs to individual documents becomes prohibitively expensive at the scale of digital libraries. This suggests an opportunity for developing strategic integration approaches that enhance topic interpretability while maintaining computational efficiency.

These challenges—balancing computational efficiency with interpretability, addressing the unique preprocessing needs of academic texts, and developing evaluation frameworks for enhanced topic representations—form the basis for the contributions presented in this thesis. The methodologies and experiments presented in subsequent chapters directly address these challenges.

Finally, while promising, integrating large language models with topic modeling requires careful consideration of computational feasibility. Simply applying LLMs to individual documents becomes prohibitively expensive at the scale of digital libraries. This suggests an opportunity for developing strategic integration approaches that enhance topic interpretability while maintaining computational efficiency.

Chapter 3

Data

3.1 Introduction

Electronic Theses and Dissertations (ETDs) represent a complex and valuable corpus for topic modeling, offering diverse metadata fields such as titles, abstracts, and disciplinary classifications. While ETD titles and abstracts generally follow standardized formats, the subject-specific terminology and varying depth of metadata elements across disciplines and institutions present challenges for computational analysis. These challenges are further compounded by the academic language used in ETDs, which often includes specialized vocabulary and intricate thematic connections across disciplines.

This chapter provides an overview of the datasets used in this research, focusing on their selection, characteristics, and relevance to the study's goals. We leverage two complementary ETD datasets—a large corpus of over 500,000 documents for scalability assessment and a curated subset of 9,400 documents for detailed methodological development. Additionally, we incorporate the widely used 20 Newsgroups dataset as provided by the scikit-learn Python library [110, 111] to benchmark the complexity of ETD metadata against general-purpose text.

By examining the unique linguistic and structural features of these datasets, we establish the foundation for our topic-modeling approach and demonstrate the need for customized

methods to address the challenges of academic text. This chapter also includes a comparative analysis of dataset characteristics, highlighting the specific complexities of ETD metadata that influence topic modeling performance.

3.2 Dataset Overview and Selection

3.2.1 Dataset Selection Strategy

Our research employed an iterative approach using three complementary datasets:

- **Large ETD Dataset**: Over 500,000 ETDs from 42 U.S. universities were used for initial feasibility assessment and final scalability validation [10].
- **Curated Dataset**: 9,400 ETDs with comprehensive metadata, used for detailed methodology development [112].
- **20 Newsgroups Dataset**: 18,846 documents across 20 categories, serving as a complexity benchmark provided by the scikit-learn library, accessible via the dataset module [110, 111].

3.2.2 Selection Rationale

The selection and use of datasets in this research followed three distinct phases: initial exploration, method development, and validation with scalability testing. Each phase addressed specific objectives and built upon the findings of the preceding stage, guided by insights from the preliminary analysis (that is discussed in Chapter 4).

1. **Initial Exploration**: The Large ETD Dataset, with metadata for over 500,000 theses and dissertations from 42 U.S. universities [10], was used to assess the potential of ETD metadata for topic modeling and to identify computational and methodological challenges. Its size and diversity provided valuable insights into academic metadata's heterogeneity.

2. **Method Development**: The Curated Dataset, comprising 9,400 ETDs [112], was selected for detailed experimentation due to its manageable size, comprehensive metadata coverage, and balanced disciplinary representation. This dataset facilitated the refinement of preprocessing techniques and modeling strategies (that are discussed in Chapters 5 and 6).

3. **Validation and Scaling**: The Large ETD Dataset was revisited to validate the developed methods and assess scalability across a larger corpus.

The 20 Newsgroups Dataset, included for comparative analysis in this chapter, served as a benchmark for contextualizing the linguistic and structural complexity of ETD metadata against general-purpose text. While not used for modeling experiments, it highlighted the unique challenges of academic content.

3.3 Primary Datasets

3.3.1 Curated Dataset

The Curated Dataset provides metadata for 9,400 ETDs across diverse academic fields. Tables 3.1 and 3.2 detail its schema and vital statistics. The dataset's distribution across departments (Figure 3.1) and decades (Figure 3.2) demonstrates comprehensive coverage of academic disciplines, essential for robust topic modeling.

The basic statistics as shown in table 3.2 reveal a dataset spanning nearly a century (1930–

2021) across 12 universities. The classification structure reflects multiple levels of academic organization: from 230 distinct disciplines at the most granular level, to 69 departments and 47 updated ETD departments, consolidating into 46 standardized ProQuest departments with corresponding numerical codes. These are further grouped into 19 general academic labels and 3 broad categories. The single school type indicates uniform institutional classification, while the near-even distribution between STEM (5,000) and non-STEM (4,400) disciplines ensures balanced representation across academic domains. The presence of 80 unique degrees reflects the diversity of academic programs captured in the dataset.

Table 3.1: Database Schema for Curated Dataset

Column Name	**Data Type**	**Description**
id	INTEGER	Primary key identifier
title	TEXT	Title of the academic work
year	INTEGER	Year of publication
abstract	TEXT	Abstract content of the work
university	TEXT	Name of the university
degree	TEXT	Degree program name
URI	TEXT	Handle/URL identifier
department	TEXT	Department name
discipline	TEXT	Academic discipline
school	TEXT	School or college name
type	TEXT	Type of document (e.g., REGULAR)
Updated_ETD_Depts	TEXT	Updated departments information
Stem_NonStem	TEXT	Classification as STEM or non-STEM
ProQuest_Depts	TEXT	ProQuest department classifications
Code	TEXT	Department/discipline codes (e.g., 544,464)
Label	TEXT	General field label (e.g., ENGINEERING)
Category	TEXT	Broad academic category

Table 3.2: Basic Statistics of the Curated Dataset

Statistic	**Value**
Year Range	1930 to 2021
Unique Universities	12
Unique Degrees	80
Unique Departments	69
Unique Disciplines	230
Unique School Types	1
Unique Updated ETD Departments	47
STEM/Non-STEM Categories	STEM: 5,000, Non-STEM: 4,400
Unique ProQuest Departments	46
Unique Codes	46
Unique Labels	19
Unique Categories	3

3.3.2 Large ETD Dataset

The Large ETD dataset comprises metadata for over 500,000 ETDs collected from more than 42 U.S. universities [10]. The collection was automated through an ETD Ingestion Framework (EIF), which harvests metadata via OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting). The data is stored in a MySQL database with PDFs organized in a hierarchical repository hosted by Old Dominion University and mirrored at Virginia Tech. Table 3.3 shows the database schema, which includes essential metadata fields such as title, author, abstract, and publication year, along with administrative data like unique identifiers and creation timestamps.

The original dataset comprised nearly 500,000 Electronic Theses and Dissertations (ETDs). Initial cleaning was guided by the missing values analysis presented in Table 3.4, which identified fields such as *Department* and *Discipline* as having a significant number of missing entries. Following this step, additional cleaning addressed rows with inappropriate information, such as HTML tags and special characters, which were prevalent across various fields.

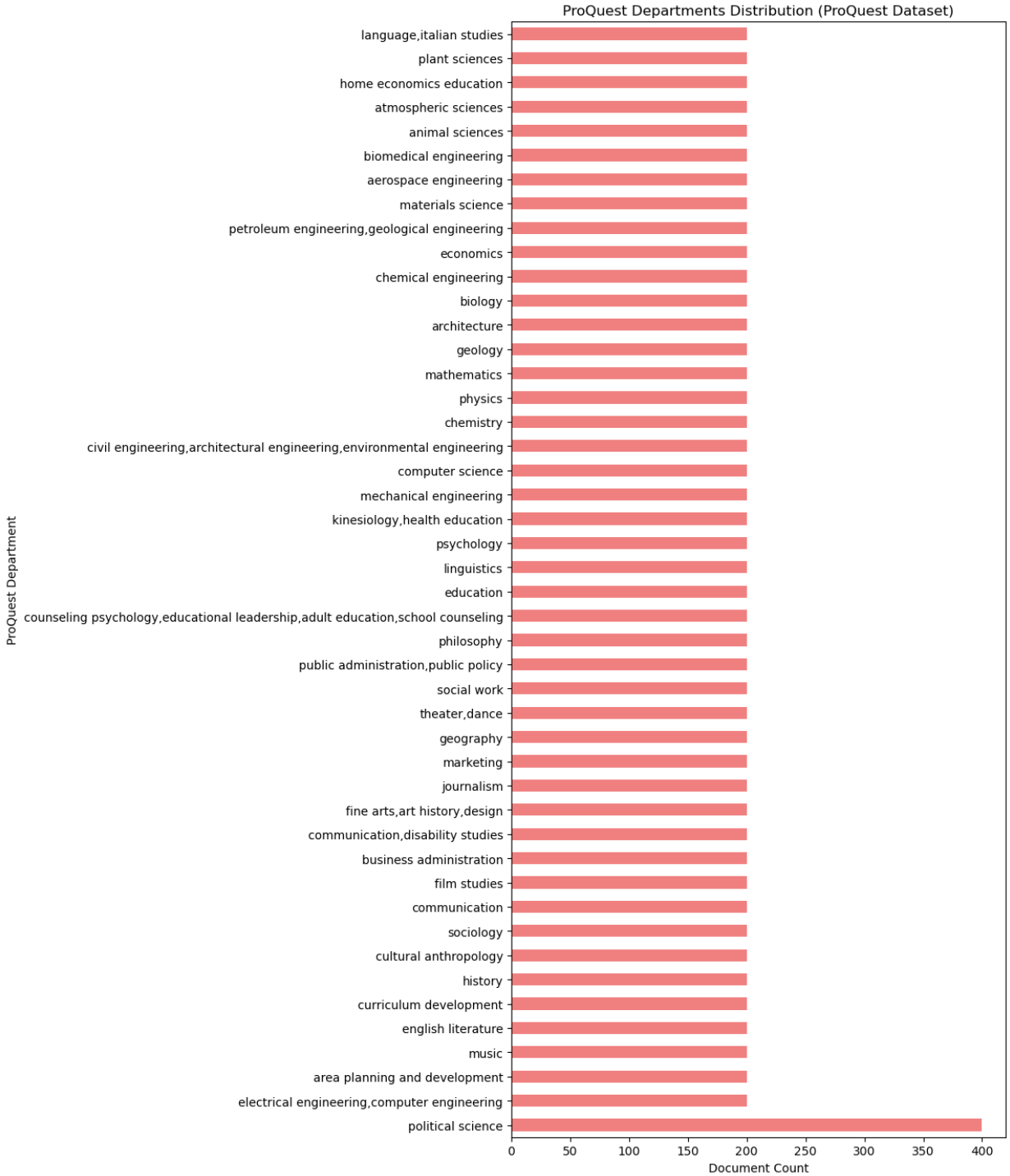

Figure 3.1: ProQuest Department Distribution (Curated Dataset)

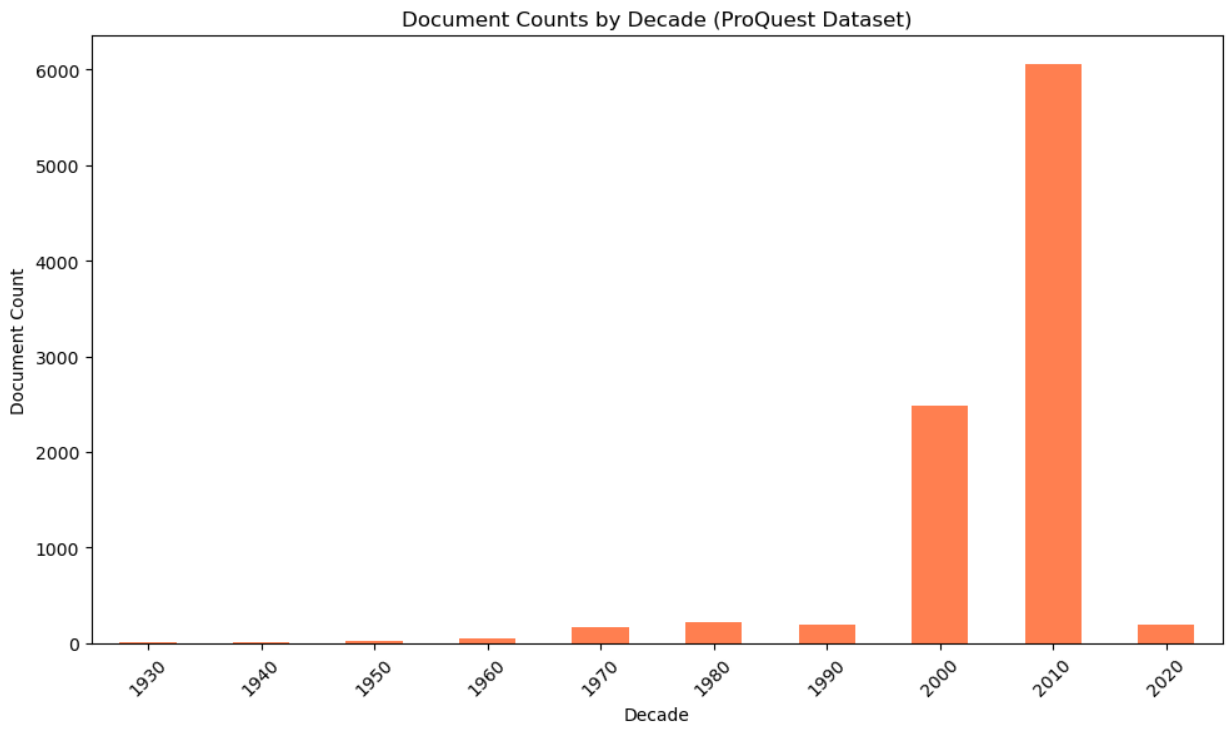

Figure 3.2: Document Counts by Decade (Curated Dataset)

Table 3.3: Database Schema for Large ETD Dataset

Column Name	**Data Type**	**Description**
id	INTEGER	Primary key identifier for the ETD record
title	TEXT	Full title of the thesis or dissertation
parent_id	INTEGER	Foreign key reference to parent record
author	TEXT	Full name of the thesis/dissertation author
abstract	TEXT	Complete abstract text of the thesis/dissertation
year	INTEGER	Year of publication/completion
URI	TEXT	Unique resource identifier/URL to access the document
language	VARCHAR(3)	ISO language code of the document (e.g., 'eng')
created_at	DATE	Record creation timestamp in the database

After this process, we identified additional rows containing HTML tags and special characters that lacked sufficient information in the *Title* and *Abstract* fields to support meaningful topic modeling. These rows were subsequently removed, resulting in a refined dataset of 333,867 records. This refined dataset provides a robust foundation for downstream analysis, with

key fields cleaned and curated for usability.

Table 3.4: Data Quality Statistics of the Large ETD Dataset

Field	Missing Count
Year	65,955
Advisor	112,748
Department	232,653
Discipline	166,690
Subjects	52,579
Abstract	99,583

Source: Adapted from Uddin et al. [10]

3.3.3 20 Newsgroups Benchmark Dataset

The 20 Newsgroups Dataset contains 18,846 documents distributed across 20 distinct categories, naturally formed from Usenet newsgroups where the messages were originally posted [110, 111]. These categories (e.g., 'comp.graphics', 'rec.sport.baseball', sci.med') reflect the structure of the newsgroups, where users self-categorized their posts. This dataset contrasts with Electronic Theses and Dissertations (ETDs), which use specialized academic language, by offering general-purpose text and straightforward categorization. Its simplicity and structure make it a widely-used benchmark for text classification, topic modeling, and other natural language processing tasks.

Table 3.5: Database Schema for 20 Newsgroups Dataset

Column Name	Data Type	Description
`text`	TEXT	Contains the content of a post or document
`group`	VARCHAR	Indicates the newsgroup category of a post

3.4 Dataset Characteristics and Complexity

Our analysis revealed several key distinctions between ETD and general-purpose datasets as seen from Table 3.6. It highlights the unique challenges of academic text processing.

Table 3.6: Comparison of Basic Textual Statistics and Readability

Statistic	Curated	Large	20 Newsgroups
Number of Documents	9,400	333,867	18,846
Vocabulary Size	57,560	414,539	93,622
Avg. Document Length	176.20	178.05	143.90
Median Document Length	181.00	168.00	88.00
Std. Dev. Document Length	89.58	96.83	281.70
Avg. Sentence Length	15.25	14.69	9.80
Avg. Word Length	7.50	7.51	6.05
Flesch-Kincaid Score	15.39	15.12	8.34

Vocabulary Characteristics: The ETD datasets ('Curated' and 'Large') demonstrate extensive vocabulary sizes, with the Large dataset exceeding 400 thousand unique words. This indicates the breadth of technical and disciplinary language. In contrast, the '20 Newsgroups' dataset, while containing a sizable vocabulary, focuses on general topics with possibly less specialized terminology.

Document Structure: The ETD datasets exhibit consistent structural patterns, with average document lengths close to 200 words and relatively low standard deviations. This reflects the formal, standardized nature of academic writing. In contrast, '20 Newsgroups' shows much higher variability, with a median length of 119 words and a large standard deviation, indicative of its less formal and diverse content.

Language Complexity: The academic rigor of the ETD datasets is evident in their high Flesch-Kincaid [113] scores (above 15), suitable for graduate-level readers. This complexity is driven by intricate sentence structures and specialized vocabulary. By comparison, '20 Newsgroups', with a Flesch-Kincaid score of 8.34, employs simpler language designed for a

general audience.

These differences in vocabulary, structure, and complexity necessitated the development of specialized techniques to handle academic text effectively.

3.5 Summary

This chapter introduced the datasets used in this research, highlighting the unique challenges posed by Electronic Theses and Dissertations (ETDs). The Curated Dataset, with its rich metadata and disciplinary breadth, was used for detailed method development, while the Large ETD Dataset dataset validated the scalability of the approaches. The 20 Newsgroups Dataset served as a benchmark for underscoring the higher complexity of academic texts in terms of vocabulary richness, structure, and linguistic intricacy.

The comparative analysis revealed the limitations of traditional preprocessing techniques and topic modeling algorithms in addressing the heterogeneity and complexity of ETD data. These observations informed the need for tailored methodologies capable of handling academic text effectively. The next chapter presents the findings from a preliminary analysis, which examined traditional approaches and identified critical shortcomings, setting the stage for developing more effective solutions.

Chapter 4

Preliminary Analysis

4.1 Overview

The CS 5604 [114] course (Fall 2023) at Virginia Tech, guided by Dr. Edward A. Fox, initiated a project to develop an ETD-specific information retrieval system. In connection with this class, the research reported in this thesis began, with an exploratory phase that addressed key challenges in topic modeling, including enhancing topic coherence, improving interpretability, and refining preprocessing methods. Insights from this phase informed the methodology in later stages.

4.2 Initial Dataset and Cleaning

The initial ETD corpus consisted of approximately 500,000 records spanning the years 1845 to 2020, including metadata such as titles and abstracts [10]. However, the dataset evidenced several quality issues, such as missing values, placeholders, and improperly formatted fields. A systematic cleaning process addressed these issues, removing irrelevant entries and ensuring that metadata fields like titles and abstracts were retained. The cleaned dataset comprised 333,867 records, creating a reliable subset for experimentation.

4.3 Model Testing and Evaluation

4.3.1 BERTopic with HDBSCAN Clustering

BERTopic, which uses BERT embeddings for contextual understanding, was tested with HDBSCAN [108] for clustering. While BERTopic captured nuanced semantic relationships, HDBSCAN over-consolidated clusters, reducing thematic granularity, and leading to a significant "outlier" topic with over 169,000 documents. These issues necessitated exploring alternative clustering methods, such as k-means [115], to improve topic differentiation and coherence.

4.3.2 Latent Dirichlet Allocation (LDA)

LDA, a probabilistic model based on word frequencies, was tested to compare with BERTopic's embedding-based approach. LDA generated balanced topic distributions, representing 73 distinct topics with clear keywords. While its topics were interpretable and practical for academic search, LDA struggled to capture the nuanced relationships identified by BERTopic.

4.3.3 Key Insights from Model Testing

The initial experiments highlighted strengths and limitations in both approaches. BERTopic's contextual embedding captured semantic nuances but required improved clustering methods to enhance granularity. LDA provided interpretable topics but lacked the semantic depth of embedding-based methods. These findings motivated refinements, including testing k-means clustering with BERTopic.

4.4 Pilot Study and System Usability Analysis

A pilot study carried out as part of the CS5604 course assessed the usability of a topic modeling toolkit for navigating ETD topics generated by BERTopic with HDBSCAN clustering.

4.4.1 Study Objectives and Tasks

The study aimed to evaluate system usability and the interpretability of BERTopic-generated topics for academic search. Participants performed two tasks: 1) exploring ETD topics to assess thematic navigation and 2) conducting keyword searches to evaluate accuracy and relevance.

4.4.2 System Usability Scale Evaluation

The pilot study, conducted from November 20, 2023, to December 9, 2023, included 11 participants and achieved a System Usability Scale (SUS) score of 77.5, indicating a "Good" usability rating, i.e., above the industry average of 68.

4.4.3 Key Findings and Recommendations

The system excelled in ease of use and learnability, with 90.9% of participants rating it above industry standards. However, feedback emphasized the need for more interpretable topic labels and improved search accuracy. Key recommendations included:

- Replacing generic keywords with descriptive topic labels.
- Refining clustering algorithms for better topic differentiation.

- Enhancing system consistency and onboarding support for new users.

4.5 Key Insights and Next Steps

This exploratory phase underscored several key challenges and opportunities:

- **Granular Topic Representation:** The need for more granular clustering techniques led to testing k-means clustering with BERTopic.
- **Enhanced Preprocessing:** The complexity of ETD metadata required advanced preprocessing strategies, including improved stopword removal and keyword extraction guiding enhanced input for topic models.
- **User-Driven Refinements:** User feedback informed refinements to topic representations and system design.

These findings informed the methodologies and evaluations presented in subsequent chapters. The next chapter focuses on preparing data for topic modeling, highlighting the importance of noise reduction, metadata enrichment, and text normalization techniques to improve topic coherence and interpretability.

Chapter 5

Enhancing Data for Topic Modeling

5.1 Introduction

Preprocessing is critical in transforming raw text data into a format suitable for further computational analysis. For Electronic Theses and Dissertations (ETDs), this process requires specialized techniques to handle the unique characteristics of academic writing, such as complex sentence structures and domain-specific terminology. The preprocessing pipeline developed for this research significantly improves topic coherence (10% higher than that with basic preprocessing) by systematically cleaning, normalizing, and enhancing the text.

The pipeline consists of modular steps, including text normalization, tokenization, LLM-assisted stopword removal, lemmatization, and keyword extraction. These stages ensure that the processed text captures the richness of academic content while reducing noise and redundancy. Additionally, a novel **Enhanced Text** representation combines key metadata fields with extracted keywords, offering a holistic input format for topic modeling.

Figure 5.1 provides an overview of the pipeline, illustrating the sequential transformation of ETD data. The following subsections describe each step in detail, emphasizing its contribution to preparing high-quality inputs for topic modeling tasks.

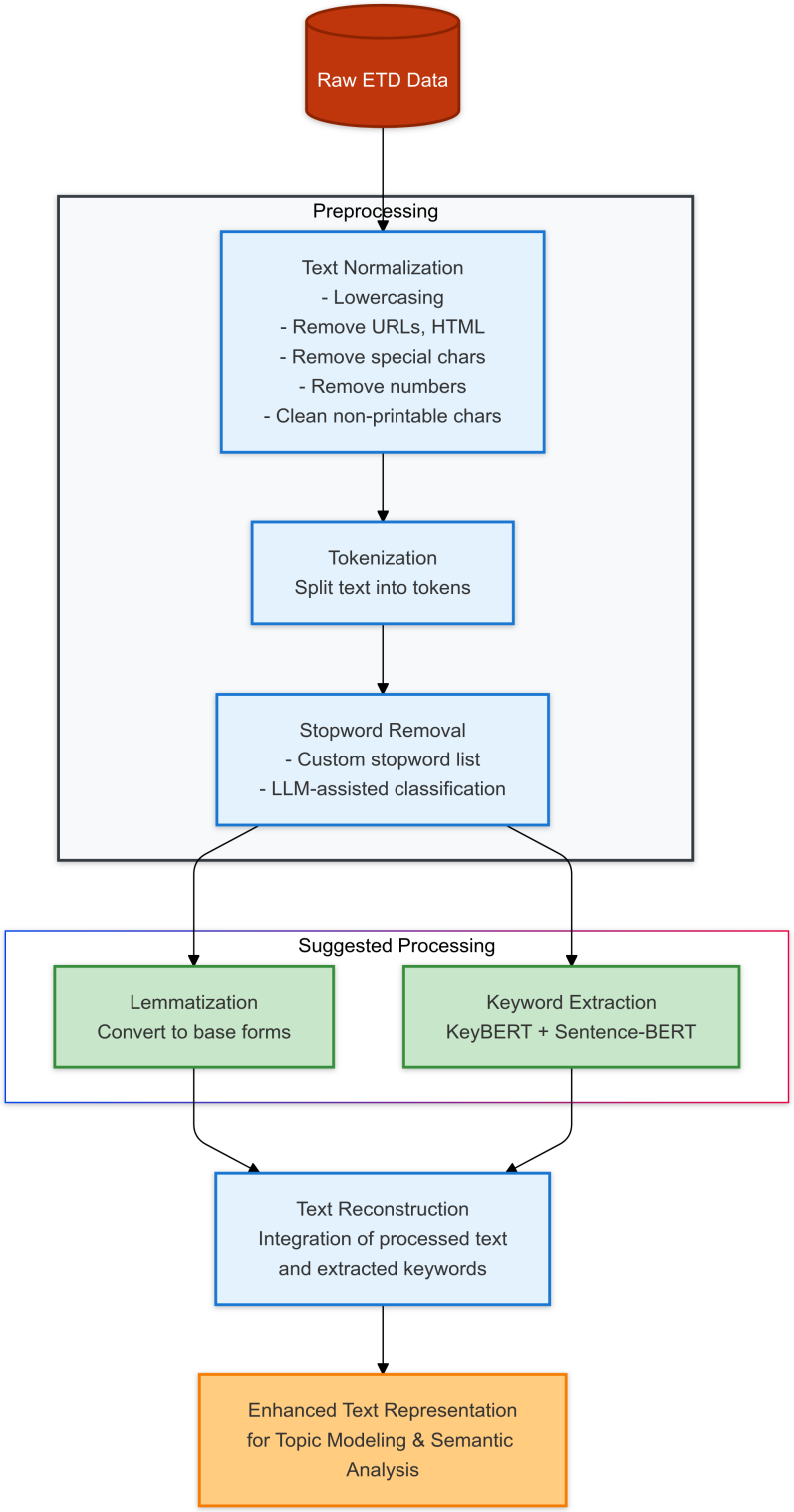

Figure 5.1: Preprocessing Pipeline: Sequential transformation of ETD text through normalization, tokenization, custom stopword removal, lemmatization, and keyword extraction stages

5.2 Core Preprocessing Steps

Through modular steps, the preprocessing pipeline transforms raw ETD text into a format optimized for topic modeling. Each step addresses specific challenges associated with academic texts, such as noise reduction, standardization, and thematic enrichment. An overview of these steps is provided below.

- **Text Normalization:** Standardizes the text to ensure uniformity and removes non-relevant elements, such as special characters and accents.
- **Tokenization:** Splits text into individual units (tokens) to enable structured word-level operations.
- **Stopword Removal:** Employs both traditional and LLM-assisted methods to identify and remove non-relevant terms tailored to academic writing.
- **Text Enhancement:** Combines metadata fields with BERT-extracted keywords to create a holistic representation of each document.

These steps collectively improve the quality of inputs for downstream tasks, such as topic modeling, while preserving the semantic richness of academic texts. The following subsections provide detailed explanations of each stage.

5.2.1 Text Normalization

Text normalization is the foundational step in the preprocessing pipeline, ensuring consistency and standardization across the dataset. This step addresses the diverse formatting and structural variations in ETD text. Key normalization processes include:

- **Lowercasing:** Converts all text to lowercase, ensuring uniformity and reducing redundancy (e.g., treating “Data” and “data” as the same token).
- **Removing Noise:** Eliminates non-relevant elements that do not contribute meaningful content, such as URLs, HTML tags, special characters, and non-printable characters.
- **Number Removal:** Excludes numerical values that are unlikely to add semantic value in topic modeling tasks.
- **Accent Normalization:** Converts accented characters (e.g., “café”) to their unaccented equivalents (e.g., “cafe”), ensuring consistency in multilingual datasets.
- **Reducing Repeated Characters:** Handles cases where repeated characters (e.g., “coooool”) might inflate vocabulary size or introduce noise.

These normalization steps reduce noise, standardize the dataset, and ensure the text is ready for downstream analysis. By addressing inconsistencies and non-relevant content, normalization significantly improves the quality of the input data, laying the groundwork for effective topic modeling.

5.2.2 Tokenization

Tokenization is a crucial step in the preprocessing pipeline that breaks text into manageable units, known as tokens, to enable structured analysis. Each token typically represents a word, providing a foundation for subsequent word-level operations such as stopword removal and lemmatization. In this research, tokenization was performed using the `word_tokenize` function from the NLTK library [99]. This process offers several key advantages. It facilitates granular processing by allowing precise application of preprocessing steps at the word

level, supports the creation of n-grams (e.g., bi-grams and tri-grams) to capture meaningful multi-word phrases essential for identifying themes, and reduces complexity by converting continuous text into standardized format, thereby simplifying downstream operations.

5.2.3 Stopword Removal

Stopword removal reduces noise in the dataset by filtering out commonly used words that do not contribute to the semantic meaning of the text. These words, such as "the," "is," and "and," often appear frequently but provide little value in topic modeling.

Traditional Stopword Removal:

- **Base Stopword List:** A foundational list from libraries like NLTK [99] and Gensim [106] was used to remove general-purpose stopwords.

- **Customization:** Generic lists were extended to include terms commonly found in academic texts (e.g., "study," "research") that are frequent but contextually non-informative.

5.3 LLM-Assisted Stopword Identification and Removal

While traditional stopword lists from libraries like NLTK and Gensim provide a foundation for text preprocessing, they do not adequately address the domain-specific nature of academic writing in ETDs. To overcome this limitation, we developed a semi-automated, LLM-assisted process that extends beyond standard stopword lists while maintaining precision through human validation.

5.3.1 Methodology

Our approach began with extracting the 2,500 most frequent words from the preprocessed corpus. These words were systematically divided into five sets of 500 words each to accommodate GPT-4's token limitations and ensure thorough processing. The process was guided by a carefully crafted prompt to identify academic-specific stopwords while preserving domain-critical terminology.

5.3.2 Prompt Engineering

The system prompt was designed to balance comprehensiveness with precision:

> **System**: You are a language processing assistant specializing in academic texts. Your task is to identify general and domain-specific stop words from a given list. Only include words that are frequent and do not add significant or unique meaning. Avoid words that may carry context-dependent importance, like 'education,' 'learning,' 'state,' or 'materials,' as they might appear in meaningful phrases (e.g., 'machine learning'). Follow the user's instructions exactly and provide the output in the requested format.
>
> **User Prompt**: I have a list of frequent words extracted from academic abstracts. The words are: [word list]
>
> **Instructions**:
>
> - Only include words from the provided list
> - Exclude words with potential specific meanings in academic contexts
> - Provide output as comma-separated values
> - Maintain original word forms without alterations

5.3.3 Implementation Algorithm

The stopword identification and removal process was implemented as follows:

Algorithm 1 LLM-Assisted Stopword Identification

```
Extract the top 2500 frequent words from the preprocessed ETD corpus
Divide the frequent words into five sets of 500 words for GPT-4 processing
for each set of 500 words do
    Generate the academic-specific prompt with the word set
    Query GPT-4 using the generated prompt
    Extract stopword candidates from GPT-4's comma-separated response
end for
Perform validation of identified stopwords
Combine validated stopwords with base stopword list
```

This process identified approximately 320 additional academic-specific stopwords, expanding our total stopword list to over 1,000 words. The human validation step proved crucial in preventing the removal of terms that, while common in academic writing, carry significant meaning in specific contexts. For example, GPT-4 initially identified words—like "learning," "novel," and "image"—as potential stopwords due to their high frequency in academic texts. However, these terms were preserved during validation because:

- "Learning" is critical in phrases like "machine learning," "deep learning," and "learning algorithms."
- "Novel" carries different meanings in contexts like "novel approach" versus "novel" in English literature studies.
- "Image" is essential in technical phrases such as "image processing" and "image recognition."

The human-in-the-loop approach, which involved three iterations to enhance the validity and consistency of the process, ensured that terms critical to domain-specific topics, such

as “learning” and “image,” were preserved. In contrast, generic academic terms, such as “dissertation” and “study,” were filtered out. This balance was crucial for retaining semantic richness without introducing noise.

5.3.4 Results and Impact

The preprocessing pipeline, including the enhanced stopword list, significantly improved the quality of downstream topic modeling tasks. While the overall pipeline contributed to increased coherence and interpretability of topics (as shown in Section 7.3), the stopword enhancement played a key role in reducing noise by filtering out common academic phrases. For example, terms like “dissertation” and “study” were excluded, enabling the models to focus on domain-specific vocabulary such as “learning” and “image.”

Future work will focus on integrating domain-specific knowledge bases (KBs) to enhance context-aware stopword identification and topic modeling. Given the corpus’s interdisciplinary nature, we envision utilizing multiple KBs tailored to specific domains, such as UMLS for biomedical content or DBpedia for general knowledge. For interdisciplinary content, a combination of KBs will be prioritized based on document metadata or inferred topics.

For domains lacking established KBs, we propose dynamically constructing term lists from high-quality corpora or using general-purpose KBs to infer related terms. Managing a diverse set of KBs will require addressing schema variability and API limitations by standardizing KB formats and using local storage solutions where necessary.

These efforts aim to balance the depth of domain knowledge with the practicality of managing a large-scale corpus, ensuring accurate and efficient processing even without predefined KBs.

These improvements aim to further streamline the preprocessing pipeline and ensure that the balance between automation and domain expertise continues to enhance the quality of

topic models.

5.4 Lemmatization

Lemmatization is the process of reducing words to their base or root forms while retaining their linguistic meaning [49]. For instance, variations like "running," "ran," and "runs" are lemmatized to "run." This ensures that grammatical variations are treated as the same term, reducing vocabulary complexity and improving topic coherence. In this research, lemmatization was performed using the `WordNetLemmatizer` [99, 116] from the NLTK library. Unlike stemming, which truncates words without considering linguistic rules, lemmatization produces linguistically valid base forms, making it particularly suited for academic texts. The benefits of lemmatization include standardizing different word forms under a single representation, maintaining semantic accuracy, and reducing noise in the vocabulary to better capture thematic structures. While stemming is faster, it often introduces inaccuracies by reducing words to invalid forms, such as truncating "studies" to "studi." It fails to differentiate between words with different meanings but similar stems. Given the precision required for processing academic texts, lemmatization was chosen to ensure linguistic accuracy and preserve semantic richness, providing the necessary foundation for effective topic modeling.

5.5 Reconstruction of Text

After tokenization, stopword removal, and lemmatization, the individual tokens are rejoined into a single string for each document. This step reconstructs the processed tokens into a continuous text format that can be used in subsequent analysis tasks. Reconstruction is essential because rejoining the tokens into a continuous string allows for compatibility

with models and functions that require full-text inputs, such as keyword extraction or topic modeling. By processing tokens individually and then reconstructing the text, we ensure that it is in its most informative and suitable form for further analysis.

5.6 Keyword Extraction

To enrich text representation, keyword extraction was performed using the `KeyBERT` model, which leverages a BERT-based embedding framework to identify the most relevant keywords from each document's abstract. Specifically, the `all-mpnet-base-v2` [101] model from sentence-transformers was employed to generate document embeddings, enabling the selection of the top five keywords based on their semantic importance. Keyword extraction is a crucial step as it concisely summarizes a document's main themes and concepts, providing a quick understanding of its content without requiring a full read. Moreover, these extracted keywords enhance the document's representation, making it more informative for downstream tasks such as topic modeling, search, or recommendation. The extracted keywords maintain contextual relevance by relying on a BERT-based model, ensuring that the selected terms strongly reflect the document's meaning and thematic structure [109, 117].

5.7 Enhanced Text Representation

To create a robust and comprehensive representation of each document, an additional **Enhanced Text** field was created in the experimental version of the Curated Dataset [112]. This field was constructed by concatenating the `title`, `abstract`, and extracted `keywords`.

Although not a part of the original metadata, the Enhanced Text field was added during preprocessing to facilitate better thematic analysis, improve content discovery, and enhance

interpretability in tasks like topic modeling. It served as an input format for experiments, ensuring that key textual elements were integrated for robust processing.

Building on the preprocessing steps outlined earlier, this representation leverages the diversity of academic language present in Titles and Abstracts, complemented by the precision of Keywords, to offer a holistic input format for topic modeling. Enhanced Text enables models to focus on granular and broad thematic content, improving coherence and relevance.

To validate the utility of Enhanced Text, a comparative evaluation was conducted against other input styles—Titles, Abstracts, and Keywords—using BERTopic as the baseline model. The assessment used four metrics: **Coherence Value (CV) Score**, **Normalized Pointwise Mutual Information (C_NPMI) Score**, **Embedding-based Topic coherence (ETC) Score**, and **Topic Diversity**, which collectively assess topic coherence, consistency, and breadth. Results, presented in Table 5.1, demonstrate that Enhanced Text from the Curated Dataset [112] significantly outperforms individual input styles by combining their strengths. This integration enables higher coherence and interpretability, making Enhanced Text a superior format for topic modeling.

Table 5.1: BERTopic Evaluation Metrics (Coherence and Diversity Scores) for Enhanced Text, Titles, Abstracts, and Keywords

Text Input	**CV Score**	**C_NPMI Score**	**ETC Score**	**Topic Diversity**
Enhanced Text	0.778	0.151	0.746	0.84
Title	0.6757	0.0681	0.6557	0.84
Abstracts	0.7172	0.1084	0.6750	0.8533
Keywords	0.7017	-0.0296	0.6705	0.8667

The CV Score measures topic coherence based on co-occurrence probabilities, while the C_NPMI Score reflects normalized pointwise mutual information, a key metric for human interpretability. Higher scores indicate better semantic coherence. Topic diversity assesses the spread of topics, ensuring that the model does not generate overly similar clusters.

5.8 Choice of Embedding Models

The high quality of BERTopic, KeyBERT, and the word embedding-based evaluation metrics heavily depends on the choice of embedding models used in this research. Embedding models transform text into dense vector representations, capturing both syntactic and semantic information. Modern topic modeling and keyword extraction approaches rely on these embeddings to produce coherent and meaningful topics, making the choice of embedding models critical to the success of this work.

5.8.1 Embedding Models by Use Case

This research employed multiple embedding models, each tailored to a specific use case, ensuring robust and contextually accurate embeddings for keyword extraction, topic modeling, and evaluation of LLM-based representations.

Keyword Extraction

The `all-mpnet-base-v2` [101] model from the Sentence-Transformers library was used for keyword extraction. Its strong performance in generating high-quality sentence embeddings made it ideal for identifying the top five most relevant keywords from document abstracts.

Topic Modeling and Representations

Two embedding models were employed for generating and refining topic representations:

- `bert-base-uncased` [104]: Used during the initial setup for BERTopic and CTM while comparing different topic modeling architectures and for embedding the top 10

keywords of each topic to calculate word embedding-based coherence scores.

- `cde-small-v1` [102]: Selected for improving topic coherence in BERTopic, providing memory-efficient, high-performance embeddings specifically tuned for semantic tasks.

Evaluation of LLM-Based Representations

The `Alibaba-NLP/gte-large-en-v1.5` [103] model was chosen for evaluating LLM-based representations using word embedding-based similarity scores. Its ability to process long sequences (up to 8192 tokens) and its robust performance across datasets made it well-suited for evaluating the semantic quality of LLM-generated topic labels and descriptions.

5.9 Summary

The preprocessing pipeline developed in this research involves several critical steps, including text normalization, tokenization, stopword removal, lemmatization, keyword extraction, and enhanced text representation. Each step transforms raw ETD data into a format suitable for further processing, ensuring that the unique complexities of academic text are preserved while reducing noise and redundancy. By including extracted keywords and the enhanced text column, the pipeline ensures that each document is represented with its most essential features, enabling better thematic exploration and content discovery.

This chapter outlined the preprocessing techniques integral to this research, emphasizing the role of LLM-assisted stopword removal and enhanced text representation in producing high-quality inputs for topic modeling. The next chapter delves into the topic modeling methodology, describing how these preprocessed inputs are leveraged to extract meaningful thematic representations from ETD metadata, and how LLMs can enhance the process.

Chapter 6

Development and Enhancement of Topic Modeling Approach

6.1 Introduction

This chapter outlines the framework (see Figure 6.1) for enhancing topic modeling in heterogeneous academic datasets, particularly Electronic Theses and Dissertations (ETDs). The proposed methodology addresses critical challenges, such as improving interpretability by integrating traditional statistical models with modern Large Language Models (LLMs).

The framework is developed and evaluated using the curated ETD dataset (described in Section 3.3.1), enabling systematic experimentation and refinement of preprocessing, modeling, and evaluation strategies. Once validated, this methodology will be extended to a larger dataset (described in Section 3.3.2) to assess its scalability and generalizability. By combining traditional topic models with LLM-driven enhancements, the proposed approach ensures a robust and interpretable pipeline for analyzing academic text collections.

This chapter is organized into three main sections. The first section describes the design and evaluation of topic modeling approaches, detailing the impact of preprocessing, optimization of the number of topics, and the systematic evaluation of different algorithms. The second section introduces the integration of LLMs to augment topic models, focusing on

improving topic representations' clarity and semantic richness. The final section provides technical specifications of the software and hardware setup, ensuring the reproducibility of the methodology.

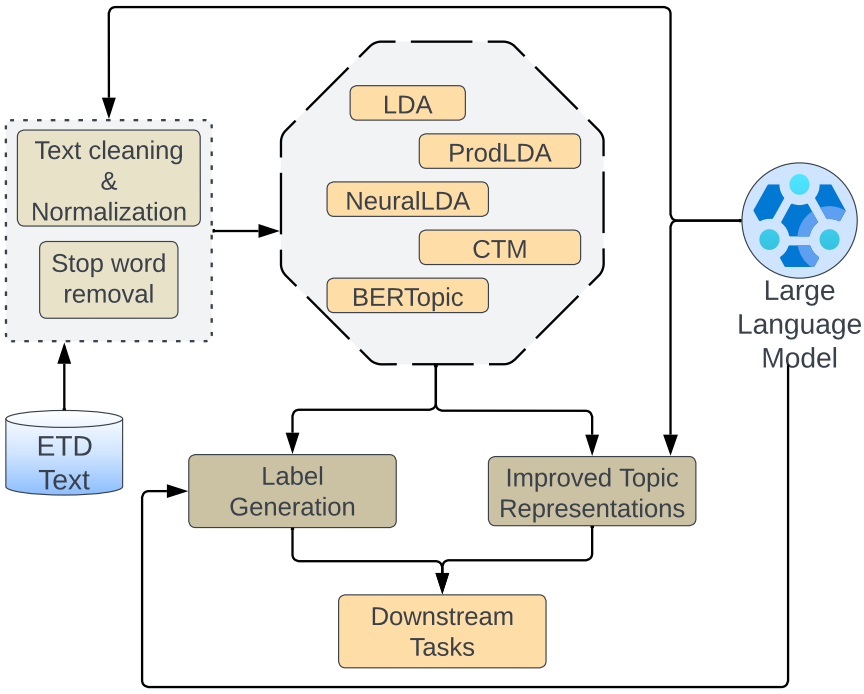

Figure 6.1: Topic Modeling Architecture showing preprocessing, model comparison, and LLM enhancement stages

6.2 Preprocessing Overview

Preprocessing plays a critical role in shaping the input for topic modeling. As described in Section 5 and Table 5.1, two distinct pipelines were used: a standard pipeline and a custom pipeline tailored to ETD data as seen in Table 6.1.

The standard pipeline includes common techniques such as tokenization, stopword removal, lemmatization, and basic text cleaning. In contrast, the custom pipeline incorporates domain-specific adjustments, including an expanded academic stopword list, using KeyBERT

for keyword extraction, and enhancing the text by combining these keywords with the title and abstract fields. These modifications address the heterogeneous nature of academic texts, ensuring the data is optimized for topic modeling.

Table 6.1: Comparison of Standard and Custom Preprocessing Pipelines for Topic Modeling

Step	**Standard Pipeline**	**Custom Pipeline (Tailored to ETD Data)**
Tokenization	Included	Same as standard pipeline
Stopword Removal	Common stopword list (e.g., NLTK, SpaCy)	Expanded academic stopword list tailored to ETD content
Lemmatization	Included	Same as standard pipeline
Text Cleaning	Basic cleaning: removing punctuation, numbers, and converting to lowercase	Extended text standardization
Keyword Extraction	Not included	Using KeyBERT to extract domain-specific keywords
Text Enhancement	Not included	Combining keywords with title and abstract fields
Target Use Case	General-purpose text preprocessing	Optimized for heterogeneous academic texts in ETD metadata

This chapter focuses on the impact of these preprocessing pipelines on topic modeling performance. Subsequent sections systematically compare the results of applying these pipelines.

6.3 Topic Modeling Development

6.3.1 Experimental Setup

Figure 6.1 shows the experimental framework for topic modeling. Five different topic modeling approaches are evaluated through systematic comparison. Each model is tested with and without custom preprocessing to assess its impact on performance. This phase aims to identify the best performing model based on topic coherence, diversity, and word embedding

metrics.

Evaluation Scope

- **Dataset:** Curated ETD dataset, described in Section 3.3.1.
- **Preprocessing Pipelines:** Standard and custom pipelines as summarized in Section 6.2.
- **Comparison Metrics:** Topic coherence, topic diversity, and embedding-based metrics as detailed in Section 6.3.5.

6.3.2 Topic Modeling Approaches

Five topic modeling approaches are evaluated, representing a spectrum of traditional probabilistic methods and modern transformer-based architectures. Each technique was implemented using established Python libraries such as Gensim [106], OCTIS [107], and BERTopic [27].

1. **LDA [21] (Gensim):** A traditional probabilistic topic modeling approach using Bayesian inference. Documents are modeled as mixtures of topics, where each topic is a probability distribution over words. This interpretable model serves as the baseline.
2. **NeuralLDA [118] (OCTIS):** A neural network-based variant of LDA that employs variational autoencoders to learn document-topic and topic-word distributions. It captures complex relationships in the data while maintaining interpretability.
3. **ProdLDA [26] (OCTIS):** An advanced neural LDA variant utilizing a product of experts' architecture. It sharpens topic distributions compared to the traditional mixture

model, potentially yielding more distinct topics.

4. **CTM [25] (OCTIS):** The Contextual Topic Model integrates neural topic modeling with transformer-based embeddings (BERT) to capture contextual relationships between words, enhancing topic coherence.

5. **BERTopic [27]:** A transformer-based approach that uses BERT embeddings for document clustering and class-based TF-IDF for topic extraction. This method identifies topics by focusing on document similarity in the embedding space.

6.3.3 Implementation Details

To ensure fair comparisons and reproducibility, as mentioned in Table 6.2, the following measures were implemented:

1. **Random Seed Control:** All experiments were initialized with a consistent random state (`random_state = 42`).

2. **Training Configuration:** Models were trained for up to 100 iterations or epochs, with early stopping applied when applicable.

3. **Preprocessing Consistency:** All models used identical input data prepared through both preprocessing pipelines.

Table 6.2: Model Configurations and Parameters

Model	**Library**	**Key Parameters**
LDA	Gensim	• `alpha = 'symmetric'` • `passes = 100` • `random_state = 42`
NeuralLDA	OCTIS	• `hidden_dim = 100` • `learning_rate = 0.2` • `batch_size = 64`
ProdLDA	OCTIS	• `hidden_dim = 100` • `dropout = 0.2` • `batch_size = 64`
CTM	OCTIS	• `bert_model = 'bert-base-uncased'` • `hidden_dim = 100`
BERTopic	BERTopic	• `embedding_model = 'bert-base-uncased'` • `clustering = 'k-means'`

6.3.4 Topic Number Optimization

Determining the optimal number of topics is critical for achieving high-quality results. Systematic experimentation was conducted with topic numbers (T) ranging from 10 to 100:

$$T = \{10, 15, 25, 50, 100\} \tag{6.1}$$

For each value of T:

1. Train all five models with identical preprocessing pipelines.
2. Evaluate performance using multiple metrics (see Section 6.3.5).

6.3.5 Evaluation Metrics

We employ several evaluation metrics to assess model performance comprehensively, each chosen to address specific aspects of topic quality. These metrics provide a balanced evaluation of all models, incorporating traditional statistical measures and modern embedding-based approaches. All metrics are applied uniformly across all models, ensuring consistency and comparability.

- **Topic Coherence:** Topic coherence is a widely used metric for evaluating the semantic consistency of topics. It measures how frequently the top words in a topic co-occur in the underlying dataset. For this study, we use two coherence metrics:
  - C_v [119]: Combines indirect cosine similarity measures with a sliding window and normalized pointwise mutual information (NPMI). This metric typically ranges from 0 to 1, where higher values indicate better coherence and stronger semantic relationships among topic words.

  - NPMI [120]: Directly measures the strength of co-occurrence between words in a topic. NPMI values generally range from -1 (no co-occurrence) to 1 (perfect co-occurrence), with values closer to 1 being desirable for coherent topics.

  These metrics provide insights into the semantic structure of topics and are applied to all models, including statistical approaches like LDA and neural approaches like CTM.

- **Topic Diversity:** Topic diversity evaluates the range of distinct words across all topics, ensuring minimal redundancy between topics. It is calculated as:

$$TD = \frac{|\bigcup_{t \in T} \text{top}_k(t)|}{k \cdot |T|}$$

  Here k is the number of top words per topic, and T is the set of all topics. For this study, we set $k = 10$. Topic diversity ranges from 0 to 1, with values closer to 1 indicating that topics are distinct and share minimal overlap in their top words. This metric is applied to all models to evaluate how effectively each method captures unique aspects of the dataset.

- **Embedding Coherence:** Embedding coherence, specifically Word Embedding-Based Pairwise Similarity (WEPS), measures the semantic similarity between topic words using pre-trained transformer embeddings. This method evaluates the cohesiveness of the top keywords within each topic by computing pairwise cosine similarity between their embeddings.

  This study uses the `bert-base-uncased` model to generate embeddings for each topic's top 10 keywords. The WEPS metric is calculated as follows:

$$\text{WEPS} = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} \cos(\text{BERT}(w_i), \text{BERT}(w_j))$$

Here:

- w_i and w_j are the top 10 keywords within a topic.
- cos represents cosine similarity between the embeddings of w_i and w_j.
- n is the total number of keywords per topic.

Higher WEPS values indicate that the top keywords in a topic are semantically cohesive. WEPS values range from -1 to 1, where values closer to 1 signify stronger semantic relationships between keywords.

The same embedding model (`bert-base-uncased`) is used across all topic modeling approaches to ensure consistency and comparability. This uniform embedding framework allows for direct comparison between statistical and transformer-based topic modeling methods.

By incorporating WEPS as an evaluation metric, we provide a robust embedding-based measure of semantic coherence that complements traditional metrics such as C_v and NPMI.

6.3.6 Model Selection for LLM Integration

The results from the initial evaluation (detailed in Chapter 7) demonstrate that **BERTopic** outperforms other models across key metrics, including C_v, C_NPMI, and Embedding-Based Topic Coherence (ETC). Its superior performance and ability to generate contextually rich and semantically coherent topic clusters position BERTopic as the most suitable *baseline model* for further enhancement using Large Language Models (LLMs).

While BERTopic was chosen for this study due to its strong performance in this specific use case, the proposed LLM-enhancement framework is designed to be model-agnostic. This

flexibility enables the enhancement of topic representations from any topic modeling approach. The methodology can be adapted to various use cases and datasets, depending on the specific requirements and performance characteristics of the chosen topic modeling technique. By demonstrating the framework's effectiveness using BERTopic, we establish its potential applicability to other models, such as LDA, NeuralLDA, or Contextualized Topic Models (CTM). This adaptability broadens the scope and impact of the proposed approach, offering a pathway for future research and applications.

To further improve the quality and interpretability of topic representations generated by BERTopic, we incorporated an embedding model optimized for semantic tasks. The selection process was guided by rankings from the Massive Text Embedding Benchmark (MTEB) leaderboard [105], focusing on balancing deployment feasibility, memory efficiency, and performance. The embedding model `cde-small-v1` was selected based on these criteria. Table 6.3 summarizes its performance characteristics.

Table 6.3: Embedding Model Metrics (MTEB Leaderboard, November 18, 2024)

Metric	**Value**
Rank on MTEB	29
Model Size (Million Parameters)	143
Memory Usage (GB, FP32)	0.53
Embedding Dimensions	768
Maximum Tokens Processable	512
Average MTEB Score (56 Datasets)	65
Classification Average (12 Datasets)	81.71
Clustering Average (11 Datasets)	48.32
Pairwise Classification Average (3 Datasets)	84.69

`cde-small-v1` demonstrates robust performance across diverse semantic tasks with an average MTEB score of **65**. Additionally, its lightweight memory usage (0.53 GB in FP32) and embedding dimensionality of 768 makes it well-suited for scalable applications, particularly for large academic datasets such as Electronic Theses and Dissertations (ETDs). Its token

limit of 512 ensures compatibility with most abstracts and shorter documents, supporting the diverse text characteristics inherent in academic collections. For longer texts, such as full documents or extended abstracts, a chunking approach can be employed, where the text is split into overlapping segments of up to 512 tokens. This ensures that all critical content is included while preserving context, and aggregated embeddings provide a comprehensive representation.

Alternatively, models with higher token limits, such as those supporting 1024 or 2048 tokens, can be explored for handling longer inputs directly. However, these models typically require greater memory and computational resources, which must be carefully managed to balance efficiency with the complexity of academic collections. The choice between chunking and higher token limit models depends on the specific requirements of the task and the available computational resources.

By integrating `cde-small-v1` with BERTopic, the enhanced pipeline leverages state-of-the-art embeddings optimized for semantic consistency and coherence. This combination improves the interpretability of topic representations and retains scalability and computational efficiency, making it a robust solution for analyzing complex academic datasets.

6.4 Enhancement Using LLMs

After determining the optimal topic modeling approach through systematic experimentation, we employed Large Language Models (LLMs) to enhance the interpretability and quality of the generated topics. By developing meaningful and concise outputs, LLMs can bridge the gap between statistical topic modeling and interpretable contextual representations. Unlike approaches that integrate LLMs at the document processing stage [121], our method focuses on post-processing, making it scalable and adaptable to various topic models, since the

computation is performed on a small set of topics rather than a large set of documents. This section details the selection of LLMs, the iterative prompt engineering process, and the evaluation framework employed.

6.4.1 LLM Selection and Comparison

For this study, we evaluated a diverse set of state-of-the-art LLMs to identify models best suited for enhancing topic representations. The models selected included GPT-4 (version GPT-4o) [6], Llama 3.1-70B [8], Llama 3.2-90B-vision, Llama 3.2-3B [7], and Mixtral-8x7B [81]. Each model was chosen based on its unique capabilities, ranging from high-capacity reasoning (GPT-4) to resource-efficient deployment (Llama 3.2-3B) and fast inference (Mixtral-8x7B).

The characteristics of these models, including their architecture, context lengths, and strengths, are detailed in Table 2.1 in the Literature Review (Chapter 2).

The selected models enabled a systematic evaluation across several dimensions:

- **Scale Variation:** From resource-efficient models such as Llama 3.2-3B to high-capacity models like GPT-4, allowing performance assessments between different sizes.
- **Architectural Diversity:** Including traditional transformers (Llama models), advanced multimodal capabilities (GPT-4), and a sparse mixture of experts (Mixtral-8x7B).
- **Version Comparison:** Evaluating architectural improvements within the same model family, such as Llama 3.1-70B, Llama-3.2-3b and Llama 3.2-90B-vision.

To ensure reproducibility, these models were accessed using two platforms:

- **OpenAI API:** For GPT-4o, with version 2024-08-06.
- **Groq Playground:** For open-source models, ensuring a consistent inference environment.

This strategy ensured that other researchers could reproduce and validate our methodology while enabling a comprehensive assessment of scalability and generalizability across diverse architectures and computational requirements.

6.4.2 Prompt Engineering

The Large Language Models (LLMs) were prompted with only the top 10 weighted keywords from BERTopic to generate various topic representations. Prompt engineering was employed to design targeted prompts for keyphrases, labels, and descriptions, ensuring each representation type met its specific interpretability goals. After independent development and validation of these prompts, a consolidated prompt was created to facilitate batch processing and cross-model comparisons.

Independent Prompt Development

Independent prompts were developed for each representation type to ensure clarity, contextual accuracy, and tailored outputs. A system-level prompt was included to set the overarching context and instruct the LLM to focus on the academic tone and semantic coherence of topic representations:

> *You are an expert in topic modeling and natural language processing. Your task is to help interpret topic model outputs from BERTopic, maintaining consistency*

and clarity while providing meaningful insights. Consider the semantic relationships between words and their natural groupings when interpreting topics. Ensure all outputs are clear, concise, and academically appropriate.

Each independent prompt followed this system instruction, focusing on generating one type of representation at a time:

Keyphrase Generation Prompt

- **Objective**: Generate concise keyphrases that represent the sub-themes and relationships within a topic.
- **Prompt**:

  Given a BERTopic topic representation with its top 10 keywords, generate 6-7 keyphrases where:

  - Each phrase is 2–3 words maximum
  - Directly combine keywords when possible
  - Avoid conjunctions (and, or) and prepositions (in, of, for)
  - Use adjective-noun or noun-noun combinations
  - Present as a single line of comma-separated phrases

Input Format: `[keyword1, keyword2, ..., keyword10]`
Output Format: *keyphrase1, keyphrase2, ..., keyphraseN*

Label Generation Prompt

- **Objective**: Create succinct, descriptive labels summarizing the topic.
- **Prompt**:

  Given a BERTopic topic representation with its top 10 keywords, create a concise topic label that:

  - Is 3–5 words long
  - Captures the main theme using the most representative keywords
  - Uses title case
  - Connects related concepts with "and" or "in" where appropriate
  - Is specific enough to differentiate from other topics

Input Format: `[keyword1, keyword2, ..., keyword10]`
Output Format: *Topic Label*

Description Generation Prompt

- **Objective**: Generate a coherent, detailed description that integrates keywords and provides an overview of the topic.
- **Prompt**:

  Given a BERTopic topic representation with its top 10 keywords, generate a 2–3 sentence description that:

- Begins with "This topic" followed by an active verb (e.g., focuses, covers, explores)
- Incorporates the main keywords naturally
- Explains how the different keywords relate to each other
- Ends with: "Other related areas, such as [example1] or [example2], might also be part of this topic."
- Maintains academic tone while being accessible

Input Format: `[keyword1, keyword2, ..., keyword10]`
Output Format: *A 2–3 sentence description*

Consolidated Prompt for Batch Processing

A consolidated prompt was developed to streamline the process and enable batch processing. This prompt combines examples of keywords, keyphrases, and descriptions for one topic and instructs the LLM to generate similar outputs for others. The consolidated prompt was tested across models such as GPT-4, Llama-3 models, and Mixtral for comparative evaluation.

> **Consolidated Prompt:** *I have a set of BERTopic results, and I need you to generate similar outputs for each topic in the following format:*
>
> - **BERTOPIC topic representation:** These are the top 10 weighted terms in the c-tf-idf format.
> - **Keyphrases:** Provide a list of descriptive keyphrases that capture the topic's essence based on the BERTopic terms.
> - **Short Label:** Provide a concise topic label/title.

- **Description:** Provide a brief description of the topic, explaining what it covers. Ensure the description is flexible by adding a sentence suggesting that other related topics might also be included.

Here's an example of the format I expect:

Example:

- **BERTOPIC topic representation:** (0, 0.034*"political" + 0.019*"state" + 0.014*"government" + 0.011*"national" + 0.011*"country" + 0.011*"international" + 0.011*"united" + 0.011*"public")
- **Keyphrases:** political systems, state authority, government structures, national governance, country policies, international relations, United Nations, public administration
- **Short Label:** Global Political Systems
- **Description:** This topic encompasses the study and analysis of political structures, governance, and policies at both the national and international levels. It includes the roles of governments, states, and international bodies, with a focus on political interactions and public administration across countries. Other related areas, such as geopolitical conflicts or international trade policies, might also be part of this topic.

Justification for Approach

Developing independent prompts for each representation ensured that outputs were tailored and optimized for their goals. The consolidated prompt was used for efficiency in batch processing and to compare the outputs across different LLMs, enabling consistency and scalability in evaluations.

6.4.3 Evaluation Metrics

To evaluate LLM-enhanced topic representations, we adopted specialized metrics that go beyond traditional coherence measures, which often fail to capture the semantic richness of LLM-generated outputs. The evaluation framework incorporates modern embedding-based approaches to assess semantic alignment and distinctiveness.

Word Embedding-Based Centroid Similarity (WECS)

Word Embedding-Based Centroid Similarity (WECS) is the primary metric used to evaluate the semantic alignment between LLM-generated representations and two key components:

1. **Topic Keywords:** To assess how well the LLM-generated representations align with the original descriptors of a topic.

2. **Metadata Columns:** To validate the relevance of LLM-generated outputs by comparing them with human-annotated metadata fields, such as departmental categorizations.

Methodology: The WECS metric involves two main steps:

- **Centroid Calculation:** The centroid of a representation (e.g., topic keywords or metadata) is computed as the mean vector of word embeddings for all words in the representation:

$$C = \frac{1}{n} \sum_{i=1}^{n} v_i$$

where v_i is the embedding vector of the i-th word, and n is the total number of words in the representation.

- **Similarity Computation:** The semantic similarity between two centroids is calculated using cosine similarity:

$$\text{WECS}(C_1, C_2) = \frac{C_1 \cdot C_2}{\|C_1\|\|C_2\|}$$

where C_1 and C_2 are the centroids of the two representations being compared. Higher WECS scores indicate stronger semantic alignment.

Implementation: The `Alibaba-NLP/gte-large-en-v1.5` model was used to generate embeddings for this study, selected for its robust performance and efficiency [103, 105]:

- **Rank:** 26th on the MTEB leaderboard [105] as of November 2024.
- **Model Size:** 434 million parameters.
- **Memory Usage:** 1.62 GB (FP32).
- **Embedding Dimensions:** 1024.
- **Token Limit:** Handles up to 8192 tokens, allowing longer representations.
- **Performance:** Achieved an average score of 65.39 across 56 datasets.

WECS Between LLM Representations and Topic Keywords

The first application of WECS evaluates the alignment of LLM-enhanced representations (e.g., labels, keyphrases, and descriptions) with the core topic keywords generated by the baseline topic model.

Purpose: This metric ensures that LLM-enhanced representations retain and enrich the semantic content of the original topic keywords.

Insights:

- **Semantic Coherence:** High WECS scores indicate that LLM-generated labels and descriptions are semantically consistent with the original keywords.
- **Interpretability:** Alignment with keywords ensures the enhanced representations are accurate and contextually relevant.

WECS Between LLM Representations and Metadata Column

The second application of WECS compares LLM-generated representations to a metadata field from the Curated ETD Dataset, specifically the `ProQuest_Depts` column. This metadata field acts as a proxy for ground truth labels, representing high-level categorizations of topics.

Methodology:

1. **Representation Selection:** Three types of LLM-generated outputs are compared,
    - Labels (concise topic descriptors)
    - Keyphrases (key concept summaries)
    - Descriptions (detailed explanations)
2. **WECS Computation:** For each representation type,
    - Generate embeddings using `gte-large-en-v1.5`.

- Compute centroid-based embeddings for both representations and metadata fields.
- Calculate WECS similarity scores.

Purpose: This evaluation validates the practical relevance of LLM-generated representations by assessing their alignment with human-annotated categorizations.

Insights:

- **Relevance to Metadata:** High WECS scores indicate that LLM-generated labels align well with departmental categories, bridging automated topic modeling and human-created metadata.
- **Contextual Validity:** The alignment validates that LLM-enhanced representations are meaningful and relevant within the academic domain.

Summary

The WECS metric provides a robust framework for evaluating LLM-enhanced topic representations' semantic alignment with machine-generated and human-annotated content. By leveraging the `gte-large-en-v1.5` model, we ensure consistency and reproducibility across evaluations, making this approach highly effective for academic datasets.

6.5 User Study: Evaluating LLM-Generated Topic Representations

We conducted a user study to evaluate topic representations' clarity, relevance, and interpretability generated by Large Language Models (LLMs). The study assessed four types of representations for academic topics, focusing on their effectiveness for Electronic Theses and Dissertations (ETDs). The pilot and main studies were conducted following approval from the Institutional Review Board (IRB# 24-973). Recruitment was conducted via email outreach using IRB-approved materials, targeting individuals with academic backgrounds, including faculty members, researchers, graduate students, and undergraduate students with relevant experience.

Further details, including the IRB approval letter and email recruitment material, are included in the Appendix (see Appendix A.1 and Appendix A.2).

6.5.1 Pilot Study and Refinements

Before implementing the main study, as part of a formative evaluation effort, a two-phase pilot study was conducted to assess the feasibility and usability of the proposed methodology. Participants were selected from a diverse pool to include individuals over 18 years old with varying academic roles and experiences, such as undergraduate students with research experience, graduate students, faculty, and professors. Efforts were made to target participants across different majors to ensure the methodology's applicability to interdisciplinary academic content.

Phase 1: Initial Pilot Study The initial pilot study involved 10 participants who evaluated 15 topics. For each topic, participants reviewed all topic representations (Extracted Keywords, LLM Keyphrases, LLM Labels, and LLM Descriptions) and were asked to generate their own labels. However, this phase revealed several challenges:

- **Time Burden:** Completing the survey took over an hour, leading to participant fatigue and reduced engagement.
- **Cognitive Load:** Participants found it challenging to evaluate 15 topics and generate labels, particularly given the domain-specific nature of ETDs.
- **Feedback from Participants:** Participants suggested reducing the number of topics and eliminating the label-generation task to improve focus and reduce workload.

Phase 2: Refined Pilot Study Based on feedback from the initial pilot study, the study design was refined to make the evaluation process more manageable:

- **Topic Selection Process:** Participants were asked to review 15 topics and select 10 topics they felt most knowledgeable about and comfortable evaluating.
- **Reduced Cognitive Load:** The label-generation task was removed to minimize participant effort and allow greater focus on evaluating pre-generated representations.
- **Streamlined Evaluations:** For each selected topic, participants first read two abstracts randomly chosen from a pool of four and then evaluated the four pre-generated topic representations (Extracted Keywords, LLM Keyphrases, LLM Labels, and LLM Descriptions).
- **Optimized Survey Design:** The survey was restructured for clarity, targeting a completion time of approximately 45 minutes.

This refined design was piloted with 7 additional participants, who provided positive feedback. Participants reported that the updated structure was less overwhelming, more engaging, and allowed for more thoughtful evaluations. These refinements were adopted for the main study.

6.5.2 Main Study Implementation

The main study was designed based on the insights from the pilot study. Recruitment was conducted via email outreach, targeting individuals with academic backgrounds, including faculty members, researchers, graduate students, and undergraduate students with experience in reading and interpreting ETDs.

Participant Demographics: As of December 25th, 2024, the main study has enrolled 10 participants, with recruitment efforts ongoing to reach the target sample size of 40 participants. From Table 6.4, all participants are over 18 and have academic backgrounds relevant to the study. Among the current participants, 5 are from the field of Computer Science, along with 1 each from Biochemistry, Civil Engineering, Computer Engineering, Veterinary Science, and Industrial Engineering. This diversity in academic disciplines reflects the study's aim to include individuals with varied yet relevant expertise to enhance the quality and applicability of the findings.

Study Procedure: The study was conducted in three main stages, following participant consent:

1. **Consent and Instructions:** Participants reviewed an information sheet detailing the study's purpose, procedures, and estimated duration. After providing consent as part

Participant	Academic Status	Age> 18	Major	Familiarity with Academic Content
P1	Graduate Student	Yes	Computer Science	Extremely Familiar
P2	Graduate Student	Yes	Biochemistry	Moderately Familiar
P3	Graduate Student	Yes	Computer Engineering	Extremely Familiar
P4	Graduate Student	Yes	Computer Science	Slightly Familiar
P5	Graduate Student	Yes	Computer Science	Moderately Familiar
P6	Graduate Student	Yes	Industrial Engineering	Slightly Familiar
P7	Graduate Student	Yes	Veterinary Science	Very Familiar
P8	Graduate Student	Yes	Computer Science	Moderately Familiar
P9	Graduate Student	Yes	Computer Science	Moderately Familiar
P10	Graduate Student	Yes	Civil Engineering	Moderately Familiar

Table 6.4: Participant Demographics (N=10)

of the first stage, they proceeded to the study tasks.

2. **Topic Selection:** Participants were presented with 15 topics and asked to select 10 topics they felt most knowledgeable about and comfortable evaluating.

3. **Evaluation of Representations:** For each of the 10 selected topics, participants:

   - **Read Abstracts:** Read two abstracts randomly selected from a pool of four prepared for the topic. This gave participants context and an understanding of the topic's scope and content.
   - **Evaluate Representations:** Evaluated four pre-generated topic representations:
     - **Extracted Keywords:** Automatically extracted keywords from the topic model.
     - **LLM Keyphrases:** Expanded keyphrases generated by the LLM.
     - **LLM Labels:** Concise, human-readable labels summarizing the topic.
     - **LLM Descriptions:** Narrative providing a detailed topic overview.

Each representation was rated on a 3-point Likert scale across three dimensions: clarity, relevance, and accuracy. Participants evaluated the representations based on how well they reflected the content of the abstract. The rating scale for each dimension was as follows:

- **Clarity**: Is the representation easy to understand?
  - **1 (Not clear)**: The representation is vague or ambiguous and does not effectively capture the main ideas of the abstract.
  - **2 (Moderately Clear)**: The representation provides some insight into the abstract's content but may lack precision or detail.
  - **3 (Very Clear)**: The representation is precise, well-defined, and clearly reflects the abstract's content.
- **Relevance**: How closely does the representation relate to the specific content of the abstract.
  - **1 (Not relevant)**: The representation does not align with the key aspects of the abstract or includes unrelated terms.
  - **2 (Moderately Relevant)**: The representation aligns with parts of the abstract but may include unrelated or unnecessary elements.
  - **3 (Highly Relevant)**: The representation is strongly aligned with the abstract and includes only terms relevant to its content.
- **Accuracy**: How well does the topic representation represent the main ideas or themes of the abstracts.
  - **1 (Not accurate)**: The representation misrepresents the abstract or includes incorrect information.
  - **2 (Moderately Accurate)**: The representation captures the general idea of the abstract but may contain minor inaccuracies.

    * **3 (Highly Accurate)**: The representation is precise and correctly reflects the abstract's content.

4. **Ranking Representations:** After completing the ratings, participants ranked the four representation types for each topic based on their overall effectiveness or usefulness for understanding and identifying the content from the abstracts. Here, "Effectiveness" refers to how well the representation helps participants understand and identify the topic's content from the abstracts. The ranking task helped identify which representation type participants found most helpful for quickly grasping and navigating the topic collection.

Data Collection: The survey was hosted on QuestionPro to ensure accessibility. Data collected included Likert-scale ratings, rankings of representations, and qualitative feedback on the study in the form of a text box at the end.

Analysis Approach

The collected data were analyzed to provide both aggregated metrics and topic-specific insights:

- **Quantitative Analysis:**
  - **Average Scores:** Likert-scale ratings for clarity, relevance, and accuracy were averaged across topics and participants.
  - **Ranking Preferences:** Rankings for clearness and effectiveness were aggregated to identify the most favored representation types, accounting for ties when applicable.

- **Qualitative Analysis:** Open-ended feedback, where available, was analyzed for recurring themes and participant suggestions.

- **Topic-Specific Insights:** Individual topics were examined to identify cases where certain representation types consistently performed well or were preferred.

With feedback from 14 participants in the pilot study and data from 10 participants in the main study, preliminary results indicate that the refined design effectively balances participant workload and evaluation depth. Initial findings suggest variations in preferences for representation types, emphasizing the need for further analysis as additional data is collected. The detailed results and insights will be discussed in the next chapter.

The iterative design process, informed by the pilot study, ensured the evaluation methodology was rigorous and accessible, allowing participants to engage effectively and provide meaningful feedback. Ongoing recruitment efforts aim to expand the participant pool and further refine the findings, contributing to a deeper understanding of how LLMs can enhance topic modeling for academic datasets. The next chapter presents detailed results and preliminary conclusions of this study.

6.6 Implementation Environment

To ensure reproducibility and clearly explain our experimental setup, this section details the hardware configurations, software dependencies, and runtime environment used in our implementation.

6.6.1 Hardware Configuration

All experiments were conducted on a high-performance system equipped with dual NVIDIA Tesla P40 GPUs. Since both GPUs share identical specifications, their details are presented in a single column in Table 6.5. This configuration ensured efficient parallel processing for computationally intensive tasks.

Table 6.5: Hardware Specifications for Experiments

Specification	**GPU Specifications**
Model	NVIDIA Tesla P40
CUDA Version	11.3
Driver Version	465.19.01
Memory	22919 MiB
Power Capacity	250 W
ECC Status (Errors)	No Errors
Compute Mode	Default

6.6.2 Software Environment

The experiments were implemented using Python 3.10.5 with various software libraries for topic modeling, embeddings, and evaluation. The critical dependencies, along with their purposes and versions, are outlined in Table 6.6.

6.6.3 Runtime Environment

The experiments were conducted in the following runtime environment to ensure compatibility and optimal performance:

- **Operating System:** Ubuntu 16.04.7 LTS

Table 6.6: Software Dependencies and Versions

Library	Version	Purpose
Gensim	4.3.2	Implementation of LDA model and evaluation
OCTIS	1.13.1	Neural topic modeling (NeuralLDA, ProdLDA, CTM) and evaluation
BERTopic	0.16.4	Transformer-based topic modeling
PyTorch	1.12.1	Deep learning framework for neural models
Transformers	4.46.0	BERT model implementation
sentence-transformers	3.2.1	Document embedding generation
KeyBERT	0.8.5	Keyword extraction
scikit-learn	1.5.2	Data preprocessing and evaluation metrics

- **Python Version:** Python 3.10.15

6.6.4 Computational Requirements

The computational requirements for the experiments varied depending on the specific tasks. Details are as follows:

- **Topic Modeling Phase:**
  - **LDA:** Primarily CPU-based processing.
  - **Neural Models (NeuralLDA, ProdLDA, CTM):** Utilized GPU acceleration for faster training and experimentation.
  - **BERTopic:** Required moderate GPU memory, utilizing approximately 10–12 GB VRAM for clustering and topic extraction.
- **LLM Enhancement Phase:**
  - **Model Inference:**

    - **GPT-4:** Conducted using OpenAI's API feature for generating labels, keyphrases, and descriptions.
    - **Other Models:** Inference for models such as Llama-3 models and Mixtral was executed on the Groq Playground platform, which provides a consistent inference environment for open-source LLMs.

  - **Embedding Generation:** Performed locally using `sentence-transformers`, requiring approximately 4–6 GB VRAM per batch.

6.7 Conclusion

This chapter detailed the computational environment and experimental setup used to address the challenges of topic modeling for heterogeneous academic datasets. We developed a robust, scalable, and reproducible methodology by systematically evaluating diverse modeling approaches, including traditional and neural-based methods, and integrating Large Language Models (LLMs) for interpretability enhancement. These foundations ensure that the proposed techniques are well-suited to the complexities of analyzing Electronic Theses and Dissertations (ETDs) and similar large-scale academic datasets.

The hardware and software configurations underline the importance of balancing computational efficiency with scalability. This balance is critical for handling both traditional models and transformer-based methodologies, mainly when working with resource-intensive frameworks like BERTopic and LLM-enhanced pipelines.

Having established the methodological framework, the next chapter transitions into a detailed analysis of the experimental results. This includes comparing the performance of different topic modeling techniques, assessing the impact of preprocessing and LLM en-

hancements, analyzing timing metrics, and exploring user evaluations. These results provide critical insights into the efficacy of the proposed approaches, their computational efficiency, and their potential implications for academic text analysis.

Chapter 7

Results and Analysis

7.1 Introduction

This chapter presents the findings from the experiments conducted to evaluate the effectiveness of the proposed topic modeling framework. The results are organized into distinct sections to ensure clarity and focus:

- **Performance Comparison of Topic Modeling Approaches:** Analyzes the results of traditional and neural topic modeling methods based on quantitative evaluation metrics.

- **Impact of Preprocessing:** Examines the effect of custom preprocessing on model performance, highlighting improvements in coherence and diversity.

- **LLM-Enhanced Results:** Evaluates the contributions of LLMs to enhancing topic interpretability through labels, keyphrases, and descriptions, supported by embedding-based metrics and comparisons with dataset metadata.

- **User Study Results:** Presents the outcomes of user evaluations, combining quantitative ratings and qualitative feedback to assess the practicality and usability of the generated representations.

Through these sections, we aim to provide a comprehensive analysis of the proposed methodology, demonstrating its strengths and identifying areas for improvement. The results are further contextualized in the following chapter, discussing their implications, limitations, and potential future directions.

7.2 Performance Comparison of Topic Modeling Approaches

We evaluated five topic modeling approaches (LDA, ProdLDA, NeuralLDA, BERTopic, and CTM) using four key metrics: CV Score, C_NPMI, Embedding Topic Coherence (ETC), and Topic Diversity. Each model was tested with varying numbers of topics (k = 10, 15, 25, 50, 100) under two conditions: with and without custom preprocessing. Tables 7.1 and 7.2 present the complete quantitative results.

7.2.1 Quantitative Metrics

CV Score: BERTopic achieved the highest CV scores in most configurations, particularly with k = 15 (0.778) and k = 50 (0.768) after preprocessing. CTM showed competitive performance, reaching 0.727 at k = 50 with preprocessing, while NeuralLDA consistently showed lower scores across all configurations.

C_NPMI: BERTopic demonstrated the highest C_NPMI scores across most topic counts (reaching 0.151 at k = 15 with preprocessing), followed by CTM (maximum 0.118 at k = 10). NeuralLDA consistently produced negative C_NPMI values (ranging from -0.326 to -0.249), indicating poor topic coherence.

Embedding Topic Coherence: LDA achieved the highest ETC scores without preprocessing (0.781 at k = 15), while neural models showed varied performance. With preprocessing, NeuralLDA showed consistent improvement, reaching 0.775 at k = 15.

Topic Diversity: NeuralLDA maintained maximum diversity (1.0) across most configurations, particularly at lower topic counts ($k \leq 25$). BERTopic showed decreasing diversity as topic counts increased, from 0.91 at k = 10 to 0.700 at k = 100 with preprocessing.

7.2.2 Observations Across Models

- **BERTopic:** BERTopic consistently delivered the best performance across metrics, particularly in CV scores and C_NPMI. Its strength lies in leveraging transformer-based embeddings for clustering, enhancing semantic coherence and interpretability. The reduction in topic diversity compared to LDA and NeuralLDA suggests a focus on refining topic specificity rather than maximizing diversity.

- **CTM:** CTM showed strong performance, particularly in embedding coherence (ETC), which aligns with its use of BERT embeddings to incorporate word context. Custom preprocessing improved CTM's coherence and diversity metrics. However, its reliance on embeddings led to some topic redundancy, as seen in slightly lower topic diversity.

- **ProdLDA:** ProdLDA demonstrated notable gains in C_NPMI and ETC with custom preprocessing, indicating its ability to adapt to clean inputs. However, its performance lagged behind BERTopic and CTM, suggesting that its product-of-experts formulation may lack the capacity to capture contextual relationships fully.

- **LDA:** As a baseline model, LDA performed well in topic diversity but lagged in coherence-based metrics, highlighting its reliance on statistical word co-occurrences

without leveraging contextual embeddings. Tailored stopword removal improved its performance, but the gains were less pronounced than those for neural approaches.

- **NeuralLDA:** NeuralLDA struggled across all metrics, with consistently negative scores for C_NPMI, and lower CV scores than for other models. Custom preprocessing led to marginal improvements but did not resolve the model's fundamental issues.

7.2.3 Trade-Offs

Our results reveal clear trade-offs between coherence and diversity. BERTopic achieved the highest C_NPMI scores (0.147-0.151 for k = 15-50 with preprocessing) but showed lower topic diversity (0.748-0.840 for the same range) compared to NeuralLDA. Conversely, NeuralLDA maintained perfect topic diversity (1.0) for $k \leq 25$ but demonstrated the lowest coherence scores (C_NPMI < -0.26). CTM offered a middle ground, with moderate to high coherence (C_NPMI: 0.094-0.118) while maintaining relatively high diversity (0.87-1.00).

Algorithm	**CV Score**					**C_NPMI Score**				
	10	**15**	**25**	**50**	**100**	**10**	**15**	**25**	**50**	**100**
Without Custom preprocessing										
LDA	0.570	0.585	0.566	0.506	0.441	0.050	0.063	0.052	0.004	-0.057
ProdLDA	0.530	0.558	0.553	0.532	0.566	-0.040	-0.001	-0.024	0.020	0.034
NeuralLDA	0.325	0.303	0.281	0.290	0.305	-0.326	-0.263	-0.262	-0.249	-0.261
BERTopic	0.705	0.728	0.708	0.737	0.700	0.101	0.116	0.112	0.113	0.106
CTM	0.619	0.593	0.595	0.629	0.621	0.022	0.002	0.041	0.060	0.056
With Custom preprocessing										
LDA	0.648	0.605	0.603	0.528	0.433	0.078	0.059	0.053	-0.015	-0.075
ProdLDA	0.615	0.645	0.632	0.595	0.618	0.044	0.089	0.076	0.060	0.075
NeuralLDA	0.533	0.488	0.510	0.420	0.409	-0.317	-0.320	-0.282	-0.284	-0.274
BERTopic	0.746	0.778	0.765	0.768	0.694	0.132	0.151	0.150	0.147	0.096
CTM	0.706	0.680	0.656	0.727	0.679	0.118	0.094	0.098	0.114	0.103

Table 7.1: Comparison of CV Score and C_NPMI Score across models with and without Custom Preprocessing

7.3 Impact of Preprocessing

Preprocessing showed varying effects across models and metrics, with the most substantial improvements observed in coherence measures as shown in Tables 7.1 and 7.2:

Algorithm	ETC Score					Topic Diversity				
	10	15	25	50	100	10	15	25	50	100
Without Custom preprocessing										
LDA	0.777	0.781	0.739	0.721	0.738	0.92	0.913	0.892	0.898	0.907
ProdLDA	0.722	0.725	0.725	0.734	0.711	0.87	0.82	0.868	0.788	0.706
NeuralLDA	0.744	0.750	0.736	0.751	0.755	1.0	1.0	1.0	0.982	0.944
BERTopic	0.730	0.732	0.745	0.762	0.746	0.88	0.85	0.78	0.698	0.644
CTM	0.762	0.721	0.730	0.747	0.752	0.96	0.94	0.90	0.88	0.71
With Custom preprocessing										
LDA	0.742	0.733	0.724	0.729	0.716	0.93	0.94	0.924	0.934	0.928
ProdLDA	0.756	0.734	0.731	0.739	0.733	0.88	0.89	0.84	0.90	0.79
NeuralLDA	0.767	0.775	0.760	0.759	0.755	1.00	1.00	1.00	0.98	0.949
BERTopic	0.723	0.746	0.731	0.712	0.709	0.91	0.84	0.748	0.728	0.700
CTM	0.758	0.731	0.734	0.722	0.749	1.00	0.97	0.98	0.87	0.77

Table 7.2: Comparison of ETC Score and Topic Diversity across models with and without Custom Preprocessing

7.3.1 Quantitative Impact

- **CV Score:** BERTopic showed the largest absolute improvement with preprocessing, increasing from 0.728 to 0.778 at k = 15. CTM similarly improved from 0.593 to 0.680 at the same topic count.
- **C_NPMI:** CTM showed the most consistent improvements with preprocessing, with scores increasing by 0.092 (from 0.022 to 0.118) at k = 10. BERTopic's improvements were more modest but maintained the highest absolute scores.
- **ETC:** ProdLDA showed the most consistent ETC improvements with preprocessing, particularly at k = 10 (from 0.722 to 0.756). However, the impact varied across topic counts and models.

- **Topic Diversity:** Preprocessing had minimal impact on topic diversity for most models. NeuralLDA maintained perfect diversity (1.0) for k $\leq$ 25 regardless of preprocessing.

7.3.2 Model-Specific Insights

The impact of preprocessing varied significantly across different model architectures:

- **BERTopic:** Demonstrated the most robust performance improvements with preprocessing, particularly in coherence metrics. CV scores improved consistently across topic counts (k = 10-50), with the largest gain at k = 15 (0.050 increase).

- **CTM:** Showed substantial improvements in C_NPMI scores with preprocessing, particularly at lower topic counts. The model maintained balanced performance across all metrics, with preprocessing enhancing both coherence and diversity measures.

- **ProdLDA:** Benefited from preprocessing primarily in coherence metrics, with C_NPMI scores improving by 0.084 at k = 15. ETC scores showed modest but consistent improvements across topic counts.

- **LDA:** Showed moderate improvements with preprocessing, particularly in CV scores at lower topic counts. However, benefits diminished as topic counts increased (k > 50).

- **NeuralLDA:** Demonstrated limited improvements with preprocessing, maintaining consistently lower coherence scores compared to other models. Its primary strength remained in topic diversity, which was unaffected by preprocessing.

7.3.3 General Observations

The impact of preprocessing demonstrated complex patterns across topic counts and models. Our analysis revealed several key findings:

- **Variable Impact Across Topic Counts:** The effectiveness of preprocessing did not consistently peak at optimal topic numbers. For instance, CTM showed larger improvements in CV score at k = 50 (0.098 increase) compared to k = 15 (0.087 increase), while BERTopic's improvements were more pronounced at lower topic counts (0.050 increase at k = 15).
- **Model-Dependent Patterns:** The relationship between preprocessing impact and topic count varied by model architecture. Neural models (CTM, ProdLDA) showed more consistent improvements across topic counts, while traditional models (LDA) showed diminishing returns at higher topic numbers.
- **Metric-Specific Effects:** The impact of preprocessing varied substantially across different evaluation metrics. Coherence metrics (C_NPMI, CV) showed more pronounced improvements compared to diversity metrics, which remained relatively stable regardless of preprocessing. However, the Embedding Topic Coherence (ETC) metric showed inconsistent effects across models, with no clear trend, highlighting that preprocessing's impact on semantic coherence is complex and model-dependent.
- **Stability at Scale:** For most models, the benefits of preprocessing became less pronounced at higher topic counts (k > 50), suggesting that increased model complexity may partially compensate for the lack of preprocessing.

While custom preprocessing generally improved model performance, particularly for coherence metrics and newer architectures, the improvements were not universal and sometimes

came with trade-offs in other metrics.

Runtime Analysis

The runtime of topic modeling methods is critical for both small-scale and large-scale datasets. Table 7.3 compares the average runtime of five models—LDA, ProdLDA, NeuralLDA, CTM, and BERTopic—using standard and custom preprocessing pipelines.

The standard pipeline involves basic text cleaning, tokenization, lemmatization, and stopword removal. The custom pipeline includes additional steps like an expanded academic stopword list, domain-specific keyword extraction, and text enhancement by combining title, abstract, and keywords.

Table 7.3: Average Runtime (in seconds) for Topic Models with Standard and Custom Pipelines

Model	**Standard Pipeline (s)**	**Custom Pipeline (s)**
LDA	300	350
ProdLDA	320	380
NeuralLDA	400	460
CTM	360	420
BERTopic	1230	1442

LDA is the fastest model due to its lightweight statistical nature, while BERTopic is the slowest because of embedding generation and clustering. However, for BERTopic, embedding calculation is a one-time process. Once embeddings are generated, the topic extraction process takes only around 300–350 seconds on average. NeuralLDA, which relies on neural networks, and CTM, which leverages contextual embeddings, have moderate runtimes, with ProdLDA offering a balance between efficiency and coherence. The custom pipeline increases runtimes across all models due to enriched input data but enhances topic coherence and relevance.

7.4 LLM-Enhanced Results

7.4.1 Topic Representations: From Keywords to Enhanced Descriptions

Table 7.4: Original BERTopic Keywords for Each Topic

Topic	Original Keywords
1	narrative, cultural, identity, film, woman, artist, argue, text, culture, audience
2	health, child, social, participant, woman, adolescent, adult, intervention, individual, depression
3	political, social, woman, identity, war, cultural, politics, Mexican, government, party
4	consumer, firm, market, brand, advertising, political, price, news, social, essay
5	city, housing, planning, neighborhood, land, transportation, historic, park, sustainable, transit
6	teacher, school, education, classroom, teaching, experience, educational, participant, academic, interview
7	quantum, theorem, beam, plasma, manifold, regularity, laser, magnetic, finite, surface
8	algorithm, optimization, sensor, hardware, protocol, mission, machine, trajectory, wireless, accuracy
9	film, surface, polymer, oxide, metal, temperature, electron, catalyst, thermal, silicon
10	language, speech, speaker, English, Spanish, bilingual, child, learner, word, linguistic
11	diet, gene, plant, fed, specie, concentration, protein, crop, genetic, dietary
12	reservoir, fracture, oil, gas, numerical, CO2, temperature, particle, velocity, surface
13	basin, rock, seismic, fault, sediment, shale, reservoir, fracture, river, deposit
14	cell, protein, imaging, tissue, cancer, muscle, vivo, optical, disease, drug
15	op, doctoral, music, organ, musical, minor, treatise, lecture, song, jazz

The topic modeling results using BERTopic were enhanced through LLM-based postprocessing, which generated semantically rich and human-readable representations for each topic. Table 7.4 lists the original keyword-based topic representations generated by BERTopic. These keywords provided the foundation for developing short labels, keyphrases, and de-

scriptions using GPT-4.

Table 7.5 presents the enhanced representations, which include hierarchical outputs: short labels (concise summaries of each topic), keyphrases (providing sub-theme granularity), and detailed descriptions (capturing the scope and context of each topic). These enhanced outputs aim to improve the interpretability and usability of the topic modeling results.

Table 7.5: LLM (GPT-4)-Enhanced Representations for Each Topic

Topic	**Short Label**	**Keyphrases**	**Description**
1	Cultural Narratives in Media	narrative structures, cultural identity, film critique, female representation in media, artist expressions, textual analysis, audience reception	This topic focuses on the exploration of cultural narratives, identity, and representation in film and other media, often involving discussions around women and artistic perspectives. It includes analysis of audience reception and textual critiques. Other related areas, such as gender studies or media theory, might also be part of this topic.
2	Health and Social Wellbeing	child health, social wellbeing, adolescent care, adult interventions, women's health, depression studies, health interventions	This topic centers on the health and social wellbeing of various demographic groups, including children, adolescents, and adults, with a particular focus on interventions and issues such as depression and women's health. Other related fields, such as public health or mental health interventions, might also be part of this topic.
3	Political Identity and Social Issues	political identity, social dynamics, women's rights, war politics, cultural politics, Mexican government, political parties	This topic covers the intersection of political identity and social issues, examining the roles of women, war, and cultural dynamics within political contexts. It includes the political structures of Mexico and broader governmental and party systems. Other related areas, such as global political movements, might also be part of this topic.
4	Consumer Markets and Advertising	consumer behavior, market analysis, brand marketing, advertising strategies, pricing models, social media influence, political advertising	This topic focuses on consumer behavior and market strategies, covering aspects such as branding, advertising, and pricing models. It also touches on the influence of political advertising and news media. Other related areas, such as digital marketing or market research, might also be part of this topic.

Continued on next page

Table 7.5 – continued from previous page

Topic	Short Label	Keyphrases	Description
5	Urban Planning and Sustainability	urban planning, housing development, city infrastructure, neighborhood design, land use, transportation systems, sustainable cities	This topic addresses urban planning and sustainability, including housing, land use, and transportation within city environments. It also involves discussions around historic preservation and sustainable development. Other related fields, such as smart cities or urban policy, might also be part of this topic.
6	Teacher Education and Classroom Dynamics	teacher education, classroom practices, school systems, teaching experiences, academic participation, educational interviews	This topic revolves around teacher education, classroom practices, and the broader school systems, focusing on both teaching experiences and academic participation. Other related areas, such as educational policy or student-teacher interactions, might also be part of this topic.
7	Quantum Physics and Theoretical Models	quantum physics, beam theory, plasma dynamics, manifold structures, laser technology, magnetic fields, surface regularity	This topic covers theoretical and applied aspects of quantum physics, including theorems, beam dynamics, and plasma studies, as well as related areas such as magnetic fields and laser technology. Other related areas, such as particle physics or quantum computing, might also be part of this topic.
8	Algorithms and Optimization in Engineering Systems	algorithm design, optimization techniques, sensor integration, hardware protocols, wireless networks, mission planning, machine learning accuracy	This topic focuses on algorithm design and optimization techniques within engineering systems, particularly in areas such as sensor integration, hardware protocols, and wireless communication. It also includes machine learning and accuracy considerations. Other related areas, such as control systems or computational engineering, might also be part of this topic.
9	Materials Science and Surface Chemistry	film materials, surface chemistry, polymer coatings, metal oxides, temperature effects, electron behavior, catalytic surfaces	This topic encompasses the study of materials science, focusing on surface chemistry, polymer coatings, and metal oxides, with considerations of temperature effects and electron behavior. Other related areas, such as nanotechnology or material properties, might also be part of this topic.
10	Language Learning and Bilingualism	language acquisition, speech patterns, bilingual education, child language development, linguistic studies, Spanish-English learners	This topic covers the study of language learning and bilingualism, focusing on speech patterns, language acquisition in children, and linguistic analysis, particularly in Spanish-English learners. Other related areas, such as multilingual education or phonetics, might also be part of this topic.

Continued on next page

Table 7.5 – continued from previous page

Topic	Short Label	Keyphrases	Description
11	Dietary Genetics and Plant Biology	dietary studies, genetic research, plant species, crop production, protein concentration, gene expression	This topic focuses on the intersection of dietary studies and genetic research, exploring plant biology, crop production, and protein concentrations. It also delves into genetic variations and how they affect species development. Other related areas, such as nutrition science or biotechnology, might also be part of this topic.
12	Reservoir Engineering and Oil Extraction	oil reservoirs, gas extraction, fracture mechanics, CO2 injection, numerical modeling, particle velocity, surface temperature	This topic focuses on reservoir engineering, particularly in oil and gas extraction, including fracture mechanics and CO2 injection techniques. It also includes numerical modeling and particle velocity analysis. Other related areas, such as geothermal energy or fluid dynamics, might also be part of this topic.
13	Geological Formations and Seismic Analysis	basin analysis, seismic activity, rock formations, fault lines, sediment deposition, shale reservoirs, river deposits	This topic explores geological formations and seismic activity, focusing on basin analysis, rock formations, fault lines, and sediment deposition. It also includes the study of shale reservoirs and river deposits. Other related areas, such as earthquake studies or hydrogeology, might also be part of this topic.
14	Cellular Biology and Imaging Techniques	cell biology, protein imaging, tissue analysis, cancer research, muscle studies, in vivo experiments, optical imaging	This topic covers cellular biology and imaging techniques, with a focus on protein structures, tissue analysis, and cancer research. It also includes muscle studies and in vivo experiments using optical imaging. Other related areas, such as drug discovery or molecular biology, might also be part of this topic.
15	Doctoral Studies in Music Composition	doctoral music studies, organ compositions, musical treatises, jazz performance, music lectures, minor compositions	This topic focuses on doctoral studies in music composition, particularly organ music, jazz performance, and musical treatises. It includes the study of minor compositions and music lectures. Other related areas, such as performance theory or music education, might also be part of this topic.

As shown in Table 7.4, the original BERTopic outputs are keyword-focused and often lack sufficient contextual clarity. For instance, Topic 0 is represented by keywords such as “narrative,” “cultural,” and “identity,” which, while relevant, provide limited interpretability.

In contrast, the enhanced representations in Table 7.5 offer a more interpretable structure. For example, Topic 0 is labeled as “Cultural Narratives in Media,” with keyphrases such

as “cultural identity” and “film critique,” and a detailed description highlighting areas like gender studies and media theory. This hierarchical structure provides both breadth and depth, making the topics easier to understand and apply in academic contexts.

While the enhanced representations offer clear qualitative improvements over the original keywords, their effectiveness must also be quantitatively evaluated. To this end, we used embedding-based metrics to assess semantic coherence and alignment with dataset metadata. The results of these quantitative analyses are discussed in the following sections.

Word Embedding-Based Centroid Similarity (WECS) Analysis

To evaluate the semantic alignment of LLM-generated representations (labels, keyphrases, and descriptions) with topic keywords, Word Embedding-Based Centroid Similarity (WECS) scores were computed. These similarity scores assess how well the generated outputs align with the underlying topic keywords, quantitatively evaluating their coherence and relevance.

WECS for Labels and Topic Keywords Figure 7.1 presents the similarity matrix for labels and topic keywords. High diagonal values indicate strong semantic alignment, demonstrating that the generated labels effectively summarize the core themes of the topics. Non-diagonal values highlight potential overlaps or ambiguities between topics.

WECS for Keyphrases and Topic Keywords The similarity matrix for keyphrases and topic keywords is shown in Figure 7.2. Keyphrases, which provide more detailed representations of topics than labels, exhibit strong diagonal values, indicating effective alignment with the topic keywords.

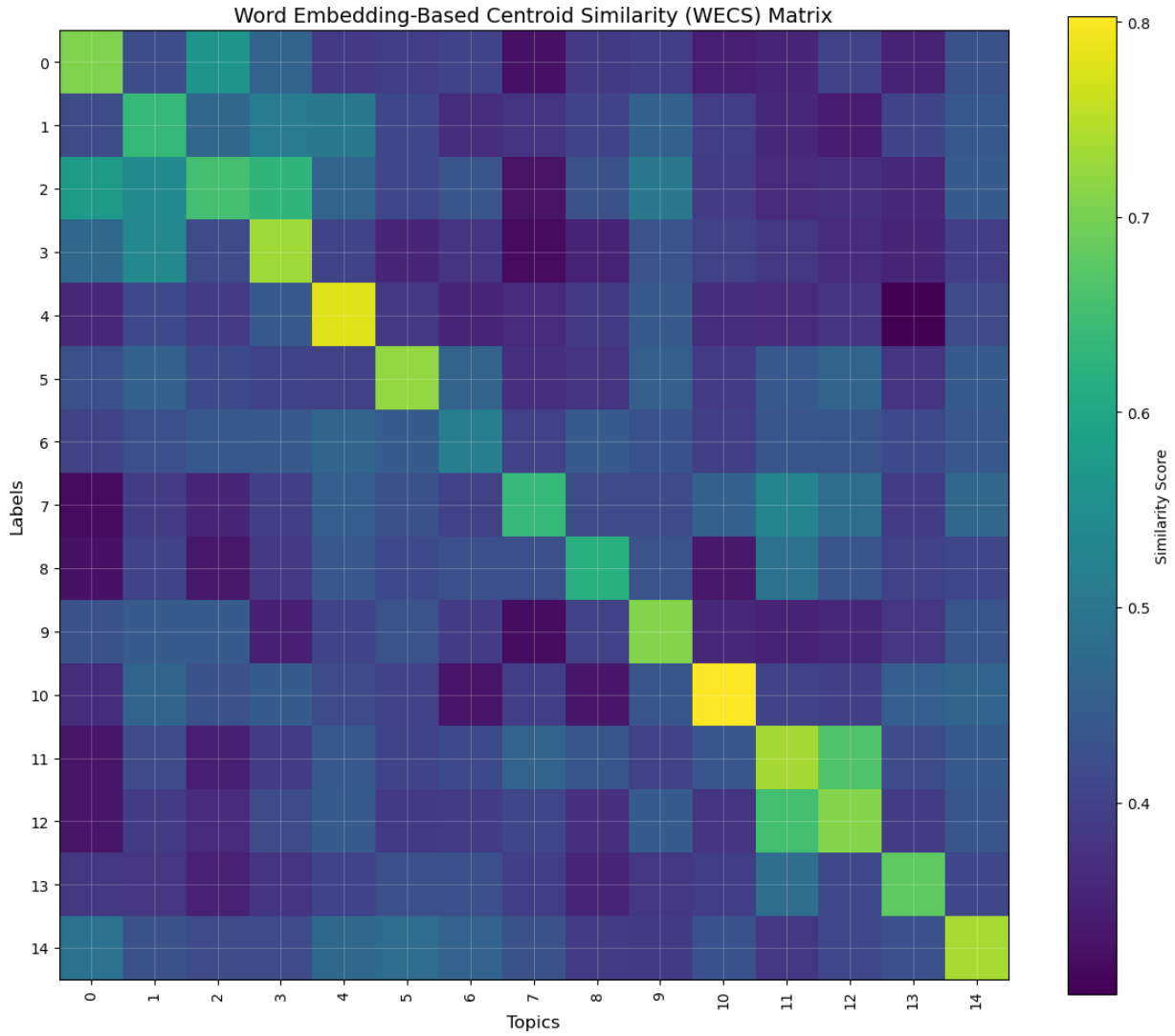

Figure 7.1: Word Embedding-Based Centroid Similarity (WECS) Matrix for Labels and Topic Keywords. The matrix shows how well LLM-generated labels align with their corresponding topic keywords. Higher diagonal values indicate stronger alignment.

WECS for Descriptions and Topic Keywords Figure 7.3 illustrates the similarity matrix for descriptions and topic keywords. Descriptions provide narrative-like summaries that encapsulate the semantic essence of the topics. High diagonal values in the matrix validate the coherence and relevance of the generated descriptions concerning the topic keywords.

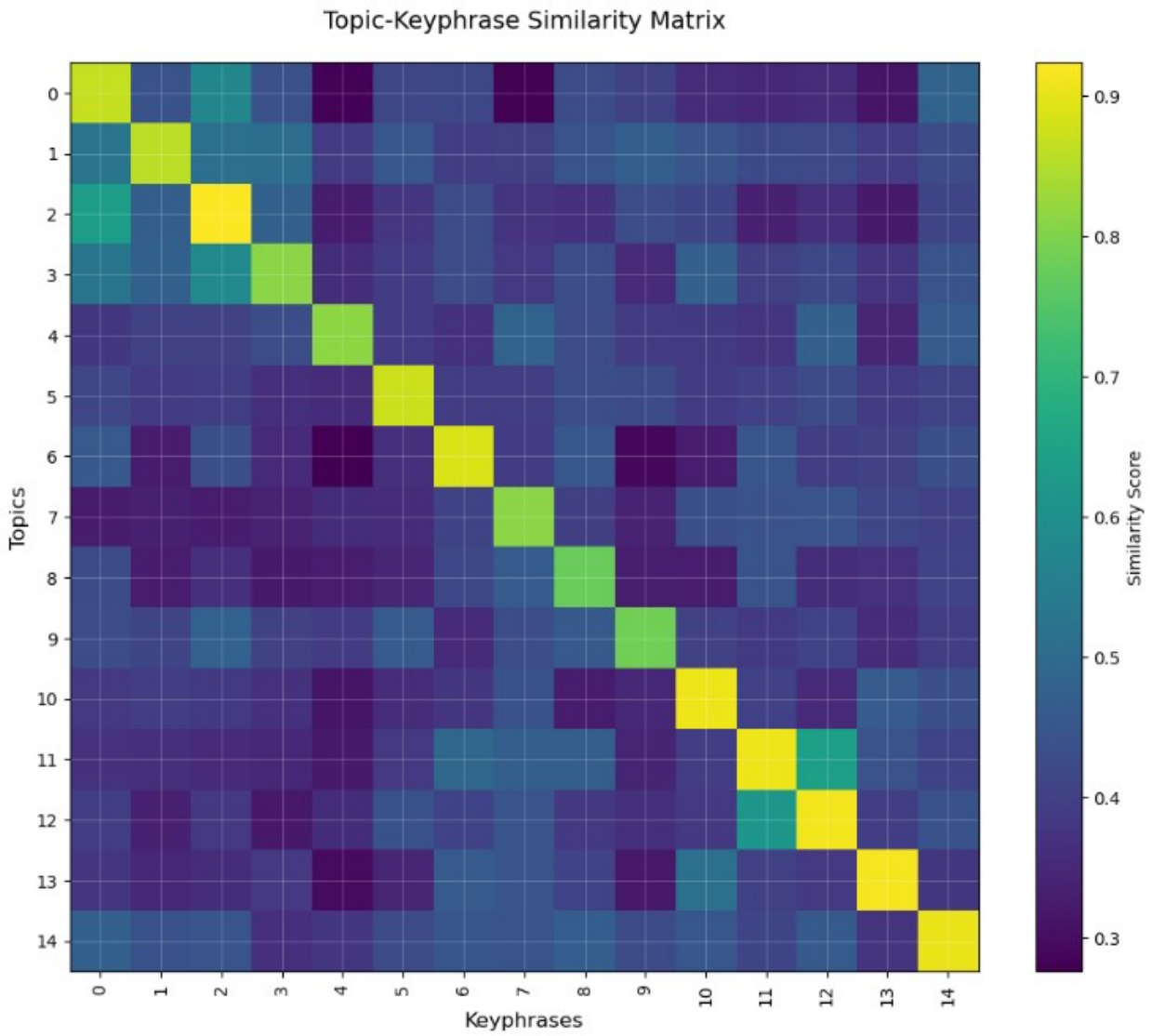

Figure 7.2: Word Embedding-Based Centroid Similarity (WECS) Matrix for Keyphrases and Topic Keywords. The matrix demonstrates the alignment of LLM-generated keyphrases with their respective topic keywords, showing strong semantic alignment.

Interpretation of Results The analysis of WECS scores across labels, keyphrases, and descriptions reveals the following insights:

- **Labels**: The WECS matrix indicates strong alignment with topic keywords, confirming that the generated labels effectively capture the core themes of the topics.

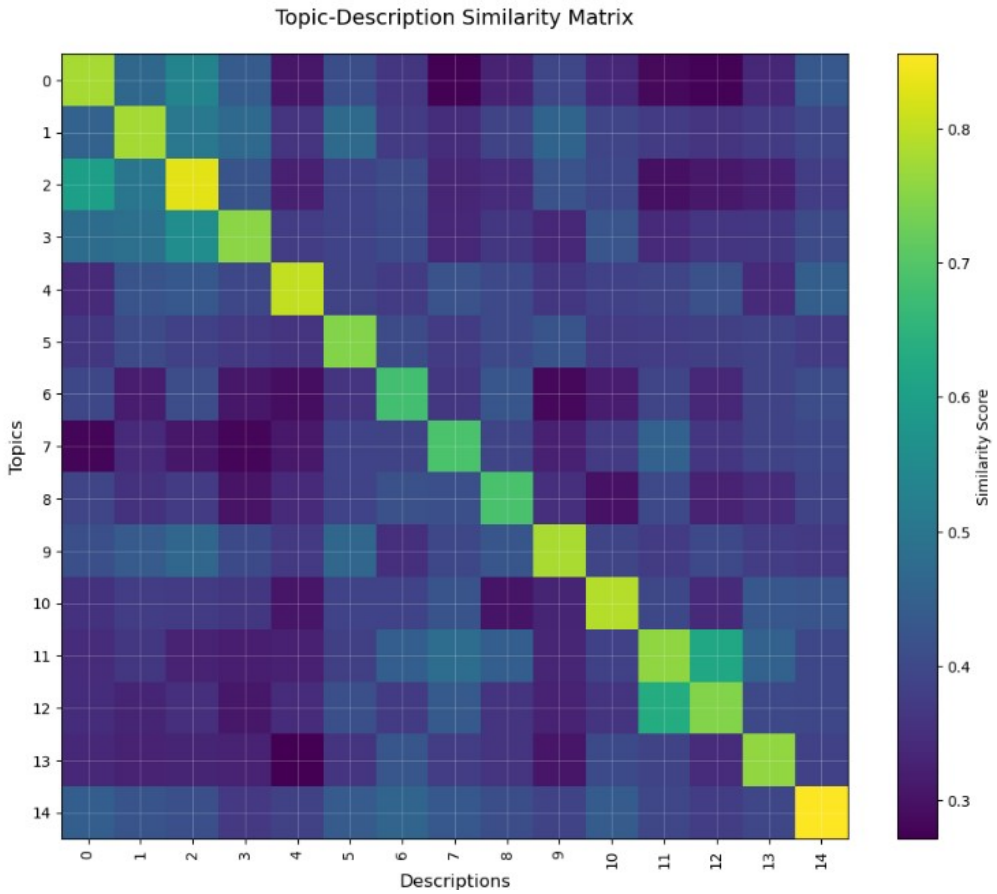

Figure 7.3: Word Embedding-Based Centroid Similarity (WECS) Matrix for Descriptions and Topic Keywords. The matrix highlights the alignment of LLM-generated descriptions with their corresponding topic keywords, demonstrating their semantic relevance.

- **Keyphrases**: The matrix shows that keyphrases offer a granular understanding of topics by capturing detailed sub-themes and relationships.

- **Descriptions**: The matrix highlights that descriptions provide coherent, contextually enriched summaries that align well with topic keywords.

These results validate the semantic coherence and relevance of LLM-generated representations, demonstrating their effectiveness in improving the interpretability of topic modeling outputs.

7.4.2 WECS Between LLM Representations and Metadata Columns

The WECS analysis between LLM-generated representations and departmental metadata revealed strong alignment across different representation types. Out of 15 topics, 11 (73.3%) showed perfect agreement across all three representation types (labels, keyphrases, and descriptions) in their top matching ProQuest department.

Representation Performance

Across all topics, descriptions showed the highest average maximum similarity (0.770), followed closely by labels (0.757), while keyphrases showed relatively lower but still substantial similarity (0.686). This pattern suggests that longer, more detailed representations may capture departmental alignments more effectively than condensed keyphrases.

Table 7.6: Topics with Perfect ProQuest Department Agreement

Topic	Matching Department	Avg. Similarity
3	Political Science	0.706
4	Marketing	0.750
6	Education	0.679
7	Physics	0.734
9	Materials Science	0.777
10	Linguistics	0.716
11	Plant Sciences	0.754
12	Petroleum Engineering	0.839
13	Geology	0.803
14	Biomedical Engineering	0.752
15	Music	0.641

Notable Cases of Variation

Four topics showed variation in departmental alignment across representation types:

- **Cultural Narratives**: While descriptions and keyphrases aligned with Film Studies (0.807, 0.733), labels matched more closely with Cultural Anthropology (0.672).
- **Health and Social Wellbeing**: Labels aligned with Sociology (0.751), while keyphrases and descriptions matched with Social Work (0.664, 0.745).
- **Urban Planning**: Showed alignment between Civil Engineering and Area Planning, reflecting the interdisciplinary nature of the topic.
- **Algorithms and Optimization**: Varied between Mechanical Engineering and Aerospace Engineering, suggesting overlap in engineering disciplines.

Table 7.7: Topics with Varied Department Alignment

Topic	Label Match	Keyphrase Match	Description Match
1	Cultural Anthropology	Film Studies	Film Studies
2	Sociology	Social Work	Social Work
5	Civil Engineering	Area Planning	Civil Engineering
8	Mech. Engineering	Aerospace Eng.	Mech. Engineering

Key Findings

The analysis revealed several important patterns:

- Strong departmental alignment across representation types, with 73.3% perfect agreement
- Higher similarity scores for descriptions (0.770) and labels (0.757) compared to keyphrases (0.686)
- Consistent identification of primary disciplinary affiliations, even in cases of variation
- Clear detection of interdisciplinary topics through varied departmental alignments

These results demonstrate the effectiveness of the LLM in generating representations that align with traditional academic departmental structures while also capturing interdisciplinary nuances.

Topic-Department Distribution Analysis

The heatmap visualization (Figure 7.4) reveals the distribution and strength of relationships between LLM-generated topic labels and academic departments. Several notable patterns emerge:

Strong Disciplinary Alignments Several topics show strong, focused alignments with specific departments:

- Topic 8 (Materials Science and Surface Chemistry) shows the strongest correlation (0.86) with its corresponding materials science department.
- Topic 11 (Reservoir Engineering) demonstrates high alignment (0.77) with petroleum engineering.
- Topic 6 (Quantum Physics) exhibits a strong correlation (0.76) with the physics department.
- Topic 9 (Language Learning) shows clear alignment (0.74) with linguistics.

Document frequency distribution Table 7.8 presents the document frequency distribution across 15 topics. Topic 1 has the highest document count with 1,048 documents, followed by Topic 2 and Topic 3 with 861 and 841 documents, respectively. On the other

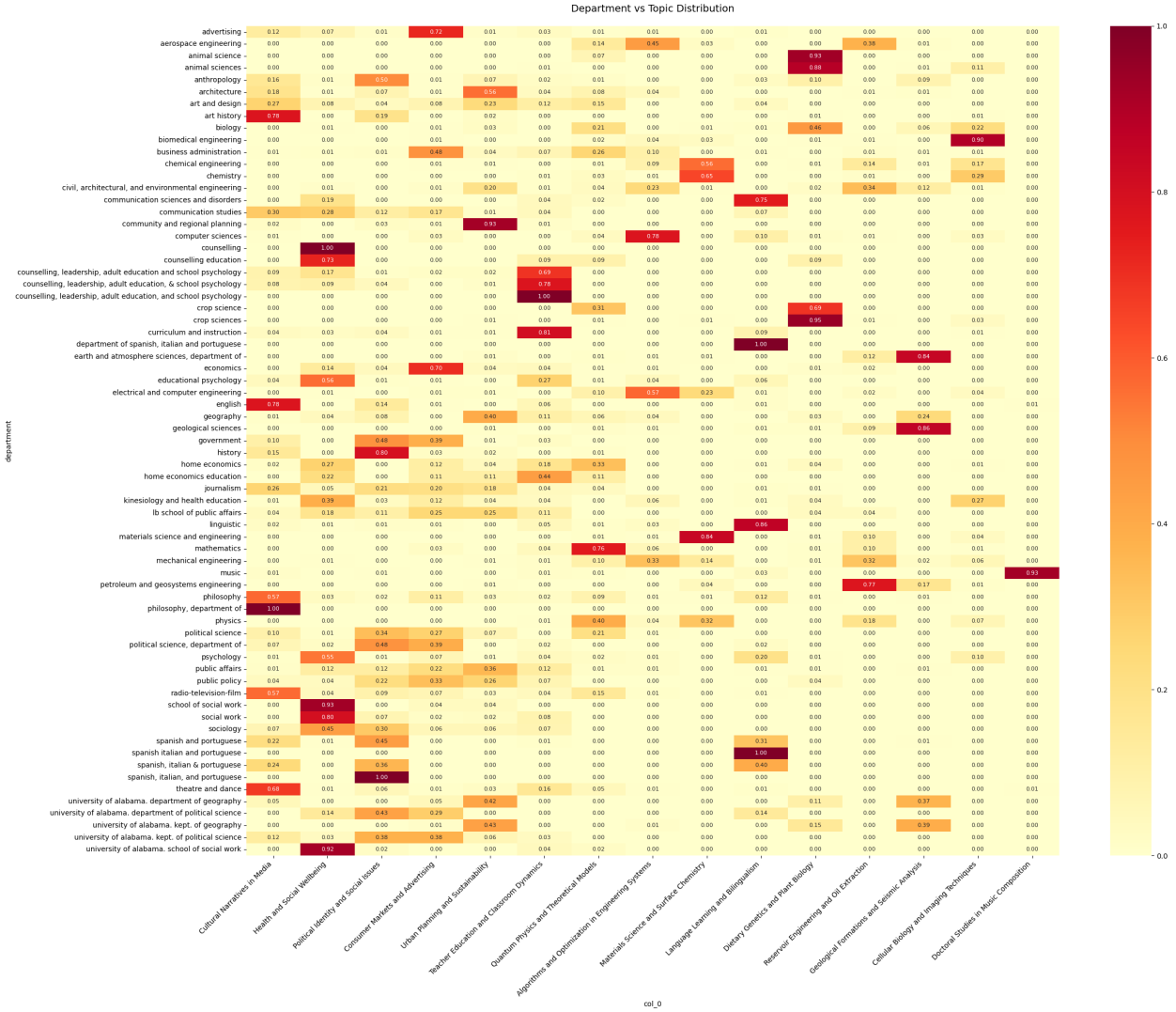

Figure 7.4: Labels vs. departments distribution

hand, Topic 15 has the lowest document count at 187. The distribution highlights a noticeable variation in topic frequencies, indicating that certain topics are more prevalent in the dataset than others.

Interdisciplinary Topics Some topics demonstrate significant correlations across multiple departments, reflecting their interdisciplinary nature:

Table 7.8: Document frequency by topic

Topic	Count
1	1048
2	861
3	841
4	796
5	649
6	634
7	633
8	593
9	573
10	552
11	533
12	528
13	504
14	468
15	187

- Topic 4 (Urban Planning) shows substantial correlations with both civil engineering (0.73) and area planning (0.69).

- Topic 7 (Algorithms and Optimization) displays strong relationships with multiple engineering departments: mechanical (0.75), aerospace (0.67), and electrical engineering (0.65).

- Topic 1 (Health and Social Wellbeing) exhibits connections across sociology (0.75), social work (0.66), and psychology (0.55).

Cross-Disciplinary Patterns The heatmap also reveals interesting cross-disciplinary relationships:

- Engineering-related topics (6, 7, 8, 11) show moderate correlations across multiple engineering departments.

- Humanities topics (0, 14) demonstrate broader, more diffuse correlations across related departments.
- Social science topics (1, 2) exhibit connections across multiple social science departments.

This distribution analysis validates the specificity of the LLM-generated topics and their ability to capture interdisciplinary relationships present in academic research.

WECS Comparison Across Models

To evaluate the performance of the five LLMs in enhancing topic representations, we compared the WECS scores for labels, keyphrases, and descriptions across GPT-4, Llama 3.1 (70B Versatile), Llama 3.2 (90B Text Preview), Llama 3.2 (3B), and MiXtral with ProQuest departments. Table 7.9 summarizes the results.

Table 7.9: WECS Comparison Across Models for LLM-Enhanced Representations

Model	**Perfect Agreement (%)**	**Labels (Avg. Similarity)**	**Keyphrases (Avg. Similarity)**	**Descriptions (Avg. Similarity)**
GPT-4	73.3%	0.757	0.686	0.770
Llama 3.1 (70B)	73.3%	**0.829**	0.741	0.755
Llama 3.2 (90B)	60.0%	0.790	**0.758**	**0.787**
Llama 3.2 (3B)	60.0%	0.790	0.743	**0.787**
Mixtral	66.7%	0.713	0.682	0.683

Analysis of Results The comparison of WECS scores reveals the following updated insights:

- **Labels:**
  - Llama 3.1 (70B Versatile) outperforms all other models with an updated average

similarity of **0.829**, emphasizing its strength in generating highly aligned and concise labels.

- GPT-4 maintains a competitive position at **0.757**, showcasing its reliability in label generation but slightly lagging behind Llama 3.1.
- Mixtral shows the lowest label similarity (**0.713**), indicating its relatively weaker capability in capturing concise thematic alignments.

- **Keyphrases:**
  - The updated scores for keyphrases reveal greater variability across models, with Llama 3.2 (90B Text Preview) achieving the highest score (**0.758**), followed closely by Llama 3.2 (3B) (**0.743**), and Llama 3.1 (**0.741**).
  - GPT-4 shows a slightly lower score (**0.686**) than these models, suggesting that open-source models may better capture granular sub-themes in keyphrase generation.
  - Mixtral performs the weakest for keyphrases at **0.682**, reflecting limited effectiveness in generating semantically rich sub-theme representations.

- **Descriptions:**
  - GPT-4 continues to generate detailed, contextually enriched descriptions with the second highest similarity score (**0.770**).
  - Llama 3.2 (90B Text Preview) and Llama 3.2 (3B) show higher similarity scores at **0.787**, demonstrating their capability to provide narrative-like summaries.
  - Mistral remains the weakest performer for descriptions with an updated score of **0.683**, highlighting the challenges of this model in producing rich, coherent contextual representations.

- **Cross-Representation Agreement:**
  - GPT-4 and Llama 3.1 (70B Versatile) achieved the highest consistency, with **73.3%** of topics showing perfect agreement across all three representations (labels, keyphrases, and descriptions) with Curated Dataset metadata.
  - Llama 3.2 models (90B and 3B) showed perfect agreement in **60.0%** of topics, indicating lower consistency across their generated representations.
  - Mistral achieved perfect agreement in **66.7%** of topics, positioning it between the highest and lowest performing models.

These updated results highlight the following:

- **Model-Specific Strengths:** Llama 3.1 dominates label generation, while GPT-4 provides the most coherent and contextually enriched descriptions. Llama 3.2 (90B Text Preview) excels in generating keyphrases.
- **Performance of Open-Source Models:** While open-source models like Llama 3.1 and 3.2 show competitive results, their performance varies more significantly across representation types compared to GPT-4.
- **Mistral's Challenges:** Mistral consistently lags across all metrics, particularly in keyphrase and description generation, underscoring its limitations in aligning to academic metadata.

7.4.3 Runtime Analysis

Table 7.10 summarizes the latency (in ms) for five models across four tasks: short labels, keyphrases, topic descriptions, and consolidated prompts. The latency comparison highlights the differences in processing efficiency for each model and task.

Table 7.10: Adjusted Latency (in ms) for Keyphrases, Short Labels, Topic Descriptions, and Consolidated Prompt

Model	Short Labels (ms)	Keyphrases (ms)	Topic Descriptions (ms)	Consolidated Prompt (ms)
GPT-4 (API)	2,850	3,325	6,650	14,250
Llama-3.2-3B	456	656	941	1,919
Llama-3.2-90B	1,272	1,402	2,903	12,171
Llama-3.1-70B	926	1,135	2,653	12,296
Mistral-8x7B	403	549	808	9,656

The results show that short labels consistently have the lowest latency for all models, with Mistral-8x7B (403 ms) and Llama-3.2-3B (456 ms) being the fastest. Keyphrases require slightly more time, while topic descriptions exhibit moderate latency, with consolidated prompts taking the longest.

Mistral-8x7B, while efficient for short labels and keyphrases, has a significantly higher latency (9,656 ms) for consolidated prompts, making it less suitable for complex tasks. In contrast, GPT-4 (API) demonstrates consistently high latency across all tasks, with the consolidated prompt taking 14,250 ms, the slowest among the models. Llama-3.2-3B offers the best overall balance of speed and task efficiency, with a maximum latency of 1,919 ms for the consolidated prompt.

7.5 Results from User Studies

This section presents preliminary findings from an ongoing user study examining topic representation preferences. The current analysis is based on feedback from **10 participants**, with recruitment continuing toward a target of 40 participants. While the small sample size limits generalizability, these initial results provide valuable directional insights.

7.5.1 Participant Distribution and Topic Selection

The preliminary results reflect contributions from participants across diverse academic fields and interests. However, the distribution of participants among topics (i.e., which of the 15 topics was selected among the 10 chosen by a participant) is uneven, leading to variations in the representativeness of results. For instance:

- Topics like *Political Identity and Social Issues* and *Urban Planning and Sustainability* were selected by **7 participants each**, as shown in Table 7.11, offering slightly more generalized insights.
- In contrast, topics such as *Dietary Genetics and Plant Biology*, with only **5 participant**, provide insights primarily reflecting limited perspectives rather than broader trends.

This disparity highlights the importance of recruiting additional participants to achieve more balanced topic coverage and reliable insights.

7.5.2 Evaluation of Topic Representations

The study evaluated topic representations based on Likert-scale metrics (accuracy, clarity, and relevance), summarized in Table 7.12. The results provide insights into how LLM-generated representations (Keyphrases, Labels, and Topic Descriptions) compare with Keywords, which are traditional topic modeling outputs.

- **Keyphrases (KP):** Performed consistently well across multiple metrics, particularly for topics such as *Health and Social Wellbeing*, where they excelled in accuracy, clarity,

Table 7.11: Topic-wise Participant Distribution

Topic	Number of Participants
1: Cultural Narratives in Media	8
2: Health and Social Wellbeing	8
3: Political Identity and Social Issues	7
4: Consumer Markets and Advertising	8
5: Urban Planning and Sustainability	7
6: Teacher Education and Classroom Dynamics	7
7: Quantum Physics and Theoretical Models	5
8: Algorithms and Optimization in Engineering Systems	6
9: Materials Science and Surface Chemistry	6
10: Language Learning and Bilingualism	7
11: Dietary Genetics and Plant Biology	5
12: Reservoir Engineering and Oil Extraction	4
13: Geological Formations and Seismic Analysis	4
14: Cellular Biology and Imaging Techniques	6
15: Doctoral Studies in Music Composition	3

and relevance. Unlike Keywords, which often lack context, Keyphrases provide richer descriptions by combining multiple related terms.

- **Labels (L):** Showed strong performance in *clearness*, with many topics ranking Labels as the clearest representation. Labels were particularly effective for topics like *Political Identity and Social Issues*, where they dominated across all metrics.
- **Topic Descriptions (TD):** Showed mixed results, performing well in some metrics but not consistently across all topics. However, Topic Descriptions provide significant depth and context beyond what Keywords alone can offer, making them particularly useful for complex topics like *Political Identity and Social Issues*.

The comparison highlights that while Keywords are helpful as a baseline, LLM-generated representations often outperform them in terms of interpretability and context. This underscores the importance of leveraging LLMs to enhance topic representations.

Table 7.12: Best Representation per Metric Across Topics. Abbreviations: K = Keywords, KP = Keyphrases, L = Labels, TD = Topic Descriptions. For more details on topic representations, refer to Table 7.5.

Topic	Accuracy	Clarity	Relevance	Clearness	Effectiveness
1	KP, TD	KP	KP	L	TD
2	L	L	L	L	L
3	KP	KP	KP	L	K
4	KP	KP	KP, L	L	KP
5	KP	TD	KP	KP	KP
6	TD	KP, L, TD	KP, L, TD	KP	TD
7	L	KP, L	KP, L	L	TD
8	L	KP, L	K, L	L	L
9	KP, TD	TD	K, KP, L, TD	TD	TD
10	K	KP	L	L	L
11	KP	KP	K	L	KP, L
12	KP	L	KP	K, TD	L, TD
13	KP	TD	K, KP, L	L	L
14	KP, L, TD	K	KP	K, KP	K
15	L	KP, L, TD	K, KP, L, TD	L	L

7.5.3 Ranking of Representations

Participant rankings for clearness and effectiveness, presented in Table 7.13, provide further insights into the preferences for topic representations. A key observation is the performance of LLM-generated representations compared to Keywords.

- **Clearness Rankings: Labels (L)** dominated the clearness rankings, appearing as the top-ranked representation in 10 out of 15 topics. **Keyphrases (KP)**, while ranking first in only two topics (Urban Planning and Teacher Education), frequently appeared as the second-best representation.
- **Effectiveness Rankings:** The variability in effectiveness rankings suggests that participant preferences depend on the complexity of the topic. For example, LLM-generated **Topic Descriptions (TD)** were particularly effective for nuanced topics

like *Urban Planning and Sustainability*, where detailed context is necessary, but Keywords performed better for straightforward topics like *Algorithms and Optimization*.

This comparison highlights that while Keywords are effective in specific scenarios, particularly for topics like *Cellular Biology and Imaging Techniques* (Topic 14), LLM-generated representations add value by providing better clarity and contextual depth. These rankings complement the Likert-scale metrics, demonstrating the versatility of LLM-based approaches.

7.5.4 Limitations and Ongoing Work

Several important caveats apply to these preliminary results:

- **Sample Size:** The current sample (n=10) provides initial insights but requires substantial expansion.
- **Topic Coverage:** Uneven distribution of participants across topics limits comparative analysis.
- **Generalizability:** Findings should be considered directional until validated with a larger sample.

Ongoing recruitment efforts focus on:

- Achieving balanced topic coverage
- Increasing statistical power
- Enabling robust comparative analysis

Table 7.13: Ranking of Representations for Clearness and Effectiveness. Abbreviations: K = Keywords, KP = Keyphrases, L = Label, TD = Topic Description. Lower ranks (1) indicate better performance. For more details on topic representations, refer to Table 7.5.

Topic	Clearness Rankings	Effectiveness Rankings
1	L > KP > K > TD	TD > KP > L > K
2	L > K > KP > TD	L > TD > KP > K
3	L > KP > K > TD	K > KP > L > TD
4	L > KP > TD > K	KP > L > TD > K
5	KP > L > K > TD	KP > TD > L > K
6	KP > L > TD > K	TD > KP > L > K
7	L > KP > TD > K	TD > KP > L > K
8	L > TD > K > KP	L > TD > K > KP
9	TD > KP > L > K	TD > KP > L > K
10	L > TD > KP > K	L > KP > TD > K
11	L > K > KP > TD	KP > L > TD > K
12	K > TD > KP > L	L > TD > K > KP
13	L > K > TD > KP	L > K > KP > TD
14	K > KP > L > TD	K > KP > L > TD
15	L > TD > K > KP	L > K > KP > TD

7.6 Results on Large Dataset

To evaluate the scalability and performance of our methodology, the BERTopic model was applied to a large dataset containing significantly more documents (333,867) and more diverse content, compared to earlier experiments. The number of topics was varied (50, 75, 100, and 125) to analyze how increased granularity affects topic coherence and diversity.

Table 7.14 presents mixed results across different metrics. While CV and NPMI scores show improvements as the number of topics increases (CV from 0.7263 to 0.7417, NPMI from 0.1151 to 0.1426), the Embedding-based Coherence remains notably low (around 0.34) across all configurations. Additionally, Topic Diversity scores show concerning trends, peaking at 75 topics (0.5827) but declining significantly at higher topic counts, reaching 0.4808 at 125 topics.

The complete list of 100 topics identified using BERTopic from the larger ETD corpus is available in Appendix B. The choice of 100 topics was made based on the balance observed between topic coherence and diversity metrics during experimentation, as detailed in Table 7.14. This configuration provides a meaningful granularity while ensuring interpretability across the diverse ETD corpus.

Topics	CV Score	NPMI Score	Embedding Coherence	Topic Diversity
50	0.7263	0.1151	0.3364	0.5260
75	0.7343	0.1304	0.3440	0.5827
100	0.7379	0.1369	0.3472	0.5260
125	0.7417	0.1426	0.3407	0.4808

Table 7.14: Comparison of Topic Models with Different Numbers of Topics. While CV and NPMI scores improve with more topics, low Embedding Coherence and declining Topic Diversity suggest limitations in the current approach.

These results reveal important limitations in BERTopic's current implementation for large-scale academic datasets. While the model demonstrates scalability and maintains reasonable CV and NPMI scores, the low embedding coherence and declining topic diversity suggest areas for improvement. These limitations likely stem from two key components: the embedding strategy and the clustering mechanism.

The current approach uses k-means clustering on embeddings, which forces all documents into distinct clusters. While this ensures complete coverage, it may not optimally capture the semantic relationships in a large, diverse academic corpus. Several potential improvements

could address these limitations:

- **Enhanced Embedding Techniques:** Investigating domain-adapted embeddings or hierarchical embedding structures could better capture the nuanced relationships in academic text, potentially improving the low embedding coherence scores.

- **Alternative Clustering Strategies:** Exploring other clustering methods or improving the performance of current clustering strategy might better preserve topic diversity at higher topic counts while maintaining coherence. Methods like hierarchical k-means could offer more flexible clustering solutions.

- **Hybrid Approaches:** Combining multiple embedding types or implementing ensemble clustering methods could help balance the trade-off between coherence and diversity.

These findings suggest that while our methodology can scale to larger datasets, significant opportunities exist for improving its performance through refined embedding and clustering techniques. Future work should focus on developing more sophisticated approaches tailored to the unique characteristics of large-scale academic corpora.

7.7 Discussion and Implications

The experimental results demonstrate several key findings regarding topic modeling of academic texts and the effectiveness of LLM enhancements. This section synthesizes these findings and discusses their broader implications.

7.7.1 Model Performance and Architecture Implications

The superior performance of BERTopic across metrics (CV scores of 0.746–0.778 with preprocessing) compared to traditional approaches highlights the advantages of transformer-based architectures for academic text analysis. The significant gap between BERTopic and NeuralLDA (CV scores of 0.409–0.533) suggests that merely incorporating neural architectures without contextual understanding is insufficient for capturing the complexity of heterogeneous datasets like ETDs.

A key strength of BERTopic lies in its ability to cluster documents in the embedding space. By leveraging transformer-based embeddings, BERTopic captures nuanced semantic relationships between documents, resulting in high coherence scores. However, this clustering mechanism may also contribute to slightly reduced diversity compared to LDA. In BERTopic, dense clustering around semantic centroids refines topics, prioritizing specificity over breadth. This contrasts with LDA's statistical approach, which often produces a broader range of topics by relying solely on word co-occurrence patterns but takes shorter computational time compared to other topic modeling approaches.

A notable trade-off emerged between coherence and diversity, suggesting that different models may be optimal for different use cases:

- **BERTopic/CTM:** The transformer-based embedding space clustering makes these models ideal for applications requiring precise, interpretable topics, such as academic search and recommendation systems.

- **Traditional LDA:** Better suited for exploratory analysis, particularly when broader topic coverage is prioritized, as it emphasizes diversity by generating more dispersed clusters.

Thus, the clustering process in BERTopic plays a dual role: enhancing semantic coherence through dense and contextually meaningful clusters while slightly constraining topic diversity. This trade-off underscores the importance of selecting models based on the application's specific requirements.

Table 7.15: Timing Results for BERTopic and Keyword Extraction on Curated ETD and Large Datasets

Method	**Dataset Size**	**Time (hh:mm:ss)**
BERTopic	9,400	00:24:00
	333,867	03:48:00
Keyword Extraction	9,400	00:22:00
	333,867	04:20:00

Table 7.15 summarizes the runtime performance of BERTopic and the keyword extraction process on both curated and large datasets. The results indicate that while BERTopic scales efficiently with increased dataset size, the time required for processing large datasets remains a critical consideration for practical applications. These findings highlight the trade-offs between scalability and computational efficiency when applying transformer-based topic modeling methods to large academic corpora.

7.7.2 Impact of Preprocessing Strategies

The differential impact of preprocessing across models reveals complex patterns that aren't simply tied to model architecture. While transformer-based models like BERTopic showed notable improvements in some configurations (e.g., CV score increase from 0.728 to 0.778 at k=15), the benefits were most apparent at lower topic counts (k<50) and near optimal topic numbers. For instance, CTM showed the strongest improvements at k=10-15, with CV score gains diminishing at higher topic counts. Traditional LDA exhibited a similar pattern, with preprocessing benefits concentrating at lower topic counts and showing diminishing returns

beyond k=50.

The implementation of preprocessing strategies, including tailored stopword lists and text enhancement techniques, showed varying effectiveness across different evaluation metrics. Coherence metrics (CV Score, C_NPMI) demonstrated more pronounced improvements at lower topic counts compared to diversity metrics, which remained relatively stable regardless of topic count. However, the Embedding Topic Coherence (ETC) metric showed inconsistent effects across models, with no clear trend, highlighting that preprocessing's impact on semantic coherence is complex and model-dependent. These findings suggest that preprocessing investments should be guided by specific performance objectives and carefully considered in relation to the chosen number of topics. The optimal preprocessing strategy should account for both the intended application context and the target number of topics, with particular attention to performance gains near the optimal k value range where improvements are most substantial.

7.7.3 LLM Enhancement Effectiveness

LLM-enhanced representations for the Curated Dataset collection demonstrated significant potential for improving the interpretability and clarity of topic modeling outputs, particularly for academic datasets. Preliminary feedback from the user study highlighted participant preferences for LLM-generated Keyphrases (KP) and Labels (L), which were perceived as more relevant and contextually rich compared to traditional Keywords (K).

The WECS analysis validated the semantic alignment of these representations with the original topics and ProQuest department labels, showcasing strengths in coherence and relevance. Among the tested models, GPT-4 excelled in generating contextually rich descriptions, while Llama variants showed strengths in label and keyphrase generation. These results highlight

the complementary strengths of different models in generating diverse topic representations.

However, the findings also underscore the importance of further validation. The small sample size in the current study limits generalizability, and ongoing recruitment efforts aim to provide more robust insights. As the survey expands, we anticipate confirming the scalability and adaptability of LLM-enhanced representations for broader datasets and applications in academic topic modeling.

In summary, the results emphasize the utility of LLMs in enhancing topic modeling outputs while pointing to the need for careful model selection and further participant validation to optimize representation effectiveness.

7.7.4 Scalability and Practical Implications

Our experiments with 333,867 documents reveal important insights about scaling topic modeling for large academic repositories. While traditional coherence metrics showed positive trends (CV scores increasing from 0.7263 to 0.7417 with higher topic counts), the consistently low embedding coherence (around 0.34) and declining topic diversity (from 0.5827 at 75 topics to 0.4808 at 125 topics) highlight critical challenges in maintaining semantic quality at scale. These contrasting results suggest that conventional evaluation metrics might not fully capture the complexities of large-scale topic modeling performance.

The divergent behavior across different metrics raises important questions about the effectiveness of current embedding and clustering approaches. The k-means clustering algorithm, while computationally efficient for large datasets, appears to struggle with preserving topic diversity as granularity increases. This limitation could be particularly problematic for academic repositories, where maintaining distinct, well-separated topics is crucial for effective content organization and discovery. These findings have broader implications for the field

of academic content management. First, they challenge the assumption that traditional topic modeling approaches can scale linearly without significant adaptation. The deterioration in topic diversity suggests that simply increasing the number of topics may not be a viable strategy for handling larger collections. Second, the consistently low embedding coherence indicates that current embedding techniques might not adequately capture the nuanced relationships present in academic texts. Our results suggest several key considerations for institutions considering large-scale implementations. Rather than focusing solely on computational efficiency, organizations might need to invest in more sophisticated clustering approaches or domain-adapted embeddings to achieve desired performance levels. The trade-offs between processing speed, topic coherence, and diversity also indicate that institutions should carefully align their implementation choices with specific use-case requirements. Future research could explore more adaptive approaches to handle these scalability challenges. This might include investigating hierarchical clustering methods that better preserve topic diversity, developing domain-specific embedding techniques, or creating hybrid approaches that combine multiple clustering strategies. Such advancements could help bridge the gap between computational feasibility and semantic quality in large-scale academic topic modeling.

7.7.5 Limitations

The current study faces several limitations that should be considered when interpreting its findings. A significant constraint lies in the framework's temporal analysis capabilities. The current implementation does not explicitly model temporal dynamics, which limits its ability to track and analyze evolving research trends over time. This limitation affects the framework's utility for understanding the progression and evolution of academic discourse across different periods.

The study's exploration of user interaction possibilities remains limited. While the framework demonstrates strong performance in automated analysis, it does not incorporate interactive refinement features or integration into a topic modeling-aided search and browse toolkit for ETDs. A user-centered design approach could enable functionalities such as merging, splitting, or renaming topics, as well as customizing search filters based on user feedback. For instance, researchers could refine topics to better align with their queries, improving the relevance of retrieved documents. Such enhancements would not only make the framework more practical but also promote user engagement, facilitating more effective exploration and discovery within ETD collections.

A notable constraint in the current evaluation is the limited number of user study participants (n=10), all of whom were graduate students. Although recruitment is ongoing, the small sample size and limited demographic diversity restrict our ability to draw definitive conclusions about user preferences and the effectiveness of different topic representations across various academic disciplines. This mainly affects our understanding of how different user groups interact with and interpret the enhanced topic representations. For example, while our preliminary observations suggest that graduate students prefer topic descriptions for STEM topics over non-STEM ones, we lack insights into how faculty members might utilize these representations in their research workflows or how librarians might leverage them for collection development. Similarly, while interdisciplinary researchers might benefit from expanded topic descriptions that highlight cross-domain connections, we need participation from this user group to validate this hypothesis. A broader participant pool, including faculty, librarians, and researchers from various disciplines, would provide valuable perspectives on how different academic roles engage with and benefit from these topic representations.

While these limitations do not diminish the current contributions, they point to important areas for future research and development. Addressing these constraints could significantly

enhance the framework's utility and applicability in academic content organization and discovery.

7.7.6 Concluding Remarks

The results demonstrate that combining transformer-based topic modeling with LLM enhancements significantly improves both technical quality and practical utility for academic text analysis. BERTopic's superior performance, particularly when enhanced with LLM-generated representations, shows the potential of integrating modern language models in topic modeling. The framework addresses critical challenges in organizing and discovering academic content by achieving high coherence and interpretability while maintaining scalability. The successful application to a large-scale metadata dataset of over 300,000 documents validates its practical viability for institutional repositories. These insights lay a strong foundation for further exploration, particularly in better integrating interactive capabilities and temporal analysis to serve the evolving needs of academic content organization.

Chapter 8

Summary, Conclusions, and Future Work

8.1 Summary and Conclusions

This thesis investigates the enhancement of topic modeling techniques for heterogeneous digital libraries, explicitly focusing on Electronic Theses and Dissertations (ETDs). The study addresses three key research questions (RQs) supported by associated hypotheses (H1, H2, H3). It contributes to developing and validating advanced methodologies for improving topic coherence, interpretability, and scalability in academic collections.

The first research question (RQ1) examines the influence of custom preprocessing techniques on the performance of various topic modeling architectures when applied to ETD corpora. The hypothesis (H1) posits that domain-specific preprocessing methods can improve coherence metrics and topic diversity. In most cases, the findings confirmed this hypothesis, revealing a more complex relationship between preprocessing and model performance. The impact was most pronounced near the optimal number of topics, with benefits diminishing at higher topic numbers. For instance, BERTopic showed significant improvements in CV scores at k=15 (from 0.728 to 0.778), while CTM demonstrated larger gains at k=50 (CV score increase of 0.098) compared to k=15 (0.087 increase). Traditional LDA showed moderate improvements at lower topic counts, with diminishing returns as topic counts increased be-

yond k=50. These results suggest that the effectiveness of preprocessing strategies is closely tied to the chosen number of topics and varies across model architectures. A specialized preprocessing pipeline was developed to observe these nuanced relationships in academic texts, with its design and impact detailed in Section 7.3.

The second research question (RQ2) explores the comparative strengths and limitations of traditional, neural, and transformer-based topic modeling approaches in handling ETD corpora. Hypothesis H2 asserts that transformer-based methods, such as BERTopic, achieve higher coherence scores than traditional probabilistic models or neural architectures. This hypothesis was validated through a comprehensive evaluation of five topic modeling approaches. BERTopic consistently achieved the highest coherence scores (CV: 0.746–0.778), demonstrating its superior capability to capture semantic relationships. However, traditional LDA maintained better topic diversity (>0.9), illustrating the trade-offs between coherence and diversity. Additionally, the scalability of the proposed methods was demonstrated by applying them to a large dataset of 333,867 ETD records. The performance analysis is presented in Sections 7.2 and 7.6.

The third research question (RQ3) focuses on how Large Language Models (LLMs) can enhance topic representations while maintaining semantic alignment with original topics. Hypothesis H3 suggests that integrating LLMs improves topic interpretability without compromising semantic coherence. A novel framework was developed to combine topic modeling with LLMs, yielding solid results. For instance, topic representations showed a 73.3% perfect agreement with departmental metadata, while semantic alignment achieved an average similarity score of 0.770 for topic descriptions. Preliminary user studies, though limited in scale (n=10), provided encouraging feedback on representation effectiveness, with participants showing a preference for LLM-generated representations over traditional keyword representations for complex academic topics. The study further evaluated the strengths of different

LLMs, such as GPT-4, Llama 3 variants, and a Mistral (MoE) model for various topic representations. Embedding-based evaluation metrics, such as Word Embedding-based Centroid Similarity (WECS), were introduced to measure these improvements quantitatively. The details of these experiments and results are discussed in Section 7.4.

Contributions: This research makes several significant contributions. A robust preprocessing framework tailored to academic texts was developed, improving topic modeling performance across multiple architectures. The comparative evaluation of traditional, neural, and transformer-based models revealed distinct trade-offs between coherence and diversity, providing valuable insights into their strengths and limitations. Integrating LLMs into topic modeling workflows demonstrated their potential for enhancing topic interpretability and semantic alignment. Though limited in scale (n=10), preliminary user studies initially validated the improved topic representations, with participants showing preferences for LLM-generated keyphrases and short labels over traditional keyword representations. The quantitative evaluation through WECS analysis complemented these qualitative insights, demonstrating strong semantic alignment between enhanced representations and original topics. Finally, the scalability of the proposed methods was validated on a large-scale ETD dataset, showcasing their practicality for real-world applications. These findings lay the groundwork for further research into advanced topic modeling methods and their applications in heterogeneous digital libraries.

8.2 Future Work

While this research demonstrates significant improvements in topic modeling for digital libraries, our findings and limitations suggest several promising directions for future research.

8.2.1 Temporal Analysis of Research Trends

The current framework could be extended to incorporate temporal dynamics in topic evolution. We could track research themes' emergence, evolution, and potential decline within academic collections by analyzing how topics change over time. This temporal modeling would be particularly valuable for understanding research trend trajectories and could help identify emerging interdisciplinary areas. Additionally, analyzing citation patterns about topic evolution could provide insights into the influence and spread of research themes across disciplines.

8.2.2 Integration with Real-Time Services

Implementing this framework to enhance real-time searching, browsing, and recommendation capabilities is a crucial next step. The enhanced topic representations could be used to develop more intuitive navigation interfaces, allowing users to explore related works through topic-based clustering. Furthermore, personalized recommendation systems could leverage these improved topic representations to suggest relevant academic works based on users' research interests and browsing patterns. This integration would require developing efficient updating mechanisms to handle new documents and user interactions in real time.

8.2.3 Advanced Embedding Models and Representations

Future work could explore integrating more powerful embedding models as they become available. For instance, domain-specific academic language models or specialized scientific embeddings could improve topic coherence and interpretation. Investigating multi-modal embeddings incorporating figures, tables, and mathematical notation could provide richer

topic representations, which is particularly valuable for STEM disciplines. This could lead to a more comprehensive understanding of academic content beyond text-based analysis.

8.2.4 Advanced Prompting and Representation Strategies

While our current framework effectively uses structured prompting to generate labels, keyphrases, and descriptions, future work could explore more advanced representation strategies. Future research could investigate dynamic prompt routing systems that automatically select the most appropriate LLM for different aspects of topic representation based on their demonstrated strengths. Additionally, the framework could include more specialized academic representations, such as methodology-focused descriptions for STEM topics or theoretical framework summaries for social sciences. This could involve developing discipline-specific prompt templates that capture different academic fields' unique characteristics and terminology while maintaining a solid departmental alignment.

8.2.5 Enhanced User Interaction and Evaluation

Building on our preliminary user study, future work should explore more comprehensive user interaction mechanisms. This includes expanding the evaluation to a larger participant base across diverse academic disciplines, developing interactive topic refinement interfaces, and investigating how user groups interact with enhanced topic representations. Such research would provide valuable insights into the practical effectiveness of different topic representation strategies and help optimize the framework for various academic contexts.

These future directions aim to enhance our framework's practical utility and advance the broader field of digital library systems. By addressing these areas, we can work toward more accessible, efficient, and comprehensive academic content discovery systems.

Bibliography

[1] Institute of Museum and Library Services (IMLS). Improving Digital Access and Inclusion, 2024. URL https://www.imls.gov/our-work/priority-areas/digital-initiatives. [Accessed: March 15, 2023].

[2] Virginia Tech. ETDs: Virginia Tech Electronic Theses and Dissertations, 2023. URL https://hdl.handle.net/10919/5534. [Accessed: August 10, 2023].

[3] Kyle Williams, Jian Wu, Zhaohui Wu, and C. Lee Giles. Information Extraction for Scholarly Digital Libraries. In *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*, JCDL '16, pages 287–288, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342292. doi: 10.1145/2910896.2925430. URL https://doi.org/10.1145/2910896.2925430. [Accessed: January 2, 2024].

[4] David Newman, Youn Noh, Edmund Talley, Sarvnaz Karimi, and Timothy Baldwin. Evaluating topic models for digital libraries. pages 215–224, 06 2010. doi: 10.1145/1816123.1816156.

[5] J. Cain. Using Topic Modeling To Enhance Access To Library Digital Collections. *Journal Of Web Librarianship*, 10:210–225, 2016. doi: 10.1080/19322909.2016.1193455.

[6] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, and et al. GPT-4 Technical Report, 2024. URL https://arxiv.org/abs/2303.08774. [Accessed: May 2, 2024].

[7] Meta AI. LLaMA 3: Vision and Edge AI for Mobile Devices. https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/, October 2024. [Accessed: November 6, 2024].

[8] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and et al. The Llama 3 Herd of Models, 2024. URL https://arxiv.org/abs/2407.21783. [Accessed: September 10, 2024].

[9] Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mixtral of experts, 2024. URL https://arxiv.org/abs/2401.04088.

[10] Sami Uddin, Bipasha Banerjee, Jian Wu, William A. Ingram, and Edward A. Fox. Building A large collection of multi-domain electronic theses and dissertations. In Yixin Chen, Heiko Ludwig, Yicheng Tu, Usama M. Fayyad, Xingquan Zhu, Xiaohua Hu, Suren Byna, Xiong Liu, Jianping Zhang, Shirui Pan, Vagelis Papalexakis, Jianwu Wang, Alfredo Cuzzocrea, and Carlos Ordonez, editors, *2021 IEEE International Conference on Big Data (Big Data), Orlando, FL, USA, December 15-18, 2021*, pages 6043–6045, Orlando, Florida, USA, 2021. IEEE. doi: 10.1109/BigData52589.2021.9672058. URL https://doi.org/10.1109/BigData52589.2021.9672058.

[11] Satvik Chekuri, Prashant Chandrasekar, Bipasha Banerjee, Sung Hee Park, Nila Masrourisaadat, Aman Ahuja, William A. Ingram, and Edward A. Fox. Integrated digital

library system for long documents and their elements. In *2023 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 13–24, 2023. doi: 10.1109/JCDL57899.2023.00012.

[12] Ike Vayansky and Sathish A.P. Kumar. A review of topic modeling methods. *Information Systems*, 94:101582, 2020. ISSN 0306-4379. doi: https://doi.org/10.1016/j.is.2020.101582. URL https://www.sciencedirect.com/science/article/pii/S0306437920300703.

[13] Bhagyashree Vyankatrao Barde and Anant Madhavrao Bainwad. An overview of topic modeling methods and tools. In *2017 International Conference on Intelligent Computing and Control Systems (ICICCS)*, pages 745–750, 2017. doi: 10.1109/ICCONS.2017.8250563.

[14] Levent Bolelli, Seyda Ertekin, Ding Zhou, and C. Lee Giles. Finding Topic Trends In Digital Libraries. In *Proceedings Of The 9th ACM/IEEE-CS Joint Conference On Digital Libraries*, JCDL '09, pages 69–72, New York, NY, USA, 2009. Association for Computing Machinery. ISBN 9781605583228. doi: 10.1145/1555400.1555411. URL https://doi.org/10.1145/1555400.1555411.

[15] Federico Bianchi, Silvia Terragni, and Dirk Hovy. Pre-training Is A Hot Topic: Contextualized Document Embeddings Improve Topic Coherence. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 759–766, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-short.96. URL https://aclanthology.org/2021.acl-short.96.

[16] Adji B. Dieng, Francisco J. R. Ruiz, and David M. Blei. Topic Modeling in Embedding

Spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453, 2020. doi: 10.1162/tacl_a_00325. URL https://aclanthology.org/2020.tacl-1.29. [Accessed: November 25, 2023].

[17] Petros Karvelis, Dimitris Gavrilis, George Georgoulas, and Chrysostomos Stylios. Topic recommendation using Doc2Vec. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–6, 2018. doi: 10.1109/IJCNN.2018.8489513.

[18] Dhiraj Vaibhav Bagul and Sunita Barve. A Novel Content-Based Recommendation Approach Based On LDA Topic Modeling For Literature Recommendation. In *2021 6th International Conference on Inventive Computation Technologies (ICICT)*, pages 954–961, 2021. doi: 10.1109/ICICT50816.2021.9358561.

[19] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990. doi: https://doi.org/10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9. URL https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291097-4571%28199009%2941%3A6%3C391%3A%3AAID-ASI1%3E3.0.CO%3B2-9. [Accessed: March 20, 2023].

[20] Thomas Hofmann. Probabilistic Latent Semantic Indexing. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '99, pages 50—57, New York, NY, USA, 1999. Association for Computing Machinery. ISBN 1581130961. doi: 10.1145/312624.312649. URL https://doi.org/10.1145/312624.312649. [Accessed: March 20, 2023].

[21] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet Allocation. *J. Mach. Learn. Res.*, 3:993—1022, Mar 2003.

[22] Kamal Nigam, Andrew McCallum, Sebastian Thrun, and Tom Mitchell. Text Classification from Labeled and Unlabeled Documents using EM. *Machine Learning*, 39: 103–134, May 2000. doi: 10.1023/A:1007692713085.

[23] Yee Teh, Michael Jordan, Matthew Beal, and David Blei. Hierarchical Dirichlet Processes. *Machine Learning*, pages 1–30, Dec 2006. doi: 10.1198/016214506000000302.

[24] David M. Blei and John D. Lafferty. Dynamic Topic Models. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 113—120, New York, NY, USA, 2006. Association for Computing Machinery. ISBN 1595933832. doi: 10.1145/1143844.1143859. URL https://doi.org/10.1145/1143844.1143859.

[25] Federico Bianchi, Silvia Terragni, Dirk Hovy, Debora Nozza, and Elisabetta Fersini. Cross-lingual Contextualized Topic Models with Zero-shot Learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1676–1683, Online, April 2021. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2021.eacl-main.143. [Accessed: October 14, 2023].

[26] Akash Srivastava and Charles Sutton. Autoencoding Variational Inference for Topic Models. *arXiv preprint arXiv:1703.01488*, 2017. URL https://arxiv.org/abs/1703.01488. [Accessed: October 14, 2023].

[27] Maarten Grootendorst. BERTopic: Neural Topic Modeling with a Class-Based TF-IDF Procedure. *arXiv preprint arXiv:2203.05794*, 2022.

[28] Stuart Lloyd. Least Squares Quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982.

[29] James MacQueen et al. Some methods for classification and analysis of multivariate

observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967.

[30] Leland McInnes, John Healy, and Steve Astels. HDBSCAN: Hierarchical density based clustering. *Journal of Open Source Software*, 2(11):205, 2017. doi: 10.21105/joss.00205. URL https://doi.org/10.21105/joss.00205.

[31] Douglas Reynolds. *Gaussian Mixture Models*, pages 659–663. Springer US, Boston, MA, 2009. ISBN 978-0-387-73003-5. doi: 10.1007/978-0-387-73003-5_196. URL https://doi.org/10.1007/978-0-387-73003-5_196.

[32] Joe H Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244, 1963.

[33] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, pages 226—231. AAAI Press, 1996.

[34] Xiangyang Liu, Tianxiang Sun, Junliang He, Jiawen Wu, Lingling Wu, Xinyu Zhang, Hao Jiang, Zhao Cao, Xuanjing Huang, and Xipeng Qiu. Towards efficient NLP: A standard evaluation and a strong baseline. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3288–3303, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.240. URL https://aclanthology.org/2022.naacl-main.240.

[35] Diksha Khurana, Aditya Koli, Kiran Khatter, and Sukhdev Singh. Natural Language Processing: State Of The Art, Current Trends And Challenges. *Multime-*

dia Tools And Applications, 82(3):3713–3744, 2022. ISSN 1573-7721. doi: 10.1007/s11042-022-13428-4. URL http://dx.doi.org/10.1007/s11042-022-13428-4.

[36] Julia Hirschberg and Christopher D. Manning. Advances in natural language processing. *Science*, 349(6245):261–266, 2015. doi: 10.1126/science.aaa8685. URL https://www.science.org/doi/abs/10.1126/science.aaa8685.

[37] Ming Zhou, Nan Duan, Shujie Liu, and Heung yeung Shum. Progress in Neural NLP: Modeling, Learning, and Reasoning. *Engineering*, 2020. URL https://api.semanticscholar.org/CorpusID:213572808. [Accessed: May 21, 2024].

[38] Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heinz, and Dan Roth. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, 56:1 – 40, 2021. URL https://api.semanticscholar.org/CorpusID:240420063.

[39] Jamie Zimmermann, Lance E. Champagne, John M. Dickens, and Benjamin T. Hazen. Approaches To Improve Preprocessing For Latent Dirichlet Allocation Topic Modeling. *Decision Support Systems*, 185:114310, 2024. ISSN 0167-9236. doi: https://doi.org/10.1016/j.dss.2024.114310. URL https://www.sciencedirect.com/science/article/pii/S016792362400143X.

[40] Rob Churchill and Lisa Singh. TextPrep: A Text Preprocessing Toolkit For Topic Modeling On Social Media Data. In *Proceedings Of The 10th International Conference On Data Science, Technology And Applications*, 2021.

[41] Christine P. Chai. Comparison of text preprocessing methods. *Natural Language Engineering*, 29(3):509–553, 2023. doi: 10.1017/S1351324922000213.

[42] Susan Dumais, John Platt, David Heckerman, and Mehran Sahami. Inductive Learning Algorithms and Representations for Text Categorization. In *Proceedings of the Seventh International Conference on Information and Knowledge Management*, pages 148–155, 1998.

[43] Murugan Anandarajan, Chelsey Hill, and Thomas Nolan. *Text Preprocessing*, chapter 4, pages 45–59. Springer International Publishing, Cham, 2019. ISBN 978-3-319-95663-3. doi: 10.1007/978-3-319-95663-3_4. URL https://doi.org/10.1007/978-3-319-95663-3_4. [Accessed: June 18, 2023].

[44] Steven Struhl. *Practical Text Analytics: Interpreting Text and Unstructured Data for Business Intelligence*. Kogan Page Publishers, 2015.

[45] S Vijayarani, R Janani, et al. Text mining: open source tokenization tools-an analysis. *Advanced Computational Intelligence: An International Journal (ACII)*, 3(1):37–47, 2016.

[46] Gregory Grefenstette. *Tokenization*, pages 117–133. Springer Netherlands, Dordrecht, 1999. ISBN 978-94-015-9273-4. doi: 10.1007/978-94-015-9273-4_9. URL https://doi.org/10.1007/978-94-015-9273-4_9. [Accessed: December 2023].

[47] W John Wilbur and Karl Sirotkin. The Automatic Identification of Stop Words. *Journal of Information Science*, 18(1):45–55, 1992.

[48] Gerard Salton. Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information. *Reading: Addison-Wesley*, 1989.

[49] Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. *Introduction to Information Retrieval*, volume 39. Cambridge University Press, Cambridge, 2008.

[50] Maryam Nasseri and Philip McCarthy. Structural factor analysis of lexical complexity constructs and measures: A quantitative measure-testing process on specialised academic texts. *Journal of Quantitative Linguistics*, 30:280–303, 2023. URL https://api.semanticscholar.org/CorpusID:264989518.

[51] Gert Faustmann. Improved learning of academic writing - reducing complexity by modeling academic texts. In *International Conference on Computer Supported Education*, 2018. URL https://api.semanticscholar.org/CorpusID:13681330.

[52] Alexandra Schofield, Måns Magnusson, Laure Thompson, and David M. Mimno. Pre-Processing for Latent Dirichlet Allocation. 2017. URL https://api.semanticscholar.org/CorpusID:38886757. [Accessed: June 19, 2023].

[53] S Vijayarani, Ms J Ilamathi, Ms Nithya, et al. Preprocessing techniques for text mining-an overview. *International Journal of Computer Science & Communication Networks*, 5(1):7–16, 2015.

[54] Alper Kursat Uysal and Serkan Gunal. The impact of preprocessing on text classification. *Information processing & management*, 50(1):104–112, 2014.

[55] PyPDF Contributors. PyPDF: A Pure-Python PDF Library. https://pypdf.readthedocs.io/, 2024. Version 3.0.0.

[56] Apache Software Foundation. Apache Tika. https://cwiki.apache.org/confluence/display/tika, 2024. [Accessed: October 24, 2024].

[57] Ray Smith. An overview of the Tesseract OCR engine. *Proceedings of the Ninth International Conference on Document Analysis and Recognition*, 2:629–633, 2007.

[58] Michael Röder, Andreas Both, and Alexander Hinneburg. Exploring the Space of Topic Coherence Measures. In *Proceedings of the Eighth ACM International Conference on*

Web Search and Data Mining, WSDM '15, pages 399—408, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450333177. doi: 10.1145/2684822.2685324. URL https://doi.org/10.1145/2684822.2685324. [Accessed: November 12, 2023].

[59] Emil Rijcken, Kalliopi Zervanou, Pablo Mosteiro, Floortje Scheepers, Marco Spruit, and Uzay Kaymak. Topic Specificity: A Descriptive Metric for Algorithm Selection and Finding the Right Number of Topics. *Natural Language Processing Journal*, 8:100082, 2024. doi: https://doi.org/10.1016/j.nlp.2024.100082. URL https://www.sciencedirect.com/science/article/pii/S294971912400030X. [Accessed: November 25, 2023].

[60] Hanna M. Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. Evaluation Methods for Topic Models. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 1105—1112, New York, NY, USA, 2009. Association for Computing Machinery. ISBN 9781605585161. doi: 10.1145/1553374.1553515. URL https://doi.org/10.1145/1553374.1553515. [Accessed: November 15, 2023].

[61] Jonathan Chang, Jordan Boyd-Graber, Sean Gerrish, Chong Wang, and David M. Blei. Reading tea leaves: How humans interpret topic models. In *Proceedings of the 22nd International Conference on Neural Information Processing Systems*, NIPS'09, pages 288—296, Red Hook, NY, USA, 2009. Curran Associates Inc. ISBN 9781615679119.

[62] David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. Automatic evaluation of topic coherence. In *Human language technologies: The 2010 annual conference of the North American chapter of the Association for Computational Linguistics*, pages 100–108, 2010.

[63] Nikolaos Aletras and Mark Stevenson. Evaluating Topic Coherence Using Distribu-

tional Semantics. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers*, pages 13–22, 2013. URL https://aclanthology.org/W13-0102. [Accessed: November 15, 2023].

[64] Jey Han Lau, David Newman, and Timothy Baldwin. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In Shuly Wintner, Sharon Goldwater, and Stefan Riezler, editors, *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 530–539, Gothenburg, Sweden, April 2014. Association for Computational Linguistics. doi: 10.3115/v1/E14-1056. URL https://aclanthology.org/E14-1056.

[65] Jey Han Lau and Timothy Baldwin. The Sensitivity of Topic Coherence Evaluation to Topic Cardinality. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–487, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1057. URL https://aclanthology.org/N16-1057. [Accessed: November 19, 2023].

[66] Fred Morstatter and Huan Liu. In Search of Coherence and Consensus: Measuring the Interpretability of Statistical Topics. *Journal of Machine Learning Research*, 18(169): 1–32, 2018.

[67] Shraey Bhatia, Jey Han Lau, and Timothy Baldwin. Topic Intrusion for Automatic Topic Model Evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 844–849, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1098. URL https://aclanthology.org/D18-1098. [Accessed: November 17, 2023].

[68] Dominik Stammbach, Vilém Zouhar, Alexander Hoyle, Mrinmaya Sachan, and El-

liott Ash. Revisiting Automated Topic Model Evaluation with Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9348–9357, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.581. URL https://aclanthology.org/2023.emnlp-main.581. [Accessed: November 19, 2023].

[69] Ran Ding, Ramesh Nallapati, and Bing Xiang. Coherence-Aware Neural Topic Modeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 830–836, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1096. URL https://aclanthology.org/D18-1096. [Accessed: November 17, 2023].

[70] Amin Hosseiny Marani and Eric P. S. Baumer. A review of stability in topic modeling: Metrics for assessing and techniques for improving stability. *ACM Comput. Surv.*, 56 (5), November 2023. ISSN 0360-0300. doi: 10.1145/3623269. URL https://doi.org/10.1145/3623269.

[71] Alexander Hoyle, Pranav Goel, Andrew Hian-Cheong, Denis Peskov, Jordan Boyd-Graber, and Philip Resnik. Is Automated Topic Model Evaluation Broken? The Incoherence of Coherence. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/0f83556a305d789b1d71815e8ea4f4b0-Paper.pdf. [Accessed: December 12, 2023].

[72] Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. Large Language Models: A Survey, 2024. URL https://arxiv.org/abs/2402.06196. [Accessed: October 2, 2024].

[73] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is All You Need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf. [Accessed: February 8, 2023].

[74] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*, volume 1, pages 4171—4186, 2019. URL https://aclanthology.org/N19-1423.pdf. [Accessed: 13 December, 2023].

[75] Alec Radford. Improving Language Understanding by Generative Pre-training. *OpenAI*, 2018.

[76] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language Models are Unsupervised Multitask Learners. *OpenAI Blog*, 1(8):9, 2019.

[77] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, and et al. PaLM: Scaling Language Modeling with Pathways, 2022. URL https://arxiv.org/abs/2204.02311. [Accessed: September 16, 2023].

[78] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. *Learning Internal Representations by Error Propagation*, pages 318—362. MIT Press, Cambridge, MA, USA, 1986.

[79] Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, 11 1997. doi: 10.1162/neco.1997.9.8.1735. URL https://doi.org/10.1162/neco.1997.9.8.1735. [Accessed: February 6, 2023].

[80] Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. On the Properties of Neural Machine Translation: Encoder-Decoder Approaches, 2014. URL https://arxiv.org/abs/1409.1259. [Accessed: December 19, 2023].

[81] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, and et al. Mistral 7B, 2023. URL https://arxiv.org/abs/2310.06825. [Accessed: February 23, 2024].

[82] J. Kaplan, Sam McCandlish, T. Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, S. Gray, Alec Radford, Jeff Wu, and Dario Amodei. Scaling Laws for Neural Language Models. *ArXiv*, abs/2001.08361, 2020. URL https://arxiv.org/abs/2001.08361. [Accessed: February 15, 2024].

[83] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and J. Jitsev. Reproducible Scaling Laws for Contrastive Language-Image Learning. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2829, 2022. doi: 10.1109/CVPR52729.2023.00276.

[84] Patrick Fernandes, B. Ghorbani, Xavier García, Markus Freitag, and Orhan Firat. Scaling Laws For Multilingual Neural Machine Translation. *ArXiv*, abs/2302.09650, 2023. URL https://arxiv.org/abs/2302.09650. [Accessed: February 15, 2024].

[85] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Z. Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, J. Nie, and Ji rong Wen. A Survey of Large Language Models. *ArXiv*,

abs/2303.18223, 2023. URL https://arxiv.org/abs/2303.18223. [Accessed: January 5, 2024].

[86] D.M.E. Luitse and Wiebke Denkena. The Great Transformer: Examining the Role of Large Language Models in the Political Economy of AI. *Big Data & Society*, 8, 2021. doi: 10.1177/20539517211047734.

[87] Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Bing Yin, and Xia Hu. Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond. *ArXiv*, abs/2304.13712, 2023. URL https://arxiv.org/abs/2304.13712. [Accessed: January 19, 2024].

[88] Xiangyun Zhao, Raviteja Vemulapalli, Philip Andrew Mansfield, Boqing Gong, Bradley Green, Lior Shapira, and Ying Wu. Contrastive Learning for Label Efficient Semantic Segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10623–10633, October 2021.

[89] Han Wang, Nirmalendu Prakash, Nguyen Khoi Hoang, Ming Shan Hee, Usman Naseem, and Roy Ka-Wei Lee. Prompting Large Language Models for Topic Modeling. In *2023 IEEE International Conference on Big Data (BigData)*, pages 1236–1241, 2023. doi: 10.1109/BigData59044.2023.10386113.

[90] Michael R. Douglas. Large Language Models. *Communications Of The ACM*, 66(7): 7, 2023. ISSN 0001-0782. doi: 10.1145/3606337. URL https://doi.org/10.1145/3606337. [Accessed: April 22, 2024].

[91] Alaa A. Abd-alrazaq, Rawan AlSaad, Dari Alhuwail, Arfan Ahmed, P. Healy, Syed Latifi, S. Aziz, R. Damseh, Sadam Alabed Alrazak, and Javaid Sheikh. Large Language Models in Medical Education: Opportunities, Challenges, and Future Directions. *JMIR Medical Education*, 9, 2023. doi: 10.2196/48291.

[92] Lixiang Yan, Lele Sha, Linxuan Zhao, Yuheng Li, Roberto Martínez-Maldonado, Guanliang Chen, Xinyu Li, Yueqiao Jin, and D. Gašević. Practical and Ethical Challenges of Large Language Models in Education: A Systematic Scoping Review. *British Journal of Educational Technology*, 2023. doi: 10.1111/bjet.13370.

[93] Cari Beth Head, Paul Jasper, Matthew McConnachie, Linda Raftree, and Grace Higdon. Large Language Model Applications for Evaluation: Opportunities and Ethical Implications. *New Directions for Evaluation*, 2023:33 – 46, 2023. doi: 10.1002/ev.20556.

[94] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. HuggingFace's Transformers: State-of-the-Art Natural Language Processing, 2020. URL https://arxiv.org/abs/1910.03771. [Accessed: December 19, 2023].

[95] Dennis Abts, Garrin Kimmell, Andrew Ling, John Kim, Matt Boyd, Andrew Bitar, Sahil Parmar, Ibrahim Ahmed, Roberto DiCecco, David Han, John Thompson, Michael Bye, Jennifer Hwang, Jeremy Fowers, Peter Lillian, Ashwin Murthy, Elyas Mehtabuddin, Chetan Tekur, Thomas Sohmers, Kris Kang, Stephen Maresh, and Jonathan Ross. A Software-Defined Tensor Streaming Multiprocessor for Large-Scale Machine Learning. In *Proceedings of the 49th Annual International Symposium on Computer Architecture*, ISCA '22, page 567–580, 2022. doi: 10.1145/3470496.3527405. URL https://doi.org/10.1145/3470496.3527405.

[96] Groq Inc. GroqCloud. https://groq.com/groqcloud/, 2024. [Accessed: October 15, 2024].

[97] Tianyu Ding, Tianyi Chen, Haidong Zhu, Jiachen Jiang, Yiqi Zhong, Jinxin Zhou, Guangzhi Wang, Zhihui Zhu, Ilya Zharkov, and Luming Liang. The efficiency spectrum of large language models: An algorithmic survey. *arXiv preprint arXiv:2312.00678*, 2023.

[98] Alex Smola and Shravan M. Narayanamurthy. An architecture for parallel topic models. *Proceedings of the VLDB Endowment*, 3:703 – 710, 2010. doi: 10.14778/1920841.1920931.

[99] Bird, Steven and Klein, Ewan and Loper, Edward. *Natural language processing with Python: Analyzing text with the natural language toolkit*. O'Reilly Media, Inc., 2009.

[100] Matthew Barnett. Regex: Regular Expressions for Python, 2024. URL https://pypi.org/project/regex/. Accessed: 21 December, 2024.

[101] Hugging Face. Sentence-transformers: all-mpnet-base-v2. https://huggingface.co/sentence-transformers/all-mpnet-base-v2, 2024. [Accessed: 17 November, 2024].

[102] Hugging Face. JXM: CDE-Small-v1. https://huggingface.co/jxm/cde-small-v1, 2024. [Accessed: 17 November, 2024].

[103] Hugging Face. Alibaba-NLP: GTE-Large-En-v1.5. https://huggingface.co/Alibaba-NLP/gte-large-en-v1.5, 2024. [Accessed: 18 November, 2024].

[104] Hugging Face. google-bert/bert-base-uncased. https://huggingface.co/google-bert/bert-base-uncased, 2024. Accessed: 2024-11-23.

[105] Hugging Face. Multilingual task evaluation benchmark (MTEB) leaderboard. https://huggingface.co/spaces/mteb/leaderboard, 2024. [Accessed: 20 November, 2024].

[106] Radim Rehurek and Petr Sojka. Gensim–Python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, 3(2), 2011.

[107] Silvia Terragni, Elisabetta Fersini, Bruno Giovanni Galuzzi, Pietro Tropeano, and Antonio Candelieri. OCTIS: Comparing and Optimizing Topic Models is Simple! In Dimitra Gkatzia and Djamé Seddah, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 263–270, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-demos.31. URL https://aclanthology.org/2021.eacl-demos.31. [Accessed: August 5, 2023].

[108] Ricardo J. G. B. Campello, Davoud Moulavi, and Joerg Sander. Density-based clustering based on hierarchical density estimates. In Jian Pei, Vincent S. Tseng, Longbing Cao, Hiroshi Motoda, and Guandong Xu, editors, *Advances in Knowledge Discovery and Data Mining*, pages 160–172, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg. ISBN 978-3-642-37456-2.

[109] Maarten Grootendorst. KeyBERT: Minimal Keyword Extraction with BERT, 2020. URL https://doi.org/10.5281/zenodo.4461265. [Accessed: November 19, 2023].

[110] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[111] Ken Lang. 20 newsgroups dataset, 1995. URL http://qwone.com/~jason/20Newsgroups/. Accessed: 15 March, 2024.

[112] Palakh Mignonne Jude. Increasing Accessibility Of Electronic Theses And Dissertations (ETDs) Through Chapter-Level Classification, 2020. URL http://hdl.handle.net/10919/99294. [VTechWorks; VT MS Thesis; Online; Accessed: 25-September-2020].

[113] J. P. Kincaid, R. P. Fishburne, R. L. Rogers, and B. S. Chissom. Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel. Technical Report 8-75, Naval Technical Training Command Millington in Research Branch, 1975. URL https://apps.dtic.mil/sti/citations/ADA006655. Retrieved October 15, 2024.

[114] Amr Ahmed Aboelnaga, Anushka Sivakumar, Jayanth Narla, Pradyumna Upendra Dasu, Ragul Seetharaman, Sahana Bhaskar, and Shankar Srinidhi Srinivas. Final report CS 5604: Information storage and retrieval. 2024. URL https://hdl.handle.net/10919/118665. Team 3: Object Detection and Topic Modeling (Fall 2023).

[115] S. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982. doi: 10.1109/TIT.1982.1056489.

[116] Christiane Fellbaum. WordNet: An electronic lexical database. *MIT Press google scholar*, 2:678–686, 1998.

[117] Prafull Sharma and Yingbo Li. Self-supervised contextual keyword and keyphrase retrieval with self-labelling. *Preprints*, August 2019. doi: 10.20944/preprints201908.0073.v1. URL https://doi.org/10.20944/preprints201908.0073.v1.

[118] Akash Srivastava and Charles Sutton. Autoencoding variational inference for topic models, 2017. URL https://arxiv.org/abs/1703.01488. [Accessed: November 18, 2023].

[119] Jey Han Lau and Timothy Baldwin. The Sensitivity of Topic Coherence Evaluation to Topic Cardinality. In Kevin Knight, Ani Nenkova, and Owen Rambow, editors, *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–487, San Diego, California, June 2016. ACL. doi: 10.18653/v1/N16-1057. URL https://aclanthology.org/N16-1057. [Accessed: January 17, 2024].

[120] Nikolaos Aletras and Mark Stevenson. Evaluating Topic Coherence Using Distributional Semantics. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)–Long Papers*, pages 13–22, 2013.

[121] Han Wang, Nirmalendu Prakash, Nguyen Khoi Hoang, Ming Shan Hee, Usman Naseem, and Roy Ka-Wei Lee. Prompting large language models for topic modeling, 2023. URL https://arxiv.org/abs/2312.09693.

Appendix A

IRB Documents

A.1 IRB Approval Letter

The Institutional Review Board (IRB) approved this study under protocol 24-973. The complete approval letter is included below.

Division of Scholarly Integrity and Research Compliance
Institutional Review Board
North End Center, Suite 4120 (MC 0497)
300 Turner Street NW
Blacksburg, Virginia 24061
540/231-3732
irb@vt.edu
http://www.research.vt.edu/sirc/hrpp

MEMORANDUM

DATE: October 29, 2024

TO: Edward Fox, Pradyumna Upendra Dasu

FROM: Virginia Tech Institutional Review Board (FWA00000572)

PROTOCOL TITLE: Validating Topic Labels Generated by LLMs: A User Study

IRB NUMBER: **24-973**

Effective October 29, 2024, the Virginia Tech Human Research Protection Program (HRPP) determined that this protocol meets the criteria for exemption from IRB review under 45 CFR 46.104 (d) category(ies) 2(ii).

Ongoing IRB review and approval by this organization is not required. This determination applies only to the activities described in the IRB submission and does not apply should any changes be made. If changes are made and there are questions about whether these activities impact the exempt determination, please submit an amendment to the HRPP for a determination.

This exempt determination does not apply to any collaborating institution(s). The Virginia Tech HRPP and IRB cannot provide an exemption that overrides the jurisdiction of a local IRB or other institutional mechanism for determining exemptions.

All investigators (listed above) are required to comply with the researcher requirements outlined at:

https://secure.research.vt.edu/external/irb/responsibilities.htm

(Please review responsibilities before beginning your research.)

PROTOCOL INFORMATION:

Determined As:	**Exempt, under 45 CFR 46.104(d) category(ies) 2(ii)**
Protocol Determination Date:	**October 10, 2024**

ASSOCIATED FUNDING:

The table on the following page indicates whether grant proposals are related to this protocol.

SPECIAL INSTRUCTIONS:

This is an amendment, authorized on October 29, 2024, to modify the protocol, recruitment, and consent to add in a random drawing for compensation.

Date*	OSP Number	Sponsor

* Date this proposal number was added.

If this protocol is to cover any other grant proposals, please contact the HRPP office (irb@vt.edu).

A.2 Email Recruitment Material

Below is the email recruitment material used to invite participants to the study. The IRB reviewed and approved this material as part of the study protocol.

Outlook

Invitation to participate in AI Research with Virginia Tech's Digital Library Research Lab

From QuestionPro Survey <survey@qp-mail.com>
Date Sun 11/24/2024 10:36 PM
To Dasu, Pradyumna Upendra <pradyumnaupendra@vt.edu>

Dear Students, Faculty, and Researchers,

We are conducting a research study at Virginia Tech focused on evaluating topic labels generated by large language models (LLMs). Your insights as someone with experience in academic content, particularly theses and dissertations, will be valuable for this research.

Study Details:

- **Eligibility Criteria**: Current graduate students, undergraduate students with research experience, faculty members, or researchers.
- **Purpose**: The purpose of this study is to assess the effectiveness of topic labels generated by LLMs and gather feedback to improve their relevance and clarity in academic contexts.
- **Time Commitment**: Approximately 45 minutes.
- **Procedures**: You will be asked to complete an online survey where you will evaluate topic labels and provide feedback.

We plan to recruit up to about 50 participants. After the study is complete, a randomly selected subset (at least 10% but likely more) of participants will receive a $20 Amazon gift card, that should be opened right away, but can be spent later.

Contact Information:
For more information about the study, or if you have any questions, please contact the research team at pradyumnaupendra@vt.edu.

IRB Approval:
This study has been approved by the Virginia Tech Institutional Review Board (IRB Number: **24-973**).
Please note that your participation is entirely voluntary, and you are not committing to participate by contacting the research team or expressing interest.

You will have the opportunity to review all consent information before deciding whether to volunteer for the study.
We hope you consider participating in this important research.

Thank you for your time and consideration.

If you are interested, please click on the Survey Link to start the survey.

Best regards,
Pradyumna Dasu
Virginia Tech | Computer Science
Blacksburg, VA 24061
540-231-7567 | pradyumnaupendra@vt.edu

Appendix B

BERTopic Results for Larger ETD Metadata Corpus

This appendix presents 100 topics derived from BERTopic analysis of the larger ETD corpus. These topics were generated using the same framework described in the detailed experimentation sections of the thesis. Each topic is represented by a set of 10 high-probability keywords that characterize its thematic focus, offering insights into the diverse range of subjects within the corpus.

Topic Number	**Keywords**	**Label**
0	lubbock, cotton, sorghum, vocational, mesquite, cattle, junior, diagenesis, personality, beef	Agriculture and Livestock
1	politics, colonial, empire, british, mexican, literary, narrative, indigenous, ideology, imperial	Colonial Politics and Literature
2	receptor, tumor, kinase, immune, inhibitor, apoptosis, macrophage, cellular, peptide, phosphorylation	Cellular Biology and Cancer Research

3	teacher, classroom, mathematics, math, instructional, literacy, enrolled, questionnaire, semester, english	Education and Classroom Practices
4	literary, narrative, poetry, poem, writer, fiction, poet, queer, genre, feminist	Literary Studies and Feminist Theory
5	robot, semantic, query, scheduling, robotic, runtime, adversarial, semantics, processor, kernel	Robotics and Computational Semantics
6	rna, enzyme, mutant, peptide, subunit, mrna, amino, genome, residue, trna	RNA and Molecular Biology
7	teacher, classroom, mathematics, instructional, literacy, curriculum, preservice, educator, instructor, english	Teacher Education and Curriculum Development
8	sexual, violence, racial, feminist, crime, narrative, masculinity, womens, sex, sexuality	Gender Studies and Violence
9	rna, cellular, genome, actin, transcriptional, receptor, membrane, kinase, mrna, chromatin	Genomics and Cellular Processes
10	seismic, crack, vibration, shear, fault, pavement, earthquake, deformation, rotor, stiffness	Structural Engineering and Seismic Studies

11	thermal, flame, oxidation, oxide, hydrogen, combustion, adsorption, reactor, polymer, electrochemical	Combustion and Thermal Processes
12	laser, semiconductor, polymer, quantum, graphene, electron, crystal, nanoparticles, ion, spin	Nanotechnology and Quantum Electronics
13	sexual, adolescent, african, violence, abuse, mother, racial, anxiety, sex, youth	Adolescent Psychology and Social Issues
14	shear, specimen, seismic, crack, deformation, pavement, soil, gear, stiffness, asphalt	Civil Engineering and Material Studies
15	violence, newspaper, civil, politics, refugee, right, protest, party, crime, news	Social Movements and Media Studies
16	acid, ligand, polymer, copolymer, enzyme, solvent, polymerization, ph, nmr, molecule	Polymer Chemistry and Molecular Interactions
17	robot, robotic, processor, query, simulator, scheduling, topology, java, navigation, server	Robotics and Navigation Systems
18	estimator, nonparametric, covariance, convex, asymptotic, outlier, gaussian, decoding, polynomial, imputation	Statistical Methods and Mathematical Modeling
19	turbulent, vortex, turbulence, elastic, shear, jet, reynolds, vibration, beam, seismic	Fluid Dynamics and Mechanical Vibrations

20	literary, religious, music, scholar, church, narrative, tradition, poetry, christian, religion	Religious Literature and Cultural Traditions
21	algebra, theorem, conjecture, polynomial, manifold, vertex, algebraic, integer, cohomology, lattice	Advanced Algebra and Geometry
22	robot, bug, developer, fault, processor, query, overhead, cache, scheduling, runtime	Software Faults and Robotic Systems
23	moral, literary, philosophical, narrative, ethic, nietzsche, rhetorical, poetry, rhetoric, ethical	Philosophy and Literary Ethics
24	autism, childrens, adolescent, peer, infant, comprehension, parental, disability, parenting, teacher	Child Development and Autism Studies
25	electron, quantum, atom, plasma, neutron, spin, laser, ion, scattering, spectroscopy	Quantum Mechanics and Spectroscopy
26	faculty, curriculum, leadership, teacher, administrator, superintendent, campus, career, educator, internship	Educational Leadership and Administration
27	music, musical, theatre, artist, painting, poetry, poem, dance, narrative, artistic	Performing Arts and Artistic Expression
28	music, recital, musical, composer, artist, painting, church, theatre, song, dance	Musical Composition and Performance

29	graphene, nanoparticles, polymer, semiconductor, laser, nanotube, electron, oxide, nanowires, quantum	Nanomaterials and Graphene Research
30	quantum, electron, spin, laser, graphene, semiconductor, silicon, ion, nanoscale, thermal	Semiconductor Physics and Quantum Technology
31	aircraft, radar, aerodynamic, antenna, rotor, airfoil, jet, vortex, blade, spacecraft	Aerospace Engineering and Aerodynamics
32	rna, genome, mutation, mrna, cellular, peptide, mutant, membrane, molecule, yeast	Genomics and Mutation Studies
33	estimator, theorem, manifold, polynomial, algebra, convex, semantic, asymptotic, dynamical, scheduling	Computational Mathematics and Optimization
34	polymer, membrane, thermal, coating, hydrogen, conductivity, solar, electrolyte, oxide, electrode	Energy Materials and Membrane Technology
35	tumor, mri, vivo, therapeutic, ultrasound, immune, vitro, nanoparticles, cardiac, vascular	Medical Imaging and Cancer Therapy
36	employee, faculty, election, brand, career, voter, income, customer, crime, nonprofit	Organizational Behavior and Marketing

37	teacher, classroom, mathematics, leadership, curriculum, instructional, educator, literacy, faculty, instructor	Educational Practices and Leadership
38	thermal, seismic, flame, alloy, combustion, earthquake, jet, deformation, aerosol, shale	Thermal and Seismic Engineering
39	wireless, antenna, robot, bandwidth, converter, routing, cmos, throughput, actuator, overhead	Wireless Communication and Robotics
40	nurse, nursing, healthcare, physician, medication, pain, provider, caregiver, client, diabetes	Healthcare and Nursing Studies
41	wireless, antenna, converter, vibration, mimo, mesh, actuator, beam, bandwidth, transmitter	Wireless Systems and Signal Processing
42	teacher, classroom, curriculum, instructional, leadership, educator, learner, preservice, administrator, pedagogy	Pedagogical Methods and Instruction
43	polymer, membrane, molecule, hydrogel, enzyme, scaffold, acid, nanoparticles, peptide, polymerization	Biomaterials and Polymer Science
44	hiv, breast, diabetes, infection, tumor, obesity, mortality, lung, vaccine, cardiac	Epidemiology and Public Health

45	pain, medication, sleep, infant, nurse, mother, diabetes, nursing, coping, pediatric	Pediatric Healthcare and Pain Management
46	restaurant, corporate, brand, street, marketing, employee, apparel, facebook, citizen, tourism	Corporate Branding and Marketing Strategies
47	moral, judgment, privacy, party, emotion, narrative, auditor, psychological, trust, employee	Ethics and Organizational Psychology
48	wireless, antenna, mimo, bandwidth, converter, scheduling, ghz, throughput, rf, voltage	Advanced Wireless Communication
49	leadership, customer, brand, employee, trust, stakeholder, commitment, manager, faculty, architect	Leadership and Stakeholder Management
50	habitat, cattle, bird, deer, soil, pathogen, virus, infection, nest, mortality	Wildlife and Environmental Studies
51	teacher, classroom, leadership, instructional, educator, curriculum, literacy, learner, mathematics, childrens	Classroom Instruction and Educational Leadership
52	tumor, bone, therapeutic, vivo, cardiac, mri, surgical, surgery, vitro, regeneration	Medical Research and Regenerative Therapy

53	auditory, emotion, personality, speaker, cortex, emotional, perceptual, neuron, judgment, bilingual	Neuroscience and Emotional Studies
54	tax, essay, income, earnings, investor, monetary, volatility, debt, fund, wage	Economic Theory and Taxation
55	corrosion, oxidation, sediment, wastewater, contaminant, ph, adsorption, coating, membrane, acid	Environmental Chemistry and Corrosion
56	soil, habitat, agricultural, crop, wetland, farmer, predator, invasive, nutrient, vegetation	Soil Science and Agriculture
57	processor, cache, scheduling, runtime, query, multicore, workload, bandwidth, scalable, latency	High-Performance Computing Systems
58	essay, pricing, monetary, auction, tax, incentive, income, investor, retailer, portfolio	Financial Economics and Market Analysis
59	agricultural, soil, habitat, farm, crop, farmer, vegetation, grazing, cattle, wildlife	Sustainable Farming and Ecology
60	customer, employee, institutional, manager, brand, citizen, party, corporate, incentive, news	Consumer Behavior and Corporate Governance
61	music, composer, musical, piano, song, genre, opera, artist, choral, repertoire	Classical Music and Composition

62	infection, inflammation, receptor, immune, inflammatory, cardiac, lung, tumor, cytokine, macrophage	Immunology and Inflammatory Diseases
63	anxiety, adolescent, sexual, emotion, emotional, depressive, psychological, adhd, coping, youth	Adolescent Mental Health and Psychology
64	solar, thermal, lithium, voltage, turbine, oxide, photovoltaic, alloy, renewable, electrolyte	Renewable Energy and Photovoltaic Systems
65	adolescent, teacher, youth, mother, african, parental, eating, mathematics, peer, girl	Youth Development and Education
66	adolescent, sexual, anxiety, sleep, psychological, adhd, youth, emotion, coping, mother	Adolescent Behavior and Mental Health
67	sediment, precipitation, lake, soil, watershed, rainfall, ocean, groundwater, storm, basin	Hydrology and Sedimentology
68	faculty, leadership, employee, respondent, questionnaire, demographic, manager, campus, website, news	Academic Leadership and Faculty Studies
69	soil, wetland, watershed, agricultural, vegetation, microbial, nutrient, crop, habitat, irrigation	Agricultural Ecology and Wetlands

70	bacteria, bacterial, pathogen, infection, antibiotic, biofilm, coli, virulence, salmonella, microbial	Microbiology and Pathogen Research
71	receptor, insulin, rat, inhibitor, mutant, infection, vivo, vitro, tumor, immune	Biomedical Research and Therapeutics
72	basin, fault, sandstone, deposit, sediment, facies, tectonic, volcanic, seismic, mantle	Geological Sciences and Tectonics
73	habitat, genus, music, rural, fossil, diet, soil, lake, taxon, bird	Rural Ecology and Biodiversity
74	emission, soil, precipitation, sediment, coastal, watershed, storm, flood, drought, warming	Climate Change and Environmental Impact
75	pain, caregiver, sleep, nurse, ptsd, healthcare, trauma, illness, psychological, knee	Caregiving and Psychological Trauma
76	teacher, career, faculty, leadership, enrollment, classroom, campus, youth, administrator, graduation	Career Development in Education
77	obesity, billion, emission, immigrant, income, hispanic, wage, electricity, overweight, crash	Socioeconomic Studies and Public Health
78	robot, wireless, routing, antenna, gps, packet, radar, stiffness, navigation, fault	Autonomous Systems and Wireless Routing

79	tumor, mutation, breast, virus, receptor, infection, leukemia, inhibitor, metastasis, viral	Oncology and Viral Research
80	graphene, silicon, laser, semiconductor, quantum, fabrication, polymer, transistor, genome, electron	Semiconductor Manufacturing and Nanotechnology
81	alloy, crystal, hydrogen, ion, electron, thermal, oxygen, oxidation, oxide, atom	Material Science and Crystal Engineering
82	teacher, classroom, bullying, disability, youth, adolescent, autism, childrens, peer, asd	Inclusive Education and Youth Challenges
83	seismic, earthquake, shear, corrosion, crack, precipitation, soil, fault, specimen, shale	Seismic Activity and Structural Materials
84	tourism, brand, marketing, rail, customer, corporate, airport, globalization, downtown, stadium	Tourism and Global Marketing
85	crop, agricultural, soybean, soil, corn, farmer, biomass, farm, wheat, fertilizer	Crop Science and Sustainable Farming
86	cow, diet, calf, soil, habitat, deer, cattle, steer, carcass, milk	Livestock Management and Diet Studies
87	diet, pig, cow, dietary, milk, acid, digestibility, corn, calf, beef	Nutritional Science and Animal Husbandry

88	galaxy, stellar, planet, cosmic, disk, gravitational, neutrino, galactic, telescope, supernova	Astrophysics and Stellar Studies
89	emission, renewable, sustainability, solar, electricity, pv, thermal, biomass, turbine, cooling	Sustainability and Renewable Energy
90	electron, crystal, ion, thermal, graphene, hydrogen, atom, alloy, oxidation, solar	Advanced Materials and Energy Applications
91	privacy, website, iot, query, twitter, server, networking, phone, wireless, apps	Cybersecurity and IoT Applications
92	obesity, insulin, diet, intake, dietary, obese, diabetes, glucose, metabolic, rat	Metabolic Disorders and Nutrition
93	insect, mating, bee, fly, bird, genus, spider, bat, reproductive, neuron	Entomology and Reproductive Biology
94	diet, acid, soil, sorghum, biomass, nutrient, fruit, milk, flour, soybean	Agricultural Nutrition and Biomass
95	emotion, mindfulness, emotional, wellbeing, brand, psychological, reward, mood, selfefficacy, anxiety	Mindfulness and Emotional Wellbeing
96	sonata, piano, recital, op, music, musical, concerto, orchestra, violin, composer	Classical Music and Orchestration
97	columbus, mary, marian, negro, bold, geology, mfa, crystal, music, acid	Historical Studies and Geological Research

98	pdf, poem, poetry, homemaking, beef, trust, pavement, extrusion, cont, leadership	Miscellaneous and Leadership Contexts
99	cuzco, aymara, contrastive, grammatical, del, la, una, como, por, verb	Linguistics and Cultural Studies