Educational Video Transcript Analysis with LLMs: Improving Entity Recognition and Qualitative Insights

Wei Wang
University of Tennessee
Knoxville, USA
wwang93@vols.utk.edu

Cody Pritchard
University of Tennessee
Knoxville, USA
cpritc12@vols.utk.edu

Joshua Rosenberg
University of Tennessee
Knoxville, USA
jrosenb8@utk.edu

Digital educational media platforms such as YouTube have become invaluable sources of qualitative data for educational researchers, who increasingly rely on video content to examine classroom interactions, instructional practices, and student engagement at scale. Yet, large-scale qualitative analysis of video content remains hampered by transcription inaccuracies and challenges in extracting structured information. This study introduces an integrated methodological pipeline leveraging Automatic Speech Recognition (ASR), Large Language Models (LLMs) for coreference resolution, and Named Entity Recognition (NER) correction to enhance transcript fidelity and analytical utility. We analyzed 48 episodes from CrashCourse US History, applying multiple ASR systems and LLM-based transcript enhancement to assess large-scale trends in transcription quality, entity extraction, and topic modeling. For evaluation, four episodes were selected for detailed manual annotation, serving as gold-standard benchmarks for validating NER and coreference improvements introduced by the LLM-powered pipeline. Results show that LLM-assisted coreference resolution and NER correction significantly improve the accuracy, recall, and precision of key historical entities, especially for complex event, organization, and law entities. Topic modeling analyses further reveal that LLM-cleaned transcripts yield more coherent and semantically distinct topics, both at the corpus level and in focused case studies, such as the “Reagan Revolution” episode. By comparing traditional ASR pipelines with the proposed LLM-enhanced workflow, we show the value of combining automated language technologies with qualitative research goals. The findings highlight the potential of LLMs as an artificial intelligence tool to advance educational data mining and qualitative inquiry, enabling researchers to increase the reliability of entity recognition in educational videos, facilitate thematic mapping and comparative analyses of teaching practices, classroom interactions, or policy enactment across diverse educational settings.

The code and data are available at https://github.com/wwang93/JEDM-Paper-Pipeline.git

Keywords: YouTube Transcript, Automatic Speech Recognition, Large Language Models, Coreference Resolution, Named Entity Recognition, Qualitative Research

1. INTRODUCTION

K-12 teachers and university professors are increasingly supplementing their curricula with new online digital resources from Teachers Pay Teachers, Share My Lesson, and YouTube (Polikoff, 2019; Fyfield, 2021; Breslyn & Green, 2022; Tosh et al., 2020). YouTube videos are not only used as student-facing resources to show students, but also as pedagogical resources for teachers seeking to build their content knowledge (Adu-Marfo et al., 2024). With over 20 million videos uploaded each day to YouTube (YouTube, n.d.), it is quickly emerging as a vital repository of qualitative data. Educational researchers increasingly recognize the value of online video content, including instructional videos, documentaries, and educational entertainment, for understanding diverse phenomena such as teaching methods, narrative framing, and cultural representation (Tosh et al., 2020; Polikoff, 2019; Miles et al., 2024; Fyfield, 2021; Lange, 2019).

Despite these opportunities, traditional qualitative analysis methods face significant obstacles when dealing with video-based data. Manual transcription—a foundational step for qualitative inquiry—is notoriously labor-intensive, prone to errors, and cost-prohibitive at scale (Eftekhari, 2024). Consequently, qualitative researchers have typically limited their scope to relatively small datasets, constraining the depth, breadth, and generalizability of their findings. Furthermore, once transcribed, the complexity of video narratives, which often feature numerous historical figures, events, and ambiguous references, complicate effective thematic and content analyses (Creswell, 2013).

Recent advances in Natural Language Processing (NLP) and Artificial Intelligence, including Automatic Speech Recognition (ASR), Named Entity Recognition (NER), Coreference Resolution, and Large Language Models (LLMs), have the potential to address these limitations, transforming qualitative researchers' ability to prepare, structure, and analyze large-scale video datasets. ASR enables rapid transcription of extensive video content, NER efficiently identifies and isolates key historical entities and events, and Coreference Resolution accurately associates ambiguous pronouns and indirect references with the correct entities. Additionally, LLMs provide powerful capabilities for refining transcripts, resolving textual ambiguities, and enhancing overall text quality, thereby dramatically improving the fidelity and interpretability of qualitative data.

This study introduces and evaluates an integrated methodological framework combining these advanced NLP tools, designed explicitly to facilitate large-scale qualitative analyses of educational media content. Using a corpus of 48 episodes from the CrashCourse US History video series as an illustrative case, our study presents technical innovations for qualitative data preparation and demonstrates how qualitative researchers can harness these innovations to undertake richer analyses of educational content.

2. PRIOR RESEARCH

2.1. ASR TRANSCRIPTS IN EDUCATIONAL DATA MINING

ASR technologies have increasingly become integral to educational data mining (De Vries et al., 2014), providing researchers with the capability to transcribe and analyze vast quantities of educational video content efficiently. ASR tools such as Whisper significantly reduce the resource-intensive nature of manual transcription, demonstrating their effectiveness through studies on educational video analysis (Rao, 2023). Despite these benefits, ASR-generated transcripts often contain errors that compromise downstream applications, such as sentiment analysis, racial disparities (Olatunji et al., 2023; Koenecke et al., 2020), low accuracy in accent recognition (Hinsvark et al., 2021), and speech emotion recognition (SER).

Addressing these limitations, researchers have utilized multimodal approaches, incorporating techniques such as Deep Canonical Correlation Analysis (DCCA), to enhance the robustness of sentiment classification (Dumpala et al., 2018). Furthermore, SER research has explored frameworks integrating ASR error correction and modality fusion to mitigate the adverse effects of ASR inaccuracies (Li et al., 2024). Additionally, comprehensive corpora, like the Corpus of German Speech (CoGS), illustrate the utility of geolocated ASR transcripts for linguistic and educational analyses, underscoring the expanding role of ASR in educational data mining (Coats, 2023).

2.2. Coreference Resolution as a Key NLP Step

Coreference resolution is a fundamental task in NLP, essential for a range of downstream applications, including entity linking (Kundu et al., 2018), named entity recognition (Dai et al., 2019), question answering (Bhattacharjee et al., 2020), sentiment analysis (Krishna et al., 2017; Mao & Li, 2021), and the development of dialogue systems such as chatbots (Zhu et al., 2018). Coreference resolution clusters entity or event mentions—text spans referring to the same real-world entity or event—within or across documents. This process is essential for information aggregation and supports applications such as contradiction detection, text summarization, and reading comprehension (Ferracane et al., 2016; Khashabi et al., 2018; Welbl et al., 2018).

Coreference resolution contributes to improving the coherence and interpretability of topics generated by accurately clustering references to identical entities across text segments. It clarifies ambiguities inherent in natural language discourse, thus providing clearer narrative structures and enhancing topic coherence (Teng et al., 2023). Incorporating discourse semantics, such as centering theory, into coreference systems further improves referential accuracy, particularly in resolving pronouns and maintaining contextual coherence in long-form texts (Chai & Strube, 2021). Entity-focused topic modeling methods like EntLDA utilize coreference resolution to integrate semantic entity information, resulting in topics that are contextually richer and semantically precise (Allahyari & Kochut, 2016). Consequently, these methodological advancements enable qualitative researchers to conduct more meaningful analyses, providing structured and insightful thematic representations essential for educational inquiry (Harabagiu & Maiorano, 1998; Shi et al., 2010). Traditional topic modeling methods often fail to adequately resolve referential ambiguities and typically produce overly generalized topics with limited interpretative depth, particularly in historical educational narratives.

Effective coreference resolution is essential for accurately associating pronouns and nominal references with their intended entities, a challenge exacerbated by the complexity and density of educational texts (Agarwal et al., 2019). Techniques leveraging fine-grained entity classification, memory networks, and structural information significantly improve coreference resolution, thus enhancing information extraction and supporting deeper text comprehension (Cheri & Bhattacharyya, 2017; Kong & Jian, 2019; Sonawane & Kulkarni, 2015; Wang & Li, 2020).

2.3. Named Entity Recognition

In addition to coreference resolution, named entity recognition (NER) is another important NLP step. NER is considered a sequence labeling task— one in which a system assigns entity class labels to each token within a given sequence, which refers to the identification of different types of entities, including Person, Event, and Location. These entities serve as referential anchors that structure the semantics of texts and guide their interpretation (Ehrmann et al., 2023). NER supports information retrieval by enabling entity-based indexing, which improves the precision of document search and excerpt retrieval (Guo et al., 2009; Lin et al., 2012). Empirical evidence indicates that a large proportion of search queries and content-bearing words involve named entities, highlighting their centrality in text analysis (Gey, 2000).

NER enhances qualitative research through efficient identification and categorization of entities within unstructured text, and facilitates systematic extraction of information about individuals, organizations, locations, and other key entities (Colavizza et al., 2019). This process assists data coding, thematic analysis, and cross-case comparison by providing consistent reference points across large and heterogeneous datasets. NER further contributes to multilingual and cross-cultural studies by standardizing the recognition of proper nouns and references in different languages (Savaram et al., 2024). Research showed the utility of NER encompasses diverse domains such as information extraction, question answering (Mollá et al., 2006), monitoring of media content (Steinberger et al., 2009), sentiment analysis, automated translation, summarization of texts, and clustering of documents (Escoter et al., 2017). In qualitative analysis, NER improves data organization, reduces ambiguity in entity references, and enhances the scalability and reproducibility of research, particularly for extracting and organizing information from unstructured text (Hu et al., 2024; Zhang, 2024).

NER is pivotal for effectively processing educational texts, as it facilitates precise identification and coherent linkage of entities within the discourse. Recent NER methodologies employing advanced architectures such as BERT-BiLSTM-CRF have notably improved the accuracy of entity extraction, overcoming the limitations of traditional rule-based systems (Cheng, 2023; Wei & Wen, 2021). The development of domain-specific datasets, such as EduNER, has further enhanced model performance by providing contextually rich training data tailored specifically for educational contexts (Li et al., 2023).

2.4. LLMs for Qualitative Research Assistance

LLMs have created significant possibilities for qualitative research, supporting more effective processing, analysis, and interpretation of complex, unstructured textual data (Tai et al., 2024; Hayes, 2025). LLMs trained on extensive corpora demonstrate remarkable capabilities in pattern recognition, coreference resolution (Le & Ritter, 2024), thematic extraction, and sentiment analysis, surpassing traditional manual coding methods in both speed and depth (Hayes, 2025). Recent applications highlight LLMs' effectiveness in preprocessing historical and educational texts, accurately correcting transcription errors, and extracting targeted information (Schwitter, 2025). Despite their transformative potential, the reliance on sophisticated prompt engineering and the risk of model-generated inaccuracies underscores the necessity for rigorous validation processes and ethical guidelines (Rask & Shimizu, 2024; Barros et al., 2024). Moreover, while LLMs expedite literature reviews and scientific writing, challenges persist regarding dataset biases and ethical considerations, requiring continued vigilance and methodological refinement (Boyko et al., 2023). Notably, existing research seldom explicitly integrates LLM-driven processes with established qualitative methods such as codebook development (Barany et al., 2024), thematic analysis, narrative inquiry, and discourse analysis, leaving the interpretative and contextual relevance of automated findings largely unaddressed.

2.5. Limitations of Existing Approaches and Research Gap

Despite the promising advancements outlined above, significant limitations persist in current qualitative educational video research. The scalability of traditional educational video transcript qualitative analyses remains constrained due to reliance on manual coding and transcription methods (Parameswaran et al., 2020). Furthermore, the fragmented approach in utilizing ASR, NER, coreference resolution, and LLM-based processing often leads to inaccuracies, unresolved coreferences, inconsistent entity recognition, and overly broad thematic outcomes.

Instructional video content on platforms like YouTube, including but not limited to history education—presents researchers with complex analytical challenges, such as rich narrative structures, diverse entity references, and extensive use of pronouns and implicit context. While historical educational videos exemplify these challenges due to their dense storytelling and frequent mention of people, events, and organizations, similar issues are pervasive across other forms of educational and instructional video, such as STEM lectures, language tutorials, and policy explainers. We selected CrashCourse US History videos as a case study precisely because they are widely used in classrooms and feature a high density of narrative elements and educational entities, making them an ideal testbed for developing and validating our integrated methodological framework. However, our approach—synergistically combining ASR transcription, LLM-driven transcript cleaning, entity recognition, and coreference resolution—is designed to be broadly applicable to a wide range of instructional video data beyond this case.

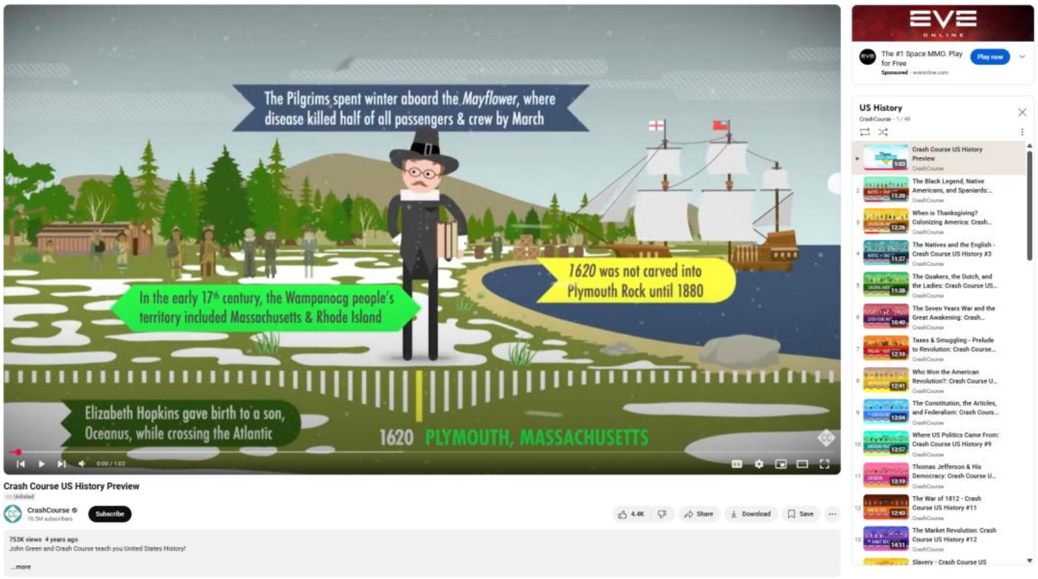
To address these critical gaps, this study presents an integrated methodological framework that combines ASR transcription, LLM-driven transcript cleaning, precise entity recognition, and coreference resolution to significantly enhance the coherence, accuracy, and interpretability of qualitative analyses in educational research contexts.

Specifically, we systematically investigate the following research questions:

RQ1: How accurate are ASR-generated transcripts of educational media compared to a human-created "gold standard"?
RQ2: What are the effects of employing LLM-based cleaning processes on ASR-generated transcripts, particularly regarding named entity recognition?
RQ3: How does LLM-enhanced NER and coreference resolution affect the thematic coherence and qualitative interpretability of narratives compared to those from unprocessed ASR transcripts?

3. DATA

This study employs the CrashCourse US History video series as its principal dataset, capitalizing on its pedagogical relevance and content richness. CrashCourse, an educational YouTube channel with over 16 million subscribers, is recognized for producing meticulously researched and visually engaging educational videos, as shown in Figure 1. The US History series has become a staple resource in both K–12 and postsecondary classrooms as well as among informal learners worldwide. Its episodes present a structured, chronological exploration of key events, figures, and themes in American history, integrating multimedia elements and narrative techniques to enhance comprehension and engagement. 48 episodes were selected to represent the breadth of major historical periods and events, ranging from the colonial era and the American Revolution to the Civil War, Reconstruction, and the Civil Rights Movement. This corpus reflects the chronological scope and the thematic complexity of US history.



The Pilgrims spent winter aboard the *Mayflower*, where disease killed half of all passengers & crew by March
1620 was not carved into Plymouth Rock until 1880
In the early 17th century, the Wampanoag people's territory included Massachusetts & Rhode Island
Elizabeth Hopkins gave birth to a son, Oceanus, while crossing the Atlantic
1620 PLYMOUTH, MASSACHUSETTS
0:00 / 1:02
EVE ONLINE
The #1 Space MMO. Play for Free
Play now
Sponsored · eveonline.com
Unlisted
CrashCourse
16.5M subscribers
Subscribe
4.4K
Share
Download
Save

Crash Course US History Preview

753K views 4 years ago
John Green and Crash Course teach you United States History!
...more

US History
CrashCourse - 1 / 49

1. Crash Course US History Preview (1:03) CrashCourse
2. The Black Legend, Native Americans, and Spaniards:... (11:20) CrashCourse
3. When is Thanksgiving? Colonizing America: Crash... (12:26) CrashCourse
4. The Natives and the English - Crash Course US History #3 (11:27) CrashCourse
5. The Quakers, the Dutch, and the Ladies: Crash Course US... (11:38) CrashCourse
6. The Seven Years War and the Great Awakening: Crash... (10:40) CrashCourse
7. Taxes & Smuggling - Prelude to Revolution: Crash Course... (12:19) CrashCourse
8. Who Won the American Revolution?: Crash Course U... (12:41) CrashCourse
9. The Constitution, the Articles, and Federalism: Crash Cours... (13:04) CrashCourse
10. Where US Politics Came From: Crash Course US History #9 (13:57) CrashCourse
11. Thomas Jefferson & His Democracy: Crash Course U... (13:19) CrashCourse
12. The War of 1812 - Crash Course US History #11 (12:43) CrashCourse
13. The Market Revolution: Crash Course US History #12 (14:31) CrashCourse
Slavery - Crash Course US

Figure 1. Example YouTube CrashCourse US History Video.

The data collection process involved a multi-step procedure. First, the selected videos were downloaded from YouTube using yt-dlp, a widely adopted open-source utility capable of extracting both video content and associated subtitle files. For each episode, the English auto-generated subtitles—produced via YouTube's embedded ASR system—were retrieved and archived as the primary text source. These ASR-generated transcripts, while immediately accessible and comprehensive, are known to exhibit variable levels of accuracy, especially regarding specialized terminology, proper nouns, and context-dependent phrases prevalent in historical discourse.

To facilitate robust evaluation and benchmarking, a subset of the corpus was subjected to manual transcription and entity annotation. For this "gold standard" subset, transcripts were produced or corrected by a researcher with subject matter expertise, who systematically cross-checked the auto-generated subtitles against the original video content. Special attention was given to the accurate transcription and identification of named entities, including historical figures, place names, dates, organizations, and legislation. Each identified entity was categorized and recorded following the established NER schema. To ensure the reliability of entity labeling, all annotations were created using the gold standard transcripts: we used the Google Gemini 2.5 Pro LLM to extract an initial list, after which a single researcher checked each named entity against both the text and historical sources. This process resulted in a manually verified NER gold standard: for the selected sample, providing a rigorous benchmark for evaluating the accuracy and performance of automated NER extraction throughout the study. The finalized dataset consists of the original ASR-generated transcripts for all 48 CrashCourse US History videos, along with manually created "gold-standard" transcript samples for four selected episodes. For these four videos, the NER gold standard was also established, including detailed classification and annotation of all named entities.

4. Method

This study employed a multi-stage methodological pipeline (see abstract for codes and data link), integrating natural language processing tools and a large language model (GPT-4o), to systematically address the three research questions. Each phase of the pipeline was designed to correspond directly to one of the research questions, ensuring clarity and methodological alignment throughout.

4.1. Evaluating ASR Transcript Accuracy (RQ1)

To investigate the first research question—how accurate ASR-generated transcripts are when compared to a manually created "gold standard"—we utilized the JiWer evaluation toolkit to quantify transcription fidelity. Multiple state-of-the-art ASR solutions were employed, including yt-dlp, Whisper, TurboScribe, Otter, and Vosk, as detailed in Table 1. These systems were selected for their prominence and representativeness in educational media transcription. For the primary demonstration and validation, we selected a representative video sample, for which ASR-generated transcripts were systematically compared against a meticulously transcribed human gold standard. Evaluation metrics, including word error rate (WER), match error rate (MER), word information lost (WIL), and character error rate (CER), were computed using the Python package JiWER (JiWER, 2025). For readers interested in comprehensive, corpus-wide ASR benchmarking, we provide access to the complete set of video subtitles on our Google Drive repository, along with scripts for reproducing all analyses

Table 1. Overview of Selected Automatic Speech Recognition Tools.

ASR	Developer	Type	Interface	Features
yt-dlp	Open-Source Community	Video/Audio Downloader + ASR	CLI (Python-based)	Integrates with various ASR backends; used for extracting and transcribing YouTube & online media audio.
Whisper	OpenAI	Deep Learning (Transformer) ASR	CLI, API, Python lib	Multilingual support; open-source; supports long-form audio transcription.
TurboScribe	TurboScribe, Inc.	Proprietary ASR SaaS	Web, Desktop App	User-friendly interface; automated diarization; supports various formats.
Otter	Otter.ai	Proprietary ASR SaaS	Web, Mobile App	Real-time transcription; speaker identification; searchable, shareable transcripts.
Vosk	Open-Source Community	Offline Deep Learning ASR	Python, Java, Node.js	Works offline; supports many languages; lightweight for embedded systems and real-time apps.

4.2. LLM-Based Coreference Resolution and NER Enhancement for RQ2

The second research question focused on the effects of employing LLM-based cleaning processes—specifically coreference resolution and named entity recognition (NER)—on improving transcript accuracy and the extraction of historical entities. To this end, we applied the state-of-the-art GPT-4o model to all 48 video transcripts in our dataset. The LLM was prompted (see Appendix A.1 for full prompt details) to perform coreference resolution, thereby clarifying ambiguous pronouns and nominal references across the transcripts. Following this step, the model was further prompted (see Appendix A.2 for prompt details) to conduct comprehensive NER, identifying and correcting named entities crucial to historical narratives, including PERSON, ORGANIZATION, GPE, LOC, NORP, EVENT, and LAW categories, as shown in Table 2.

Table 2. Named Entity Recognition Types.

Entity Type	Description	Example Entities
PERSON	Names of historical figures, individuals, or participants	Abraham Lincoln
ORGANIZATION	Groups, institutions, or companies, etc.	United Nations, IBM
GPE	Geopolitical entities (countries, cities, states)	United States, Paris, Tennessee
LOC	Non-GPE locations, mountains, landmarks, etc.	Mount Everest, Mississippi River
NORP	Nationalities, religious or political groups	American, Protestant, Democrat
EVENT	Named historical events, wars, movements	World War II, Civil Rights
LAW	Named laws or legal documents	Constitution, Civil Rights Act

For illustration, we present comparative transcript example excerpts, displaying the ASR baseline, the LLM coreference-resolved version, and the final LLM-enhanced NER version.

> ASR Baseline:
> “ *He signed the bill, but he did not anticipate how Congress and the Court would respond.”*
>
> LLM Coreference-Resolved:
> “ *President Reagan signed the bill, but President Reagan did not anticipate how Congress and the Supreme Court would respond.”*
>
> LLM NER-Enhanced:
> “ *President Ronald Reagan signed the Economic Bill of Rights, but President Ronald Reagan did not anticipate how the United States Congress and the Supreme Court would respond.”*

To evaluate NER accuracy, we randomly selected four videos from the corpus and constructed a manually validated gold standard of entity annotations. Accuracy, precision, recall, and F1-score were computed for each of the three transcript versions across the seven core NER categories, thereby enabling direct performance comparison and quantifying the improvements brought by LLM-based processing.

4.3. Thematic Modeling and Qualitative Interpretability for RQ3

The third research question explored how LLM-enhanced NER and coreference resolution affect the thematic coherence and qualitative interpretability of narratives relative to unprocessed ASR transcripts. We applied Latent Dirichlet Allocation (LDA) topic modeling to each of the three transcript versions (raw ASR, coreference-resolved, and NER-enhanced) for all 48 videos. Using pyLDAvis for visualization, we assessed differences in topic structure by examining intertopic distances and the semantic separation between clusters, with the number of topics (K) set to 5 for comparability. This approach allowed us to systematically compare the consistency, clarity, and distribution of themes across preprocessed and post-processed corpora.

Additionally, we conducted a focused video transcript case on the "Reagan Revolution" episode, matching named entities and their types from each transcript version to a human-validated list of core narrative entities. This enabled a detailed, entity-level assessment of how LLM-based processing influences the interpretability and narrative fidelity of qualitative analyses. Through these methods, we demonstrate not only the quantitative improvements in transcript accuracy and entity extraction, but also the tangible benefits for qualitative inquiry in educational research contexts.

5. RESULTS

5.1. RQ1: ASR TRANSCRIPTS ACCURACY VALIDATION

To address the first research question, we systematically evaluated the accuracy of multiple ASR-generated transcripts against manually curated gold-standard transcripts, employing established metrics including WER, MER, WIL, and CER. Our results reveal considerable variation in transcription quality across the different ASR systems. Among all systems tested, TurboScribe achieved the highest transcription fidelity, recording a WER of 2.43%, MER of 2.42%, WIL of 3.76%, and CER of 0.99%. Both Whisper and yt-dlp exhibited comparable, though marginally higher, error rates (Whisper: WER = 3.42%, MER = 3.40%, WIL = 5.24%, CER = 1.40%; yt-dlp: WER = 3.57%, MER = 3.52%, WIL = 4.89%, CER = 2.07%), confirming their suitability for large-scale educational video transcription tasks.

In contrast, Otter yielded moderately elevated error rates (WER = 4.16%, MER = 4.08%, WIL = 6.17%, CER = 1.22%), while Vosk performed noticeably worse, with a WER of 15.32%, MER of 14.69%, WIL of 22.84%, and CER of 8.86%. Notably, the overall pattern suggests that proprietary and cloud-based ASR solutions such as TurboScribe and Whisper tend to outperform open-source or offline alternatives, especially in accurately rendering educational terminology and specialized content. This highlights the importance of ASR system selection in educational media research, as even relatively small differences in error rates can have substantial downstream impacts on entity extraction and qualitative analysis. These findings establish a robust baseline for subsequent LLM-enhanced transcript processing and underscore the persistent challenges inherent in automated transcription of educational video content.

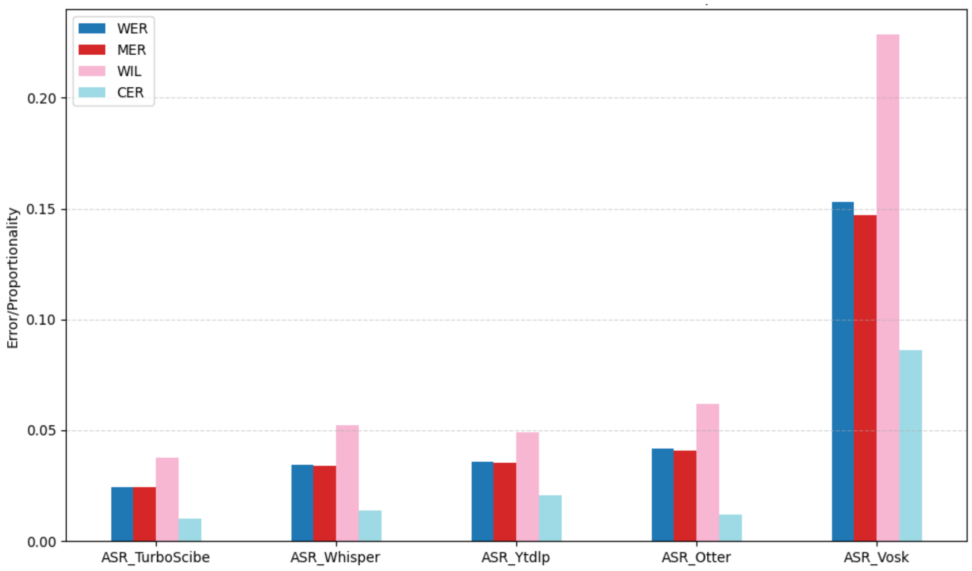


Figure 2. JiWER Validation Results across Selected ASR Systems.

5.2. RQ2: LLM-Assisted Transcript NER Validation

To address the second research question, we systematically assessed the impact of LLM-based processing on NER performance across three pipeline stages: (1) initial ASR-generated transcripts, (2) transcripts processed with LLM-assisted coreference resolution, and (3) transcripts subjected to comprehensive LLM-based NER correction. Table 3 summarizes the comparative results for each core NER category, highlighting both the quantitative gains and category-specific nuances.

A clear pattern of progressive improvement emerges across most entity types. For EVENT entities, which often present the greatest challenge due to their context-dependent and narrative nature, accuracy increased from 0.06 in the ASR baseline to 0.08 with coreference resolution, and then more than doubled to 0.16 with full LLM-based NER; correspondingly, the F1 score rose from 0.11 (ASR) to 0.15 (coreference) and ultimately 0.27 (NER). LAW and ORG categories also showed substantial advances: for LAW, accuracy improved dramatically from 0.08 (ASR) and 0.08 (coreference) to 0.57 (NER), with the F1 score climbing from 0.15 to 0.73. ORGANIZATION entities saw accuracy grow from 0.12 (ASR) to 0.16 (coreference) and 0.35 (NER), with a corresponding F1 increase from 0.21 to 0.52.

For PERSON entities, which were among the most reliably recognized, accuracy and F1 both improved, with LLM-based NER achieving an accuracy of 0.63 and an F1 of 0.78, compared to 0.57 and 0.73, respectively, in the ASR baseline. The GPE (geopolitical entities) category, which already benefited from relatively strong ASR performance, still realized gains: accuracy rose from 0.59 (ASR) to 0.70 (NER), and the F1 score from 0.74 to 0.82, suggesting that LLM post-processing can further enhance even robust baseline performance.

Not all categories benefited equally. LOC (non-political geographic locations) and NORP (nationalities, religious, or political groups) displayed more modest or mixed improvement. LOC accuracy and F1 showed little change, while NORP improved slightly with coreference resolution (accuracy 0.43 to 0.45, F1 0.60 to 0.62), but then declined with full NER correction (accuracy 0.15, F1 0.26), likely due to over-correction or model-specific challenges in distinguishing nuanced group references. Across almost all categories, the introduction of LLM-based coreference resolution led to moderate gains in precision, recall, and F1, while the integration of comprehensive LLM-based NER correction yielded the most pronounced improvements, particularly for those entity types most critical to educational and historical research, such as EVENT, LAW, PERSON, and ORGANIZATION.

Table 3. Selected Sample Video Transcripts NER Validation.

NER	**ASR**				**LLM Coreference**				**LLM NER**			
	Acc	Precision	Recall	F1	Acc	Precision	Recall	F1	Acc	Precision	Recall	F1
EVENT	0.06	0.20	0.07	0.11	0.08	0.25	0.11	0.15	0.16	0.50	0.19	0.27
GPE	0.59	0.74	0.75	0.74	0.54	0.66	0.75	0.70	0.70	1.00	0.70	0.82
LAW	0.08	0.22	0.11	0.15	0.08	0.22	0.11	0.15	0.57	0.80	0.67	0.73
LOC	0.22	0.33	0.40	0.36	0.22	0.33	0.40	0.36	0.19	0.33	0.30	0.32
NORP	0.43	0.74	0.50	0.60	0.45	0.76	0.53	0.62	0.15	0.77	0.16	0.26
ORG	0.12	0.17	0.28	0.21	0.16	0.21	0.39	0.27	0.35	0.70	0.41	0.52
PERSON	0.57	0.72	0.73	0.73	0.52	0.64	0.73	0.68	0.63	0.93	0.67	0.78

Our results demonstrate that the LLM-augmented pipeline substantially enhances the reliability and completeness of entity extraction from complex educational transcripts. This is especially significant in the context of large-scale qualitative and thematic analyses, where accurate tracking and interpretation of key historical actors, events, and organizations are foundational for robust educational research.

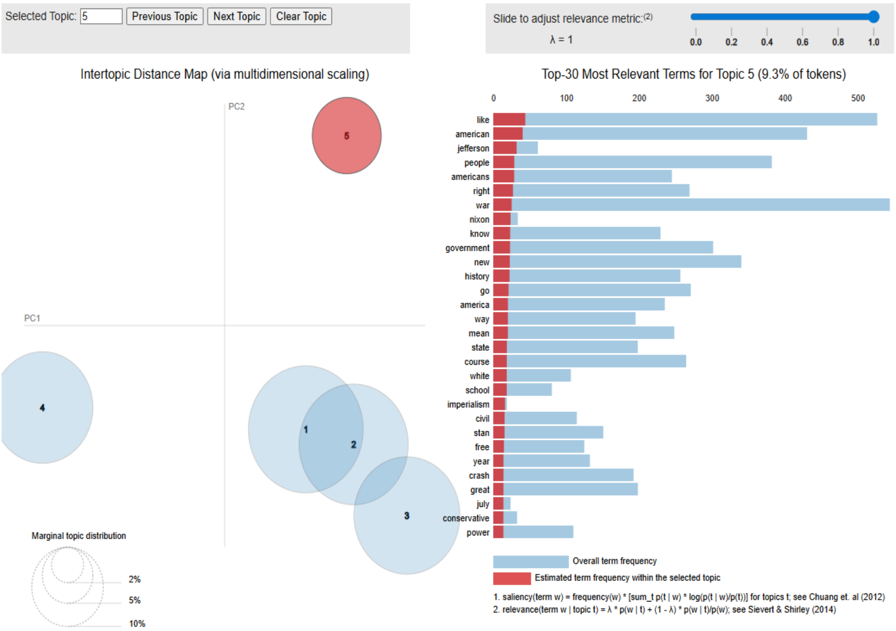
5.3. RQ3: Thematic Coherence and Interpretive Value of Topic Modeling within ASR versus LLM-Cleaned Transcripts

To address the third research question, we applied Latent Dirichlet Allocation (LDA) topic modeling to all 48 episodes' transcripts at three key processing stages: (1) the original ASR-generated transcripts, (2) the LLM-assisted coreference-resolved transcripts, and (3) the final LLM-based NER-enhanced transcripts. Figure 3, Appendix Figures B1 and B2 visualize the intertopic distance maps produced by pyLDAvis for each condition (For anyone wishing to reproduce or further explore the interactive topic modeling results, the full codebase and associated data are openly available at: https://github.com/wwang93/JEDM-Paper-Pipeline.git).

The results reveal a difference in thematic structure and coherence across the three pipelines. In the ASR baseline condition, topics are highly overlapped, as indicated by the significant clustering and intersection of topic bubbles in the left panel. This pattern reflects the semantic diffusion and lack of clear thematic boundaries typical of raw, error-prone transcripts—likely a result of unresolved pronouns and ambiguous references that obscure narrative structure.

LLM-based coreference resolution stage yields the clearest thematic separation: topics are well separated in semantic space, with almost no overlap among topic clusters. This suggests that accurately resolving referential ambiguity plays a critical role in enabling unsupervised models to disentangle underlying themes—likely because entities and events are more consistently and explicitly tracked throughout the transcript. The LLM-based NER-enhanced condition shows partial improvement relative to the ASR baseline, with reduced topic overlaps and somewhat clearer boundaries, but does not achieve the same level of separation as the coreference-only stage. This outcome suggests that, while named entity normalization helps clarify key concepts and entities, it may also introduce overcorrection or artifact, particularly if entities are forced where context is subtle or ambiguous.

Taken together, these findings indicate that LLM-powered preprocessing can enhance the clarity and interpretability of thematic structure in educational video corpora, with coreference resolution contributing most substantially to topic coherence and boundary definition. The results also highlight a nuanced trade-off: while both coreference resolution and NER improve unsupervised analysis, excessive or aggressive entity normalization may in some cases reduce thematic distinctiveness. For researchers, this underscores the importance of selecting and tuning LLM-based preprocessing steps to maximize downstream interpretive value in educational qualitative research.



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 3. ASR Transcript PyDavis Intertopic Distance Map.

To complement the global comparison, we conducted a focused, entity-level analysis using the episode "The Reagan Revolution[1]," which provides an in-depth overview of the pivotal changes that occurred in American society, politics, and international relations during the presidency of Ronald Reagan. The episode examines the ideological, economic, and cultural shifts initiated under Reagan's leadership, including tax reforms, conservative movements, landmark legislation, Cold War dynamics, and the redefinition of U.S. government and global engagement. Through discussions of key individuals, organizations, laws, and historical events, the video explores how the Reagan era reshaped both domestic policy and America's position on the world stage.

To evaluate NER systems on this topic, we constructed a gold standard list of core entities that best represent the main narrative and historical context of the Reagan Revolution period. These entities—encompassing people, organizations, locations, geopolitical regions, events, and landmark legislation—were selected and validated by humans. Table 4 presents the certified NER gold standard for this US history video.

[1] CrashCourse. (2013, August 14). *The Reagan Revolution: Crash Course US History #43* [Video]. YouTube. https://www.youtube.com/watch?v=2h4DkpFP_aw

Table 4. Human Reagan Revolution Narrative NER Gold Standard.

NER Type	Entities
PERSON	Ronald Reagan, Jimmy Carter, George HW Bush, Mikhail Gorbachev, Nancy Reagan, John Poindexter, Oliver North
ORG	Congress, Moral Majority, Supreme Court, NATO, Reagan administration, Sandinista government
GPE	America, US, Soviet Union, Illinois, New York, Western Europe, France, Lebanon, Nicaragua
LOC	Iron Curtain
NORP	American, African Americans, conservatives, religious conservatives, economic conservatives, Cold War hawks, Christian right, anti-government crusaders, Democratic, Soviet, Iranian, Middle Eastern
EVENT	Reagan Revolution, Cold War, Iran-Contra Scandal, New Deal, Great Society, Korean, Vietnam War, 1960s, 1970s, 1980s, mid-1990s, atomic age, FREEZE movement
LAW	Economic Bill of Rights, Tax Reform Act, Anti-ballistic Missile Treaty

We systematically evaluated the output of three NER pipelines: (1) the baseline ASR transcript, (2) LLMs-assisted coreference resolution, and (3) LLMs-based NER correction. Table 5 summarizes the entity-level results for each NER type across these approaches.

The ASR-generated transcript demonstrated notable limitations. For EVENT and LOC entities, the model failed to identify any relevant mentions, with accuracy, precision, recall, and F1 all scoring zero. PERSON and GPE categories achieved modest results, with accuracies of 0.24 and 0.40 and F1 scores of 0.38 and 0.57, respectively. LAW and ORG entities exhibited lower performance (e.g., ORG F1 = 0.17), and overall recall remained limited, particularly for categories central to the Reagan narrative.

The LLM-assisted coreference resolution pipeline yielded moderate improvements for several entity types. For example, PERSON recall increased substantially to 0.86, and NORP (nationalities, religious, and political groups) F1 rose to 0.44. EVENT and LOC categories, however, still showed limited gains, reflecting persistent ambiguity or insufficient disambiguation solely through coreference resolution.

The full LLMs-based NER correction stage produced marked improvements across nearly all categories. Accuracy and F1 scores for EVENT, GPE, LAW, LOC, and PERSON all increased substantially: EVENT F1 improved from 0.00 (ASR) and 0.12 (coreference) to 0.89 (LLMs NER); GPE from 0.57 to 0.95; LAW from 0.40 to 1.00; LOC from 0.00 to 1.00; and PERSON from 0.38 to 0.54. Similarly, the ORG category's F1 increased from 0.17 (ASR) to 0.58 (LLMs NER). Notably, every single entity type reached or approached perfect recall and precision in the LLMs NER stage, demonstrating the transformative effect of the advanced LLM pipeline on entity extraction fidelity.

This single-case evaluation demonstrates that, while automatic ASR systems alone fail to capture the diversity of historically salient entities, the application of LLM-based coreference resolution and NER correction can recover nearly all key figures, organizations, geopolitical regions, and events central to the episode’s narrative. These improvements are especially important for downstream qualitative analysis, as accurate and comprehensive entity extraction directly supports interpretive coding, thematic mapping, and contextualized content analysis in educational research.

Table 5. Regan Revolution Narrative NER Validation.

NER	ASR				LLM Coreference				LLM NER			
	Acc	Precision	Recall	F1	Acc	Precision	Recall	F1	Acc	Precision	Recall	F1
EVENT	0	0	0	0	0.06	0.25	0.08	0.12	0.8	0.86	0.92	0.89
GPE	0.4	0.5	0.67	0.57	0.33	0.45	0.56	0.5	0.9	0.9	1	0.95
LAW	0.25	0.5	0.33	0.4	0.25	0.5	0.33	0.4	1	1	1	1
LOC	0	0	0	0	0	0	0	0	1	1	1	1
NORP	0.24	0.36	0.42	0.38	0.29	0.4	0.5	0.44	0.6	0.6	1	0.75
ORG	0.09	0.12	0.29	0.17	0.11	0.13	0.43	0.2	0.41	0.41	1	0.58
PERSON	0.24	0.26	0.71	0.38	0.29	0.3	0.86	0.44	0.37	0.37	1	0.54

6. Discussion

As AI continues to reshape educational research methods, this study assessed the effect of integrating advanced AI-driven techniques—specifically, LLM-based coreference resolution and NER correction—within an ASR transcript pipeline for large-scale qualitative analysis of educational video content. The results indicate that this approach yields improvements in both named entity recognition and thematic modeling, which are foundational to qualitative inquiry in educational research.

6.1. Advancements in Entity Recognition and Thematic Analysis

Consistent with prior studies (Ehrmann et al., 2023; Rao, 2023), which have highlighted the challenges of entity extraction from historical documents, our study provides empirical evidence that LLM-assisted preprocessing improves the accuracy for key historical entities. Notably, these ganins are most apparent in categories such as PERSON, GPE, EVENT, LAW, and ORG, which are repeatedly underscored in the literature as vital for reconstructing historical narratives (Allahyari & Kochut, 2016; Dai et al., 2019). For instance, while baseline ASR transcripts exhibited limited ability to capture complex or context-dependent entities—especially in the case of events and legal documents—LLM-based correction consistently enabled the recovery and accurate clustering of such information, as shown in both corpus-level and single-episode ("Reagan Revolution") analyses. This resulted in a more complete extraction of relevant entities, enhanced consistency, and reduced ambiguity across transcript versions.

Topic modeling analyses reinforce these quantitative improvements, as LLM-cleaned transcripts yielded topic clusters that were more coherent, semantically distinct, and aligned with substantive historical themes. Intertopic distance mapping via pyLDAvis revealed reduced topic overlap and clearer thematic boundaries in the LLM-processed data, facilitating more robust qualitative interpretation, addressing the lack of downstream qualitative validation noted by Nelson (2020). In practical terms, this allows researchers to identify, trace, and interpret key actors, events, and policy shifts with greater confidence and granularity.

6.2. Implications for Educational Video Qualitative Research Using LLMs

Recent advances in LLMs have generated excitement for the potential in qualitative research (Hayes, 2025; Barany et al., 2024), yet prior studies often lacked a systematic integration of LLMs within pipelines for educational video analysis. Our work directly responds to this gap by automating the resolution of coreference and refining the extraction of named entities. LLMs allow for the analysis of much larger and more diverse corpora than would be feasible with manual transcription and hand-coding alone. This is particularly valuable in the context of contemporary educational research, where the volume and variety of digital video content continue to expand (YouTube, n.d).

Moreover, the increased fidelity and interpretability of LLM-enhanced transcripts can open new avenues for qualitative approaches such as thematic analysis, discourse analysis, and content mapping (Schwitter, 2025; Than et al., 2025). Researchers can more efficiently identify patterns, compare cases, and construct grounded explanations from large-scale data. The improved accuracy of entity-level extraction also facilitates mixed-methods designs, enabling more meaningful integration of quantitative and qualitative findings.

6.3. Reflection on Automation and Human Judgment

While LLMs provide substantial analytical advantages, their application in qualitative research should be contextualized within the broader interpretive process (Than et al., 2025). The automation of entity extraction and topic modeling does not replace the need for human judgment in interpreting meaning, nuance, and context. Rather, these tools can be seen as augmenting and accelerating the analytic process, supporting human researchers in focusing on higher-level interpretation, theory-building, and contextualization (Kubsch et al., 2023; Rosenberg & Krist, 2021; Nelson, 2020).

At the same time, researchers must remain attentive to the potential for errors and biases introduced by automated systems. For example, LLMs may occasionally hallucinate entities, misattribute references, or propagate systematic errors present in training data (Huang et al., 2025; Lin et al., 2024). Thus, best practice entails combining LLM-augmented analysis with validation steps such as manual audit, triangulation, and sensitivity analysis, especially for high-stakes or sensitive research questions.

6.4. Methodological Contribution and Future Directions

This study contributes an empirically validated, scalable workflow for educational video analysis that bridges advances in NLP with qualitative traditions. It demonstrates how LLM-based processing can move the field beyond the limitations of manual transcription and rule-based entity recognition (Jehangir et al., 2023), making possible more rigorous, nuanced, and scalable qualitative analysis of large and complex digital corpora, for example, adapt this LLM-augment pipeline to STEM education videos, where domain-specific Jargon and formulae pose unique challenges for ASR and NER modules (Fyfield, 2021; Coats, 2023). Similarly, applying this approach to multilingual educational content, such as bilingual history classes or international YouTube channels, would test the generalizability of LLM-based coreference and entity resolution, given the increased ambiguity in cross-lingual contexts (Savaram et al., 2024).

Another direction would be to examine teacher-generated media or informal learning resources (Lange, 2019), where content is less scripted and linguistic structure is highly variable, which may strain even advanced LLMs. As automated LLM-based analysis becomes more prevalent, future studies should investigate how hybrid human-AI coding strategies can help resolve interpretive ambiguities, particularly in sensitive domains such as civics or ethics education. Looking forward, further research should explore the application of this pipeline to other content domains, examine its adaptability across languages and contexts, and deepen the integration of automated and human-centered qualitative approaches. Such efforts will help ensure that methodological advances remain robust and relevant across the rapidly evolving landscape of educational media.

7. LIMITATIONS

First, the evaluation was conducted primarily on the CrashCourse US History series. While this corpus is representative of high-quality, content-rich educational media, results may not generalize to other subjects, video genres, or platforms with different linguistic styles, audio quality, or levels of production. Second, gold standard annotations are based on manual review and expert validation, which, despite rigorous procedures, may still reflect human subjectivity or unrecognized bias. Third, the LLMs employed—while state-of-the-art—are sensitive to prompt design and may occasionally produce errors such as hallucinated entities, false positives, or failures in complex or ambiguous contexts. Fourth, the study focused on English-language transcripts; extending this approach to other languages or multilingual content could pose additional challenges related to model coverage and cultural specificity. Finally, although the integrated workflow improves efficiency, fully automated analysis cannot replace the depth and contextual sensitivity of manual qualitative inquiry, particularly for nuanced interpretation or the analysis of implicit meaning.

8. CONCLUSION

This study presents an integrated pipeline that combines ASR, LLM-based coreference resolution, and targeted NER correction to enhance the quality of qualitative analysis for large-scale educational video data. Empirical results demonstrate that the proposed workflow substantially improves entity recognition accuracy and thematic coherence in both corpus-wide and single-episode analyses. By addressing the limitations of traditional manual transcription and rule-based extraction, the LLM-augmented pipeline can support more scalable and rigorous analyses of large-scale educational audio and video content.

DECLARATION OF GENERATIVE AI SOFTWARE TOOLS IN THE WRITING PROCESS

During the preparation of this work, the author(s) used OpenAI's ChatGPT (GPT-4.1) to optimize the language and enhance the clarity of academic writing in sections including Methods, Results, and Discussion. After using this tool, the author(s) thoroughly reviewed and edited the content as needed and took full responsibility for the content of the publication.

REFERENCES

ADU-MARFO, A. O., KWAPONG, O. A. T. F., OHENEBA-SAKYI, Y., AND MILLER-YOUNG, J. 2024. Understanding teachers' usage of YouTube as a pedagogical tool: A qualitative case study of basic school teachers in Ghana. *E-Learning and Digital Media*. Advance online publication.

ALLAHYARI, M., AND KOCHUT, K. J. 2016. Discovering Coherent Topics with Entity Topic Models. *Web Intelligence*, 26–33.

AGARWAL, O., SUBRAMANIAN, S., NENKOVA, A., AND ROTH, D. 2019. Evaluation of named entity coreference. In *Proceedings of the Second Workshop on Computational Models of Reference, Anaphora and Coreference*, 1–7.

BARROS, C., BORGES AZEVEDO, B., GRACIANO NETO, V. V., KASSAB, M., KALINOWSKI, M., DANTAS DO NASCIMENTO, H. A., AND BANDEIRA, M. C. G. S. P. 2024. Large Language Model for Qualitative Research – A Systematic Mapping Study.

BARANY, A., OLNEY, A. M., CHOUNTA, I. A., LIU, Z., SANTOS, O. C., AND BITTENCOURT, I. I. 2024. ChatGPT for education research: Exploring the potential of large language models for qualitative codebook development. In A. M. OLNEY, I. A. CHOUNTA, Z. LIU, O. C. SANTOS, AND I. I. BITTENCOURT, Eds., *Artificial Intelligence in Education. AIED 2024. Lecture Notes in Computer Science* 14830, Springer, Cham.

BOYKO, J., COHEN, J., FOX, N., VEIGA, M. H., LI, J. I.-H., LIU, J., MODENESI, B., RAUCH, A. H., REID, K. N., TRIBEDI, S., VISHERATINA, A., AND XIE, X. 2023. An Interdisciplinary Outlook on Large Language Models for Scientific Research. *arXiv.Org*, abs/2311.04929.

BRESLYN, W., AND GREEN, A. E. 2022. Learning science with YouTube videos and the impacts of Covid-19. *Disciplinary and Interdisciplinary Science Education Research* 4, 13, 1–20.

BHATTACHARJEE, S., HAQUE, R., DE BUY WENNIGER, G. M., AND WAY, A. 2020. Investigating query expansion and coreference resolution in question answering on BERT. In *International Conference on Applications of Natural Language to Information Systems*, 47–59. Springer.

CRESWELL, J. W. 2013. Qualitative inquiry and research design: Choosing among five approaches (3rd ed.). *Sage*.

CECI, L. 2024. Hours of video uploaded to YouTube every minute as of February 2022. *Statista*.

CHENG, X. 2023. Named Entity Recognition in the Education Domain Based on BERT-BiLSTM-CRF-Using Data Structures as an Example. In *2023 International Conference on Educational Knowledge and Informatization (EKI)*, 5–9. IEEE.

CHERI, J., AND BHATTACHARYYA, P. 2017. Towards Harnessing Memory Networks for Coreference Resolution. *Meeting of the Association for Computational Linguistics*, 37–42.

COATS, S. 2023. A new corpus of geolocated ASR transcripts from Germany. *Language Resources and Evaluation*.

CHAI, H., AND STRUBE, M. 2022. Incorporating Centering Theory into Neural Coreference Resolution. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2996–3002.

COLAVIZZA, G., EHRMANN, M., AND BORTOLUZZI, F. 2019. Index-driven digitization and indexation of historical archives. *Frontiers in Digital Humanities* 6, 4.

DAI, Z., FEI, H., AND LI, P. 2019. Coreference aware representation learning for neural named entity recognition. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, 4946–4953. International Joint Conferences on Artificial Intelligence Organization.

DE VRIES, N. J., DAVEL, M. H., BADENHORST, J., BASSON, W. D., DE WET, F., BARNARD, E., AND DE WAAL, A. 2014. A smartphone-based ASR data collection tool for under-resourced languages. *Speech Communication* 56, 119–131.

DUMPALA, S. H., SHEIKH, I. A., CHAKRABORTY, R., AND KOPPARAPU, S. K. 2018. Sentiment Classification on Erroneous ASR Transcripts: A Multi View Learning Approach. *Spoken Language Technology Workshop*, 807–814.

ESCOTER, L., PIVOVAROVA, L., DU, M., KATINSKAIA, A., AND YANGARBER, R. 2017. Grouping business news stories based on salience of named entities. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, 1096–1106.

EFTEKHARI, H. 2024. Transcribing in the digital age: qualitative research practice utilising intelligent speech recognition technology. *European Journal of Cardiovascular Nursing*.

EHRMANN, M., HAMDI, A., PONTES, E. L., ROMANELLO, M., AND DOUCET, A. 2023. Named entity recognition and classification in historical documents: A survey. *ACM Computing Surveys* 56, 2, 1–47.

FYFIELD, M. E. B., BA, G. D. S., AND CERT, G. 2020. Selection and use of instructional videos by secondary teachers: knowledge and context. *Doctoral dissertation*, Monash University. doi:10.26180/13697608.v1.

FYFIELD, M. 2021. YouTube in the secondary classroom: How teachers use instructional videos in mainstream classrooms. *Technology, Pedagogy and Education* 31, 2, 185–197.

FERRACANE, E., MARSHALL, I., WALLACE, B. C., AND ERK, K. 2016. Leveraging coreference to identify arms in medical abstracts: An experimental study. In *Proceedings of the Seventh International Workshop on Health Text Mining and Information Analysis*, Austin, 86–95. Association for Computational Linguistics.

GUO, J., XU, G., CHENG, X., AND LI, H. 2009. Named entity recognition in query. In *Proceedings of the 32nd international ACM SIGIR conference on Research and Development in Information Retrieval*, 267–274.

GEY, F. C. 2000. Research to improve cross-language retrieval—Position paper for CLEF. In *Workshop of the Cross-Language Evaluation Forum for European Languages*, 83–88. Springer Berlin Heidelberg.

HU, C., BAI, S., AND ZHANG, M. 2024. Research on Named Entity Recognition for Oral History Text, 368–372.

HAYES, A. S. 2025. “Conversing” with qualitative data: Enhancing qualitative research through large language models (LLMs). *International Journal of Qualitative Methods* 24, 16094069251322346.

HINSVARK, A., DELWORTH, N., DEL RIO, M., MCNAMARA, Q., DONG, J., WESTERMAN, R., ... AND JETTE, M. 2021. Accented speech recognition: A survey. *arXiv preprint* arXiv:2104.10747.

HARABAGIU, S., AND MAIORANO, S. J. 1999. Knowledge-lean coreference resolution and its relation to textual cohesion and coherence. In *The Relation of Discourse/Dialogue Structure and Reference*.

HUANG, L., YU, W., MA, W., ZHONG, W., FENG, Z., WANG, H., ... AND LIU, T. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems* 43, 2, 1–55.

JEHANGIR, B., RADHAKRISHNAN, S., AND AGARWAL, R. 2023. A survey on named entity recognition—datasets, tools, and methodologies. *Natural Language Processing Journal* 3, 100017.

JIWER: SPEECH RECOGNITION EVALUATION IN PYTHON. 2025.

KUBSCH, M., KRIST, C., AND ROSENBERG, J. M. 2023. Distributing epistemic functions and tasks—A framework for augmenting human analytic power with machine learning in science education research. *Journal of Research in Science Teaching* 60, 2, 423–447.

KRISHNA, M. H., RAHAMATHULLA, K., AND AKBAR, A. 2017. A feature based approach for sentiment analysis using SVM and coreference resolution. In *2017 International Conference on Inventive Communication and Computational Technologies (ICICCT)*, 397–399.

KHASHABI, D., CHATURVEDI, S., ROTH, M., UPADHYAY, S., AND ROTH, D. 2018. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, New Orleans, 252–262. Association for Computational Linguistics.

KONG, F., AND JIAN, F. 2019. Incorporating Structural Information for Better Coreference Resolution. *International Joint Conference on Artificial Intelligence*, 5039–5045.

KUNDU, G., SIL, A., FLORIAN, R., AND HAMZA, W. 2018. Neural cross-lingual coreference resolution and its application to entity linking. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 395–400. Association for Computational Linguistics.

KOENECKE, A., NAM, A., LAKE, E., NUDELL, J., QUARTEY, M., MENGESHA, Z., ... AND GOEL, S. 2020. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences* 117, 14, 7684–7689.

LIN, Z., GUAN, S., ZHANG, W., ZHANG, H., LI, Y., AND ZHANG, H. 2024. Towards trustworthy LLMs: A review on debiasing and dehallucinating in large language models. *Artificial Intelligence Review* 57, 9, 243.

LIN, T., PANTEL, P., GAMON, M., KANNAN, A., AND FUXMAN, A. 2012. Active objects: Actions for entity-centric search. In *Proceedings of the 21st international conference on World Wide Web*, 589–598.

LANGE, P. G. 2019. Informal learning on YouTube. In *The international encyclopedia of media literacy*, 1–11.

LE, N. T., AND RITTER, A. 2024. Are language models robust coreference resolvers? In *First Conference on Language Modeling*.

LI, Y., BELL, P., AND LAI, C. 2024. Speech emotion recognition with ASR transcripts: A comprehensive study on word error rate and fusion techniques. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, 518–525. IEEE.

LI, X., WEI, C., JIANG, Z., OUYANG, F., ZHANG, Z., AND CHEN, W. 2023. EduNER: a Chinese named entity recognition dataset for education research. *Neural Computing and Applications*, 1–15.

LIU, Y., PENG, X., CAO, J., SHI, B., SHEN, Y., ZHANG, X., SHENG, C., WANG, X., YIN, J., AND DU, T. 2024. Bridging Context Gaps: Leveraging Coreference Resolution for Long Contextual Understanding.

MAO, R., AND LI, X. 2021. Bridging towers of multi-task learning with a gating mechanism for aspect-based sentiment analysis and sequential metaphor identification. *Proceedings of the AAAI Conference on Artificial Intelligence* 35, 15, 13534–13542.

MOLLÁ, D., VAN ZAANEN, M., AND SMITH, D. 2006. Named entity recognition for question answering. In *Australasian Language Technology Association Workshop*, 51–58.

MILES, J., COMPTON, A., AND HEROLD, E. 2024. Crash Course in the classroom: Exploring how and why social studies teachers use YouTube videos. *The Journal of Social Studies Research* 48, 3, 190–203.

NELSON, L. K. 2020. Computational grounded theory: A methodological framework. *Sociological Methods & Research* 49, 1, 3–42.

OLATUNJI, T., AFONJA, T., DOSSOU, B. F., TONJA, A. L., EMEZUE, C. C., RUFAI, A. M., AND SINGH, S. 2023. Afrinames: Most AST models "butcher" African names. *arXiv preprint* arXiv:2306.00253.

POLIKOFF, M. 2019. The supplemental curriculum bazaar: Is what's online any good? *Thomas B. Fordham Institute*.

PARAMESWARAN, U. D., OZAWA-KIRK, J. L., AND LATENDRESSE, G. 2020. To live (code) or to not: A new method for coding in qualitative research. *Qualitative Social Work* 19, 4, 630–644.

ROSENBERG, J. M., AND KRIST, C. 2021. Combining machine learning and qualitative methods to elaborate students' ideas about the generality of their model-based explanations. *Journal of Science Education and Technology* 30, 2, 255–267.

RAO, A. 2023. Transcribing Educational Videos Using Whisper: A preliminary study on using AI for transcribing educational videos. *arXiv.Org*, abs/2307.03200.

RASK, M., AND SHIMIZU, K. 2024. Beyond the Average: Exploring the Potential and Challenges of Large Language Models in Social Science Research. In *2024 International Conference on Artificial Intelligence, Computer, Data Sciences and Applications (ACDSA)*, 1–5. IEEE.

SAVARAM, P., TABASSUM, S., AND BANDU, S. 2024. Multilingual approaches to named entity recognition. In *2024 3rd Edition of IEEE Delhi Section Flagship Conference (DELCON)*, 1–6. IEEE.

STEINBERGER, R., POULIQUEN, B., AND VAN DER GOO, E. 2009. An introduction to the Europe media monitor. In *Proceedings of ACM SIGIR 2009 workshop: Information access in a multilingual world*, Boston, MA, USA.

SCHWITTER, N. 2025. Using large language models for preprocessing and information extraction from unstructured text: A proof-of-concept application in the social sciences. *Methodological Innovations*.

SONAWANE, S. S., AND KULKARNI, P. 2015. Entity based co-reference resolution with name entity recognition using hierarchical classification. *IEEE India Conference*, 1–6.

SHI, S., HUANG, H.-Y., AND CHEN, R.-Y. 2010. A method of Chinese coreference resolution combined multi-features in discourse. *International Conference on Machine Learning and Cybernetics* 3, 1311–1316.

TAI, R. H., BENTLEY, L. R., XIA, X., SITT, J. M., FANKHAUSER, S. C., CHICAS-MOSIER, A. M., AND MONTEITH, B. G. 2024. An examination of the use of large language models to aid analysis of textual data. *International Journal of Qualitative Methods* 23, 16094069241231168.

TOSH, K., DOAN, S., WOO, A., AND HENRY, D. 2020. Digital instructional materials: What are teachers using and what barriers exist? Data note: Insights from the American Educator Panels. *Research Report*. RR-2575/17-BMGF/SFF/OFF. RAND Corporation.

TENG, D., ZHANG, X., XING, X., CHEN, P., AND YANG, C. 2023. Coreference Resolution Method Integrating Textual Information and Semantic Assessment, 397–403.

THAN, N., FAN, L., LAW, T., NELSON, L. K., AND MCCALL, L. 2025. Updating "The Future of Coding": Qualitative coding with generative large language models. *Sociological Methods & Research* 54, 3, 849–888.

WANG, T., AND LI, H. 2020. Coreference resolution improves educational knowledge graph construction. In *2020 IEEE International Conference on Knowledge Graph (ICKG)*, 629–634. IEEE.

WEI, K., AND WEN, B. 2021. Named Entity Recognition Method for Educational Emergency Field Based on BERT. *International Conference on Software Engineering*.

WELBL, J., STENETORP, P., AND RIEDEL, S. 2018. Constructing datasets for multi-hop reading comprehension across documents. *Transactions of the Association for Computational Linguistics* 6, 287–302.

YOUTUBE. N.D. YouTube by the numbers. *YouTube Official Blog.* https://blog.youtube/press/

ZHANG, B. 2024. Getting to Know Named Entity Recognition: Better Information Retrieval. *Medical Reference Services Quarterly* 43, 2, 196–202.

ZHU, P., ZHANG, Z., LI, J., HUANG, Y., AND ZHAO, H. 2018. Lingke: A fine-grained multi-turn chatbot for customer service. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, Santa Fe, 108–112. Association for Computational Linguistics.

APPENDIX A. PROMPT DESIGN AND API IMPLEMENTATION DETAILS

A.1 COREFERENCE RESOLUTION PROMPT AND IMPLEMENTATION

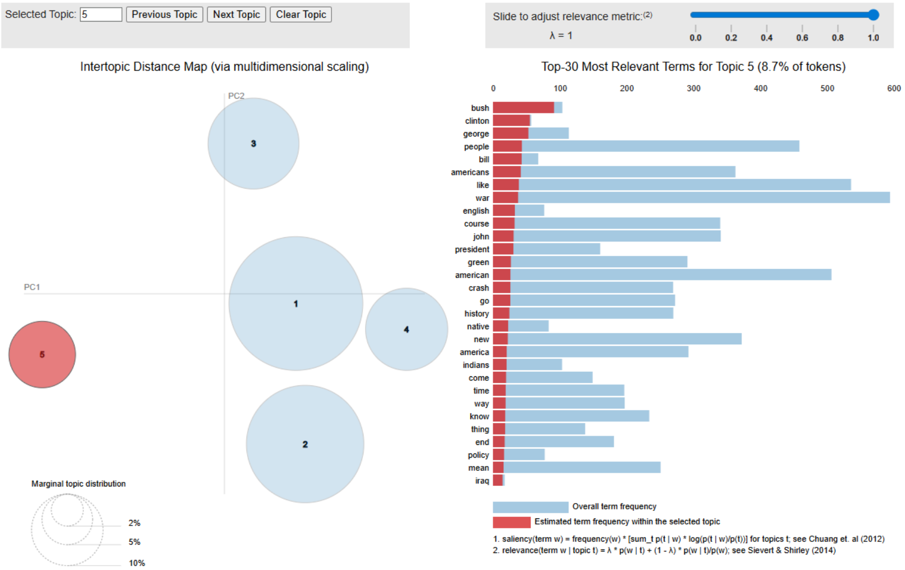
```
def resolve_coreferences_with_openai(text: str, model="gpt-4o"):
    """
    call openai api to resolve coreferences.
    """
    prompt = (
        "Below is a paragraph of text.  \n"
        "Please replace each pronoun (he, she, they, etc.) with the proper name or noun it refers to. \n"
        "Keep all other words exactly as they appear.\n\n"
        + text
    )
    try:
        response = openai.chat.completions.create(
            model=model,
            messages=[
                {"role": "system", "content": "You are an assistant skilled in coreference resolution."},
                {"role": "user", "content": prompt}
            ],
            temperature=0.0,
            top_p=1.0
        )
        return response.choices[0].message.content.strip()
    except Exception as e:
        print(f"[WARN] OpenAI API error: {e}. Returning original text.")
        return text
```

A.2 NER Enhancement Prompt

```
  prompt = (
    "You are an expert in US history and advanced Named Entity Recognition (NER). "
    "You will be given the full transcript of a video subtitle file. "
    "Your job is to carefully review the entire text and ensure that all named entities are correctly identified and accurately standardized, focusing on the following entity types:\n"
    "PERSON: Historical or public figures (e.g., 'Abraham Lincoln', 'Frederick Douglass').\n"
    "ORG: Organizations, institutions, or historical groups (e.g., 'Continental Congress', 'Supreme Court').\n"
    "GPE: Countries, cities, or regions (e.g., 'United States', 'Virginia').\n"
    "LOC: Geographic locations that are not geopolitical (e.g., 'Appalachian Mountains', 'Mississippi River').\n"
    "NORP: Nationalities, religious groups, or political groups (e.g., 'Unionists', 'Quakers', 'Republicans').\n"
    "EVENT: Named historical events (e.g., 'Civil War', 'Boston Tea Party').\n"
    "LAW: Historical legal documents or acts (e.g., 'Emancipation Proclamation', 'Bill of Rights').\n\n"
    "Instructions:\n"
    "1. For each entity above, correct any errors, misspellings, abbreviations, or ambiguous mentions, replacing them with the full, precise, and canonical historical name.\n"
    "2. Replace pronouns or vague references with their explicit entity name only when the reference is clear.\n"
    "3. Do not change any other text. Do not add explanations, comments, tags, or notes.\n"
    "4. Output only the corrected full transcript, preserving the original format and line breaks. Your output should be a fully corrected subtitle file ready for saving as a .txt file.\n\n"
    "Text to review:\n"
    + text
)
```

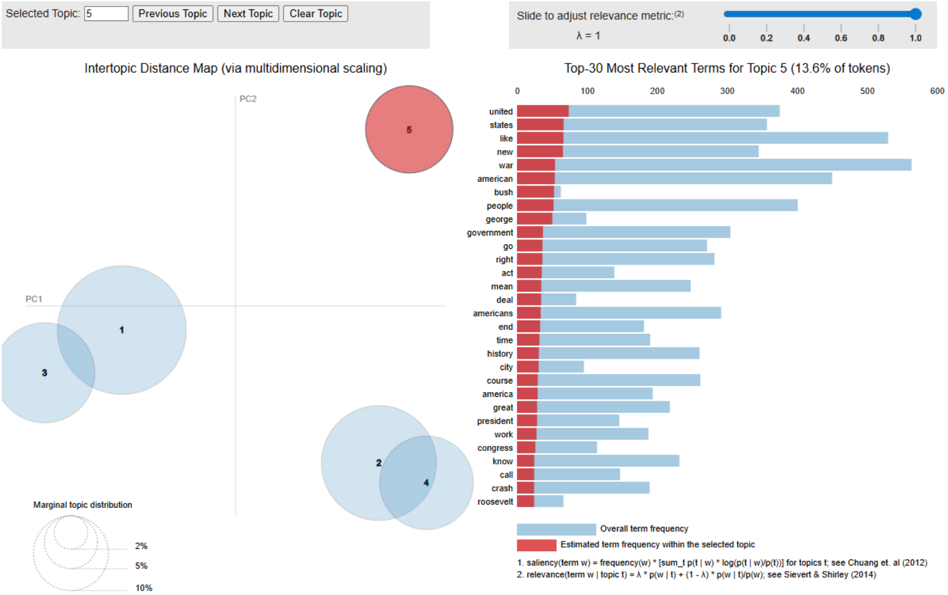
Appendix B. PyLDAvis Intertopic Distance Map Figures

Appendix Figure B1. PyLDAvis Intertopic Distance Map for LLM Coreference-Resolved Transcripts



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Appendix Figure B2. PyLDAvis Intertopic Distance Map for LLM NER-Enhanced Transcripts



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)