RESEARCH ARTICLE UHD JOURNAL OF SCIENCE AND TECHNOLOGY

LectBench-95: Preparing a University-Lecture Corpus for A/B Evaluation of Lecture-Processing Techniques

Ari A. Aziz, Aree A. Mohammed

Department of Computer Science, College of Science, University of Sulaimani, Sulaymaniyah, Kurdistan Region, Iraq.

ABSTRACT

Tools for processing educational lectures are rapidly advancing, but there is a need for a diverse, balanced, and high-quality university lecture transcript corpus. The existing datasets are either limited to K-12, tutorial styled, and lack interactivity, focus on narrow disciplines, paywalled/non-accessible, or are impractically large. We introduce LectBench-95, a publicly available corpus of 95 video lecture transcripts spanning 3 disciplines and 17 specific subjects within them. With strict filtering for high audio quality (SNR ≥25 dB), transcription confidence (Mean 0.84, Min 0.7), transcript quality controls, and a power analysis-guided sample size, the 94-h dataset aims to detect ≥20% performance differences between competing systems with 95% confidence for head-to-head A/B experiments. LectBench-95 contains 816 k words (unique ≈ 123 k), a mean Measure of Textual Lexical Diversity of 54.5, and a mean speech rate of 144 words/min, mirroring real-world university lectures. A toy A/B test on zero-shot summarization (Gemini-1.5-Flash vs. 1.5-Flash-8B) shows the corpus's utility, resulting in a statistically significant 43% win-rate gap with $P = 3 \times 10^{-5}$. Released under CC BY-NC-SA 4.0, LectBench-95 provides a modest yet statistically robust dataset for future educational natural language processing research and prototyping.

Index Terms: University Lecture Corpus, A/B Evaluation, Lecture Processing Techniques, Educational Natural Language Processing

1. INTRODUCTION

In the information age, advancing pedagogical tools are more important than ever, requiring every step in the development process to be efficient and effective. Evaluation, in particular, can strongly impact the success of such tools. For lecture processing techniques, this means a dataset that is reflective of real-world university lectures. Such techniques have been on the rise, especially with the advent of Large Language Models (LLMs) [1], [2]. Examples of educational LLM-based tools include: question answering teacher assistant [3]; screening children's language development levels [4]; providing lesson plans, activities, and materials [5]; and automatic grading of student responses [6]. However, the advancement of these tools requires a solid dataset that provides accurate feedback for pinpointing improvement areas.

Current university-lecture corpora often lack the properties needed for reliable A/B comparisons of lecture processing techniques on classroom discourse, including (1) interactive, professor-led lectures rather than tutorials and monologues; (2) multidisciplinary coverage that reveals subject-specific blind spots; (3) rigorous filtering and transcript quality controls; (4) public accessibility at a practical size. This

Access this article online

DOI:10.21928/uhdjst.v9n2y2025.pp216-230 E-ISSN: 2521-4217 P-ISSN: 2521-4209

Corresponding author's e-mail: Ari A. Aziz, Department of Computer Science, College of Science, University of Sulaimani, Sulaymaniyah, Kurdistan Region, Iraq. E-mail: ari.aryan@univsul.edu.iq

Received: 26-06-2025 Accepted: 23-08-2025 Published: 11-10-2025

limits both model evaluators who are selecting systems and researchers studying classroom discourse.

To address this, we introduce LectBench-95, a dataset of 95 carefully curated, diverse, and high-quality university lecture transcripts, sourced from various institutions. It is designed for head-to-head A/B experiments in mind, with pre-declared acceptance criteria: (a) ≥3 subjects per area across STEM/Humanities/Social Sciences; (b) SNR ≥25 dB and mean Whisper confidence ≥0.70 per lecture; (c) 30–120 min lecture durations; and (d) a statistical power-aware sample size that targeting the detection of moderate win-rate differences in pairwise tests. Each transcript includes metadata and segment-level timestamps, allowing for use cases that go beyond those outlined in this paper. this dataset is publicly available at this link https://osf.io/astqv/.

This paper provides a comprehensive overview of the LectBench-95 dataset, including its collection, processing, and the rationale behind its design choices. Detailed statistics are presented to characterize the dataset and its applications. To evaluate the statistical significance of the results obtained from experimenting with this dataset, a toy A/B experiment is conducted that demonstrates the dataset's benefits when it comes to evaluating two simple lecture summarization techniques.

The videos are collected from publicly available university lectures and transcribed using a state-of-the-art speech-to-text system. The rest of the paper details the process.

Our contributions include:

1. A curated, interactive university-lecture corpus spanning 17 subjects with timestamps and quality summaries
2. A transparent method that allows selecting and auditing for interactivity
3. A power-aware evaluation recipe for paired win-rate comparisons
4. A demonstration of zero-shot summarization showing statistically significant model differentiation.

The remainder of the paper is organized as follows: Section 2 discusses the current educational lecture datasets. Section 3 details the dataset's construction, including collection and processing, statistical design, and lecture selection criteria. Section 4 provides both a quick statistical summary and a more detailed breakdown of the dataset's characteristics, including lexical, speech rate, and quality statistics. Section 5 presents an A/B experiment using the dataset to demonstrate its utility in evaluating lecture processing techniques. Section 6 outlines the intended use cases (including implementation and analysis opportunities) and limitations of the dataset, both of which suggest future work.

2. RELATED RESOURCES

Diverse open-source datasets in the domain of university lectures are somewhat scarce, especially those that focus on traditional teaching. This scarcity creates obstacles for researchers who require rich, varied data reflective of real classroom dynamics for developing various types of educational technologies. The current corpora often focus narrowly on specific subjects or delivery formats (e.g., online tutorials); they may even focus more thoroughly on pre-university level courses. While there are also some that can tick these boxes, they may fall short in terms of accessibility or usability. In this section, such datasets are explored, and arguments will be made for the limitations that may hold true in this context.

The National Center for Teacher Effectiveness (NCTE) Transcripts [7] is one dataset, which contains 1660 elementary math classroom transcripts, each 45–60 min long, collected by the NCTE between 2010 and 2013. The lessons are from 4th and 5th grade math classes in the United States. It includes rich turn-level metadata of dialogic discourse, such as questioning and prompting. Furthermore, it includes classroom observation scores, teacher and student demographics, survey responses, and student test scores. However, the dataset is limited only to elementary math classes and does not include higher education or other diverse subjects. Another somewhat similar dataset is the MET dataset [8], which contains 2500 4–9th grade classroom recordings collected between 2009 and 2011 in the US. It spanned multiple subjects, including English Language Arts and mathematics. Its downsides are that it requires a subscription; it only focuses on 4th–9th-grade lessons, and it does not provide transcripts. The TalkMoves dataset [9] is another math classroom transcript dataset containing 567 K-12 math lectures, also collected in the US. Despite being publicly available, it is limited in terms of subjects and to K-12 education. Another dialogue dataset is the CIMA dataset [10], which is designed specifically with training deep learning models in mind for Italian tutoring. It was collected through asynchronous role-playing by crowd-workers meant to depict Italian tutoring lessons. It is ideal for tutor agent training, but is limited to Italian. Rai *et al.* [11] investigate word error rates (WER) between YouTube Automatic Captions and OpenAI's Whisper model. It focuses on NPTEL (a large

MOOC platform in India) videos, and it collects 8740 h of Indian English lecture audio covering 9800 lectures sourced from the aforementioned platform. The Technical Indian English dataset mentioned includes audio, transcripts, and speaker metadata. This dataset focuses mostly on non-native varieties of English (specifically Indian English) that are technically dense for the purpose of improving ASR systems. The Autoblog 2020 [12] and its subsequent version Autoblog 2021 [13] are two corpora—the latter is publicly available on Kaggle—that focus on online learning video lectures. The main aim of Autoblog 2020 was to convert lectures into easy-to-consume blog posts in an automated manner, while its subsequent version aimed to increase the dataset by adding manual transcription for ASR benchmarking of spontaneous speech. Autoblog 2021 includes 63 lecture videos and audio files, manual and automatic transcriptions, slide images, and more. Its limitations include that it is not a traditional in-class lecture with teacher-student engagements in a classroom; rather, the teaching is performed in a monologous manner in an online setting. Furthermore, it is limited only to Medical Engineering and Pattern Recognition. The Automatically Recognizing Lecture Highlighting Corpus [14] is another corpus meant for training speech processing models for recognizing moments where the lecturer highlights or emphasizes a certain word using vocal emphasis. The dataset features 104 different English speakers from various disciplines derived from YouTube tutorial videos. Despite the authors saying that it would be made freely available to the community, no access link was found to the dataset as of the time of writing this paper. Furthermore, it focuses on tutorials rather than formal university lecture transcripts. VT-SSum [15] is a benchmark dataset that contains 125K transcript-summary pairs derived from 9,616 video transcripts from videolectures.net across 26 different categories. It focuses on spoken language summarization and segmentation and leverages the accompanying slides as weak supervision. However, it uses Microsoft speech-to-text, while this dataset utilizes the more modern, accurate, and open-source OpenAI Whisper [16]. The AVLectures dataset [17] is another dataset that focuses on audio-visual lecture segmentation and summarization. It has 86 courses (15 of them manually segmented), encompassing more than 2350 lectures (2,200 h total) spanning STEM fields such as EECS, Physics, and Mathematics. Despite being publicly accessible and despite the authors claiming that you can access each of the course's tarballs individually, the dataset webpage[1] provides 3 tarballs only where one is 340 GB and the other two are 79 GB and 1 GB, respectively. This makes it impractical to use for researchers seeking smaller or modest-sized datasets. This argument is true for the Slide Speech dataset [18] as well.

In addition, several other datasets comprise MOOC lecture transcripts, which do not align with the goal of collecting traditional university lectures. For example, the Khan Academy Corpus [19] contains a large collection of transcripts from Khan Academy videos, which are primarily tutorial style and also require affiliation with certain universities to access. Similarly, the TED-LIUM [20] and TED-LIUM 3 [21] datasets contain a large amount of audio and transcripts from TED talks, which are also not traditional university lectures. Lee *et al.* [22] also have the problem of including untraditional tutorial-style lectures with no teacher–student interactions.

Collectively, these resources highlight many shortcomings with the current status quo of university-level educational corpora and the lack of accessible, rich, and interactive dialogue within traditional university lecture settings. This work aims to fill these gaps by curating LectBench-95. Table 1 summarizes the limitations of the aforementioned datasets for better readability.

3. DATASET CONSTRUCTION

In this section, a detailed overview of data collection, processing pipeline and the statistical planning that went into the dataset construction is provided. The criteria used for lecture selection will also be discussed.

3.1. Collection and Processing Pipeline

The videos were downloaded, and their audio was extracted using the yt-dlp[2] library. Which is then passed on to OpenAI's Whisper[3] [23] library. The choice of Whisper was guided by its robustness and low WER [24]. This pipeline yields a segmented transcript containing the text along with other useful metadata such as the start–end times of a segment, temperature, tokens, and segment IDs. However, only the text itself, the start- and end-times, and the segment IDs are needed. The segments are concatenated in the following format:

[Segment ID] [Start Time - End Time] Text Content

where:

- Segment ID: A zero-padded, three-digit identifier for

1 https://india-data.org/dataset-version/52e67bb5-5acd-438d-a13b-f255cee17432/5f9005e0-8a6b-4180-bc52-fdda80a6917a

2 https://github.com/yt-dlp/yt-dlp

3 https://github.com/openai/whisper — Specifically the small variant.

TABLE 1: Limitations of existing educational lecture datasets

Dataset	Limit. Subjects	Pre-Univ.	Access Restricted	No Transcripts	Non-Interactive	Lang. Limit.	Large Size
NCTE Transcripts	✓	✓					
MET Dataset		✓	✓	✓			
TalkMoves	✓	✓					
CIMA						✓	
Technical Indian English						✓	
Autoblog 2021	✓				✓		
Automatically Recognizing Lecture Highlighting Corpus			✓		✓		
VT-SSum					✓		
AVLectures	✓				✓		✓
Slide Speech					✓		✓
Khan Academy Corpus			✓		✓		
TED-LIUM					✓		
Lee et al.							

Limitations key:
1. Limited Subjects: Focuses on specific subjects (e.g., only math/STEM)
2. Pre-University: Primarily K-12 education level
3. Access Restricted: Requires subscription, affiliation, or unavailable
4. No Transcripts: Only audio/video available
5. Non-Interactive: Monologue without teacher–student interactions
6. Language Limitations: Focuses on specific languages or dialects
7. Large Size: Impractical for modest computing resources

the segment (e.g., 001, 002, etc.). It is used to uniquely identify each segment and to allow the LLM to reference it during the evaluation

- Start Time and End Time: The timestamps indicating the beginning and end of the segment, formatted as MM: SS.ss, where MM represents minutes, SS represents seconds, and ss represents fractional seconds rounded to two decimal places
- Text Content: The transcribed text of the segment, with leading and trailing whitespace removed.

Because interactivity was an afterthought for this dataset, and its value became apparent after going through the literature review, it was necessary to recheck all the videos to determine whether or not they are interactive. The traditional way would be to open each video and check if the lecturer is in a traditional classroom setting with students present, classifying it as interactive. Conversely, a MOOC-style video would be considered non-interactive. In some cases, however, the lecture may be in a Zoom-like setting where the lecturer is talking to a camera without any students present, but the lecturer may still be interactive with the audience. In these cases, we would still classify it as an interactive lecture[4]. Regardless, this method is cumbersome and time-consuming; therefore, we adopted a different approach. A Google Sheet was created with the video links, titles, broad subject, and specific subject, and a column was created to extract the video's link using the following cell formula:

=REGEXEXTRACT(B2, "v=([a-zA-Z0-9_-]+)")

Where B2 is the cell containing the video link. Afterward, the video ID is used to create 4 other cells that fetch the video's storyboard from YouTube servers; each cell fetches a different storyboard image using the following formula:

=IMAGE("https://i.ytimg.com/vi/" & D2 & "/0.jpg", 2)

Here, D2 is the cell containing the video ID, and the 0.jpg in the URL represents the storyboard's image index. The other 3 cells use 1.jpg, 2.jpg, and 3.jpg, respectively. Then, another column is created to hold a boolean value indicating whether the video is interactive or not, as judged by the storyboard images. If the storyboard images show a classroom setting with students present, it is marked as TRUE (i.e., interactive); otherwise, it is marked as FALSE (i.e., non-interactive). After that, Google Sheets' data-validation feature was used to create a Checkbox for that column, allowing toggling the interactivity status with a click instead of editing. This way, determining whether the video is interactive or not is done in a matter of seconds, without the need to open each video, waiting for it to load, and skipping to the relevant parts.

4 These videos make up a very small portion of the dataset, and were only included in order to meet the subjects quota. Preference was given to videos within physical classrooms

This approach is not perfect, but it is a good approximation that saves us a lot of time and effort. A screenshot of the Google Sheet is shown in Appendix Fig. 1. (Appendix A).

Now, for those videos that appear to be in a zoom-like setting, the storyboard images may not be enough to determine interactivity; manual checks were performed to determine their interactivity. If a video is non-interactive, it is simply discarded, and an interactive alternative on the same subject is sought.

Appendix B describes the attempts made at anonymizing the dataset. However, anonymization was not feasible given the current resources and time constraints we had. Furthermore, the lectures are all publicly available on YouTube, thus anonymization is not strictly necessary in this case. As a cautionary note, users are strongly encouraged to anonymize the dataset if their use case requires it.

3.2. Statistical Design

Each transcript is considered a single sample and is used in a contest between the different techniques. With 95 transcripts, each pairwise comparison between two techniques is performed across all 95 transcripts, providing 95 independent contests per pair. The following paragraphs explain how this number is determined and why it is sufficient to distinguish the performance superiority of a technique in an A/B test.

The sample size was not randomly chosen; the aim was to have a sample size that would be large enough to be statistically significant and small enough to be feasible to collect and process. Thus, a power analysis was used to determine the sample size needed to achieve 80% power with a significance level of 0.05, which corresponds to a 95% confidence level. Where the win-rate of technique A is $p_1 = 0.4$ and technique B is $p_2 = 0.6$, i.e., 20% effect size. p_1 and p_2 were chosen to cover the most demanding case in terms of sample size (more on this next). The main goal here is to ensure (with 95% confidence) that a 20% difference in the win rates of a pair of techniques will be detected if it exists. Larger effects (e.g. 30% or 40%, etc.) are inherently easier to detect as they require smaller sample sizes.

A two-proportion z-test is used as the basis to determine the sample size needed, the best fit for our scenario, as it assumes a binomial distribution of binary outcomes (i.e., win/lose scenario of an A/B experiment) and independent samples. Originally, the two-proportion z-test is used to determine the statistical significance of the difference between two proportions. However, by rearranging the formula[5], it can be used to determine the sample size needed to achieve a certain level of statistical significance. Equation 1 shows the formula used to calculate the sample size needed for each group.

Sample size calculation formula:

$$n = \frac{(Z_{\alpha/2} + Z_{\beta})^2 . (p_1(1-p_1) + p_2(1-p_2))}{(p_1 - p_2)^2} \quad (1)$$

Where:

- n is the sample size needed for each group
- $z_{\alpha/2}$ is the z-score corresponding to the significance level (0.05). Which is 1.96 for a two-tailed test
- z_{β} is the z-score corresponding to the power (0.80). Which is 0.84 for a power of 80%
- p_1 is the baseline proportion (e.g., win rate of technique A). Set to 0.4 in this case
- p_2 is the expected proportion (e.g., win rate of technique B). Set to 0.6 in this case.

The choice of a 20% effect size was not arbitrary; rather, it was made to strike the aforementioned balance between statistical significance and data collection feasibility derived, as shown in Fig. 1. The Figure illustrates the different effect sizes and baselines along with their corresponding sample sizes. As can be seen from the plot, the required sample size increases when the baselines are near 0.5, where it is the most difficult to distinguish which technique is superior, thus more samples are needed. In contrast, when the baseline is near 0 or 1, the required sample size is smaller as it is easier to detect a difference. Typically, the smallest effect size with the largest sample size is sought to ensure the most demanding case is covered. In this case, that would be the 10% effect size with a baseline of 0.5, which requires a sample size of $n = 385$. However, collecting 385 samples would be time-consuming and resource-intensive. Due to the impracticality of gathering such a big sample size, the 20% effect size with a baseline of 0.4, which requires a sample size of $n = 95$ will be chosen. This should give us a good balance between statistical significance and data collection feasibility.

3.3. Data Selection Criteria

To ensure the transcripts are relevant, high-quality, and diverse, the following criteria were put in place:

- The transcripts should be only traditional, professor-led

5 Assuming equal sample sizes for both proportions, i.e., $n_1 = n_2$ which is standard for experiment designs.

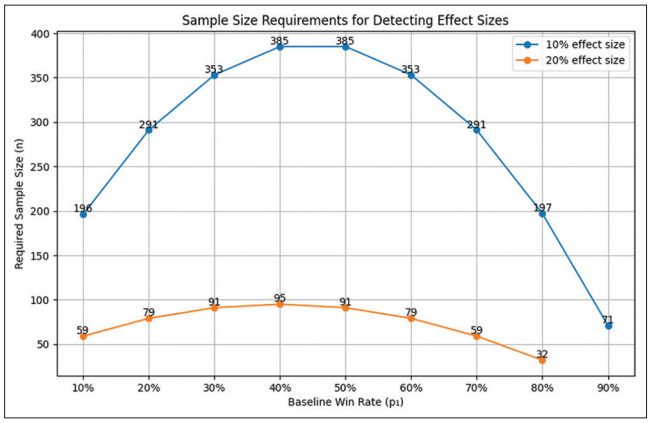


Fig. 1. Required sample size for different effect sizes and baseline win-rates ($p1$). Note that p2 = p1+effect size.

academic lectures with a clear presence of both students and teachers in the classroom–be it physical or virtual. No tutorials or non-academic content were included. This is to ensure that the content is relevant to the intended use cases

- The transcripts should be in English, as the majority of the academic content in the public domain is in English
- The transcripts range from 30 min all the way up to 2 h, which is the typical length of a university lecture
- The transcripts must have clear audio quality evidenced through a high signal-to-noise ratio (SNR) and Whisper's own segment-level average confidence scores (derived from avg_logprob). The thresholds are ≥25 dB for SNR and ≥70% for the average confidence level across all the segments. This is to ensure that mistranscription does not skew any results
- The transcripts should be from a diverse set of universities (public and private) and professors across different academic fields. To be more precise, a distribution of 40% STEM (38 transcripts), ≈ 30% Humanities (28 transcripts), ≈ 30% Social Sciences (29 transcripts) was targeted. Fig. 2 shows the full breakdown of the transcripts across different academic fields.

4. DATASET ANALYSIS

In this section, an overview of the LectBench-95 dataset will be provided, making clear its composition, characteristics, and key statistics.

4.1. General Stats

Table 2 provides a quick overview of the dataset, showing the key statistics such as the total number of lectures, total audio duration, and average lecture duration, among others.

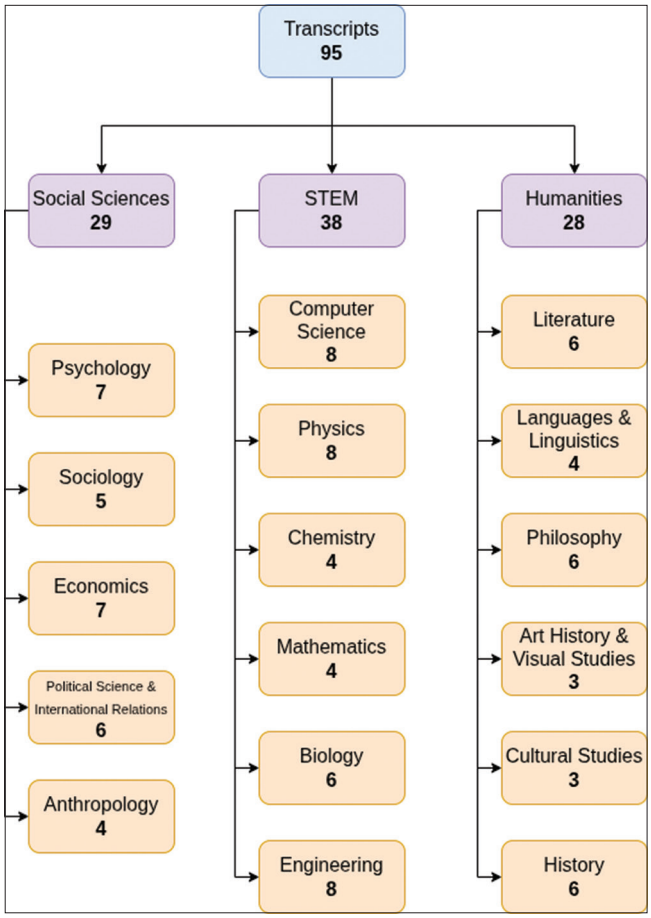


Fig. 2. Distribution of the transcripts across different academic fields.

Meanwhile, Table 3 provides a more detailed breakdown of the dataset's statistics, including temporal characteristics, speech characteristics, lexical diversity, and quality metrics.

The dataset contains 95 lecture transcripts, each with segments and metadata such as video publisher, title, URL, duration (in minutes), broad subject (i.e., discipline), and specific subject. With a total lecture audio duration of 94.06 h and an average lecture duration of 59.41 min (SD: 15.67), three lectures exceeded 90 min. The lectures span 3 disciplines and 17 subject areas, with each subject area having at least 3 lectures, as shown in Fig. 2. Furthermore, the dataset is lexically rich, with a total word count of 816,214 and a unique word count of 122,696. The Measure of Textual Lexical Diversity (MTLD) [25] is mainly used to assess lexical diversity and yields a mean score of 54.49 in this case. The flow of speech in the dataset is also noteworthy, with an average speech rate of 144.3 words per minute (WPM) and a standard deviation of 21.0 WPM, which aligns with typical speech rates in academic settings [26]. The transcripts were

generated using OpenAI's Whisper model, which provides an average log probability for each segment, from which the confidence scores are derived. The mean confidence score across all transcripts is 0.8402 (SD: 0.0508, Min: 0.7018), indicating a solid level of transcription quality. Segment-level timestamps are also available, allowing for any temporal analysis or alignment tasks.

These quick statistics should give insights into the dataset's scale and composition. The upcoming sections will dive even deeper into the dataset's statistics, including lexical, speech rate, and quality metrics.

TABLE 2: LectBench-95 dataset summary statistics

Metric	Value
General characteristics	
Total number of lectures	95
Total audio duration	94.06 h (5,644 min)
Average lecture duration	59.41 min (SD: 15.67)
Disciplines	3
Subject areas (all ≥3 lectures)	17
Lexical statistics	
Total word count	816,214
Unique word count	122,696
Vocabulary richness	MTLD: 54.49 (mean), TTR: 0.1578 (mean)
Speech rate	144.3 WPM (SD: 21.0)
Transcript quality	
Confidence score (Whisper)	Mean: 0.8402 (SD: 0.0508, Min: 0.7018)
Segment-level timestamps	Available

4.2. Statistical Breakdown and Quality Validation

In this section, the quality of the dataset will be validated by comparing the statistics against existing literature and benchmarks. This will tell us whether or not the dataset is reflective of real-world university lectures. The dataset validation is approached from three main angles: lexical statistics, speech rate statistics, and quality metrics. For each, a general overview and a subject-specific breakdown will be provided.

4.2.1. Lexical statistics

In terms of lexical statistics, the dataset shows a high degree of vocabulary richness, as evidenced in Fig. 4, which categorically illustrates the distribution of MTLD scores across lectures. As can be seen, around 97.9% of the lectures fall within the medium to high diversity categories; the majority being high diversity (64.2%), with only 2.1% classified as low-diversity.

Besides being an indicator of transcript quality, lexical diversity is also crucial for assessing the ability of lecture processing tools to handle the rich and varied vocabulary common in academia.

Turning our attention to subject-specific analysis, notable variations in lexical diversity are observed across different academic disciplines and specific subject areas within them.

TABLE 3: Full descriptive statistics of the LectBench-95 dataset

Metric	n	Mean	Standard	Min	Q1	Median	Q3	Max
Duration and temporal								
Duration (minutes)	95	59.41	15.75	33.05	47.81	54.82	72.20	104.57
Duration (hours)	95	0.99	0.26	0.55	0.80	0.91	1.20	1.74
Segment Count	95	803	366	299	553	700	955	2181
Avg segment duration (seconds)	95	5.05	2.04	2.13	3.53	4.42	6.07	13.02
Speech and linguistic rate								
Speech rate (WPM)	95	144.3	21.1	96.9	130.6	141.7	156.6	201.2
Syllable rate (SPM)	95	203.6	27.4	127.6	184.7	205.8	219.0	266.6
Word count	95	8591	2693	3965	6623	7960	10340	15585
Avg word length	95	4.19	0.25	3.66	4.03	4.18	4.34	4.95
Lexical diversity								
MTLD score	95	54.49	11.62	32.28	46.42	52.47	60.34	90.35
Type-Token ratio	95	0.158	0.043	0.074	0.127	0.153	0.187	0.282
Vocabulary size	95	1291	345	513	1086	1254	1484	2381
Hapax ratio	95	0.079	0.031	0.024	0.057	0.075	0.098	0.173
Sophistication ratio	95	0.165	0.033	0.098	0.146	0.163	0.183	0.282
Transcription quality								
Avg confidence	95	0.840	0.051	0.702	0.816	0.846	0.882	0.912
Low confidence ratio	95	0.060	0.100	0.00	0.008	0.025	0.058	0.467
Min confidence	95	0.552	0.163	0.004	0.448	0.580	0.647	0.830
Max confidence	95	0.924	0.043	0.748	0.913	0.937	0.951	0.970

Notes: *n*=Sample size, Q_1=First quartile, Q_3=Third quartile. WPM: Words per minute, SPM: Syllables per minute, MTLD: Measure of textual lexical diversity. Duration metrics measured across 95 educational video transcripts. Low confidence ratio indicates the proportion of transcript segments with confidence<0.7

Fig. 3 presents a ranking of subjects based on their mean MTLD scores, showing how diverse each subject's vocabulary is. For instance, art History and Visual Studies exhibit the highest mean MTLD score of 88.49 (±1.75), whereas Mathematics is the lowest with a mean MTLD score of 40.26 (±2.67), indicating a more specialized vocabulary. A trend that can easily be observed is that the more technical the subject, the lower the MTLD score. This trend is also backed by the discipline-based analysis conducted, in which Humanities and Social Sciences both have higher mean MTLD scores (58.5 and 58.3, respectively) compared to STEM subjects (48.6). This trend is consistent with existing literature that suggests PhD dissertations in the Humanities and Social Sciences tend to have higher lexical diversity compared to those in STEM fields, where more specialized and repetitive words are used [27].

Appendix Table 1 (Appendix C) provides a table with more statistics on the lexical diversity of the dataset.

4.2.2 Speech rate statistics

The dataset also exhibits substantial diversity when it comes to speech rate. Fig. 6 shows the distribution of lectures across different speech rate categories. As is evident, the majority of lectures (approximately 70%) fall within the Normal (120–160 WPM) range. In addition, a good portion of lectures fall in the slow (<120 WPM) and fast (>160 WPM) categories, with approximately 10% and 20%, respectively. This diversity in speech rate allows researchers both to evaluate the performance of lecture processing techniques on the most common speech rates (i.e., Normal) and to test their robustness on the tails of the distribution (i.e., Slow and Fast). This is crucial for a comprehensive evaluation of a certain technique.

As shown in Fig. 5, the variance of the average speech rate across different subjects can be noted. The subjects with the fastest speech rates are Computer Science (mean: 170.3 WPM, SD: 25.1), followed by Languages and Linguistics (mean: 161.2 WPM, SD: 12.3), whereas the slowest are Art History and Visual Studies (mean: 129.9 WPM, SD: 8.7) and Physics (mean: 134.0 WPM, SD: 22.9). It can be seen that the two fastest and two slowest subjects both fall in STEM and Humanities. This variation tells us that consistency cannot be detected in speech rate across disciplines, as the speech rate seems to be more dependent on the specific subject area. Other factors such as the lecturer's personal style, the complexity of the subject matter, and the intended audience may also play a significant role in determining speech rate. Looking at a subject such as Art History and Visual Studies, under the light of the previous lexical diversity analysis, wherein it was shown to have the richest vocabulary, coupled with the slowest speech rate, it can be concluded that this subject is more likely to be delivered in a more deliberate and measured manner, the lecturers are most likely very selective in their words and phrases, and might allow pauses for reflection, both of which which may contribute to the slower speech rate and higher lexical diversity. This might be in contrast to subjects such as Computer Science and Languages and Linguistics, which were ranked fifth and sixth, respectively, in terms of lexical diversity, but have the fastest speech rates. Mathematics and Physics, on the other hand, fall into the slow speech rate and low lexical diversity quarter. We point out such

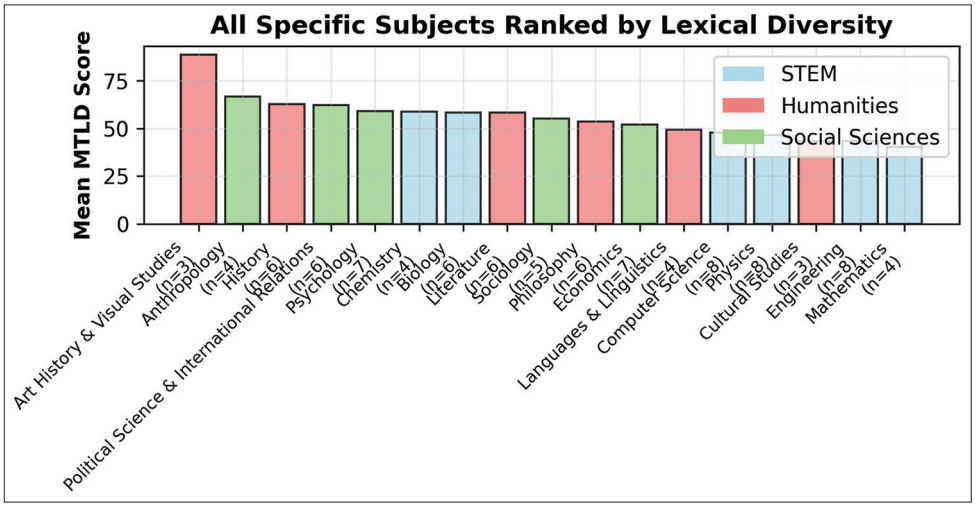


Fig. 3. Ranking of subjects by measure of textual lexical diversity scores.

analysis in the Intended Use Cases section (6.1) and leave it for future work.

Appendix Table 2 (Appendix D) contains further statistics.

4.2.3. ASR quality statistics

The distribution of average confidence scores across the dataset is shown in Fig. 8. This figure illustrates the spread of confidence scores, highlighting the overall quality of the transcriptions. The minimum confidence score is 0.7018, while the mean is 0.8402. It is evident that majority of the transcripts are clustered around the higher end of the confidence scale, indicating a generally high level of transcription accuracy.

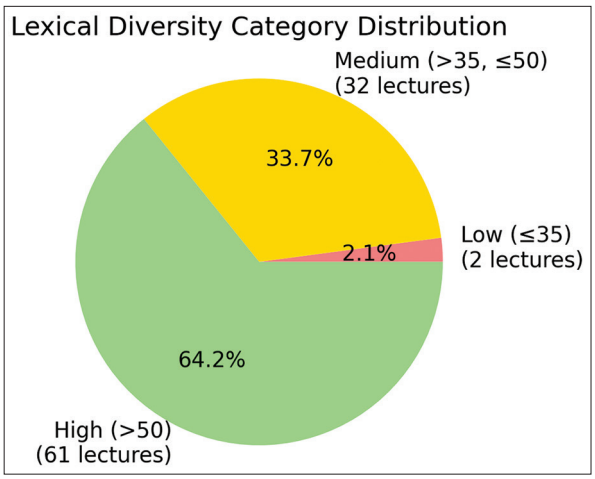


Fig. 4. Categorical breakdown of lectures by measure of textual lexical diversity scores.

As for the subject-specific average confidence scores, Fig. 7 shows the average confidence scores for each subject area. It is evident that most subjects have high average confidence scores, with most falling between 0.8 and 0.9. This indicates that the transcript reliably reflects the spoken content. Furthermore, it supports confidence in evaluations or analyses that will be conducted on the dataset.

Further statistics are available in Appendix Table 3 (Appendix E).

5. TOY A/B DEMONSTRATION

To validate the effectiveness of the dataset in providing statistically significant results on head-to-head comparisons between the techniques, a toy A/B evaluation is conducted using two different LLM models on the task of zero-shot summarization of the transcripts. The two models used were Google's Gemini-1.5-Flash and Gemini-1.5-Flash-8B. This choice was guided by Google's claim that, despite the Gemini-1.5-Flash-8B model being smaller than the Gemini-1.5-Flash model, it is more efficient, faster, and nearly as capable as the larger model[6]. Because of the inherent performance similarity of the two models, distinguishing the superior model is harder, making this an appropriate test of the dataset's ability to detect small performance differences. Fig. 9 illustrates the prompt used for the zero-shot summarization task.

BERTScore [28] was used to evaluate the faithfulness of the

6 https://developers.googleblog.com/en/gemini-15-flash-8b-is-now-generally-available-for-use/

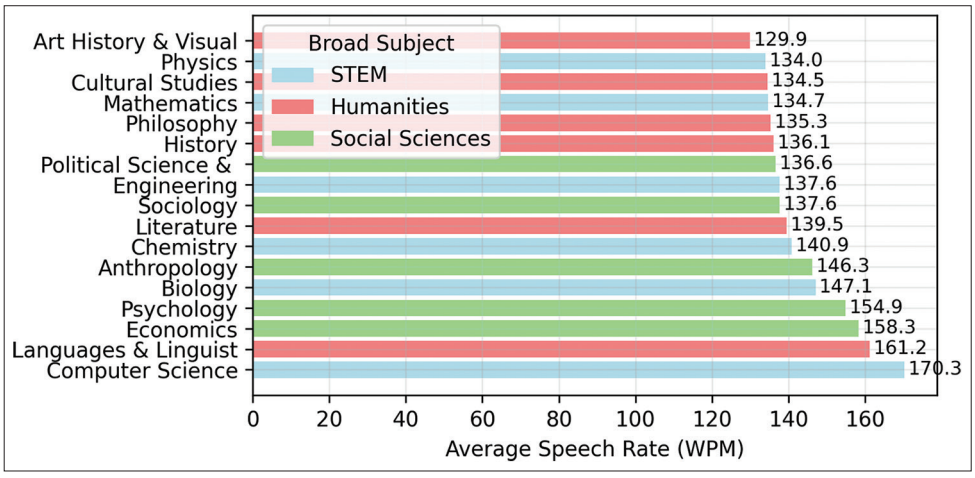


Fig. 5. Average speech rate by subject.

summaries generated by each of the two models.

The results of the A/B evaluation are shown in Fig. 10. As can be seen, the Gemini-1.5-Flash model outperformed the Gemini-1.5-Flash-8B model in terms of the number of wins, with a win-rate of 71.6% (68 wins out of 95 comparisons). The Gemini-1.5-Flash-8B model had a win-rate of 28.4% (27 wins out of 95 comparisons). An effect size of 43.16% was observed between the two models, with 95% confidence intervals for the win-rates being (0.6251, 0.8065) and (0.1935, 0.3749), respectively. The two-tailed binomial test was used to determine the statistical significance of the results with a null hypothesis of the win-rate being 50% (i.e., no difference between the two models). The *P*-value was found to be 0.00003114, which is statistically significant at the 0.05 level. Therefore, the null hypothesis can be rejected, and we conclude that the two models differ significantly in terms of zero-shot summarization task performance.

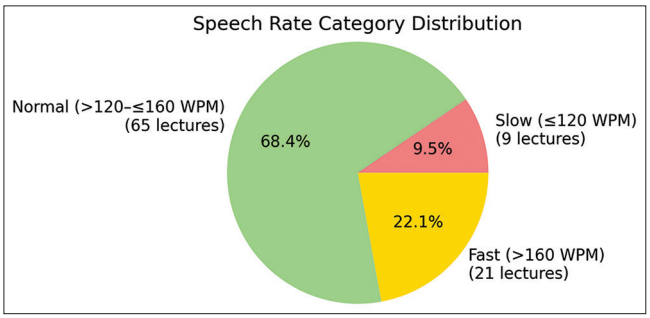


Fig. 6. Categorical breakdown of lectures by speech rate.

This all demonstrates that the dataset is capable of detecting statistically significant differences in performance between different techniques. In this case, far beyond the 20% effect size aimed for. The narrow confidence intervals further confirm the statistical significance of the results.

While BERTScore gave us F1 scores that could be used in a t-test to compare the magnitude of differences in average faithfulness, but instead the win-rate was favored as it looks at the consistency of a model's superiority over the other. We believe that the consistent superiority of a tool across lectures demonstrates more practical educational value than marginal average gains.

6. DISCUSSION

Here, the dataset's practical use cases and limitations are discussed, each of which provides valuable insights into avenues for independent research and future work.

6.1. Intended Use Cases

6.1.1. Evaluation of lecture processing techniques

The primary intended use case for this dataset is the evaluation of different lecture processing techniques in A/B settings. The dataset's diversity should allow for robust comparisons (e.g., testing technique performance across STEM vs. humanities lectures) while revealing the blind spots a technique might have in academic fields. Lecture processing techniques can include summarization, question generation, feedback generation, automated grading, and more.

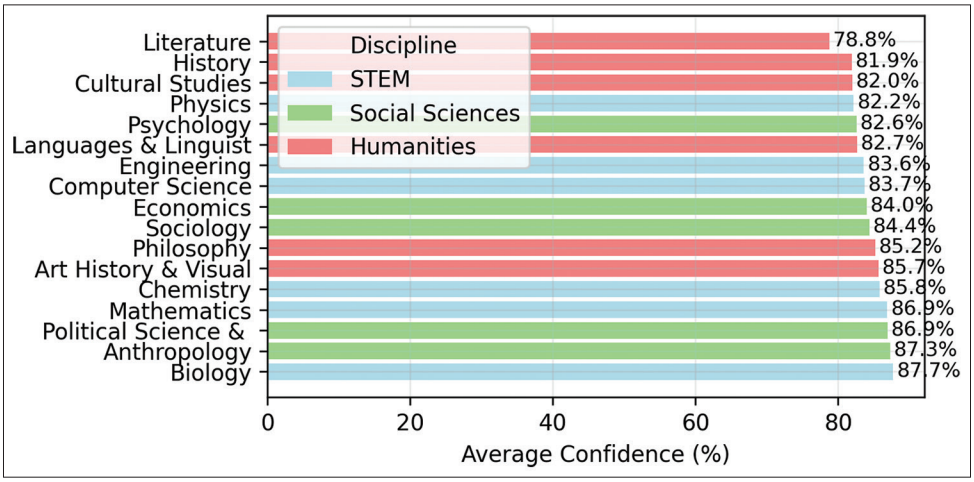


Fig. 7. Average confidence scores by subject.

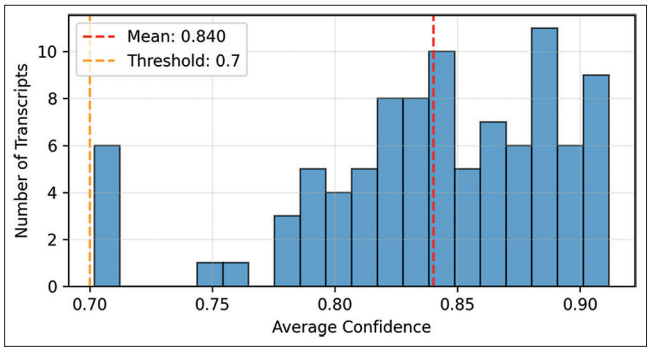


Fig. 8. Distribution of average confidence scores.

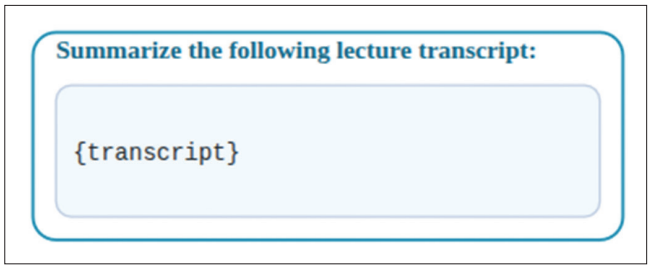


Fig. 9. Prompt template for lecture transcript summarization.

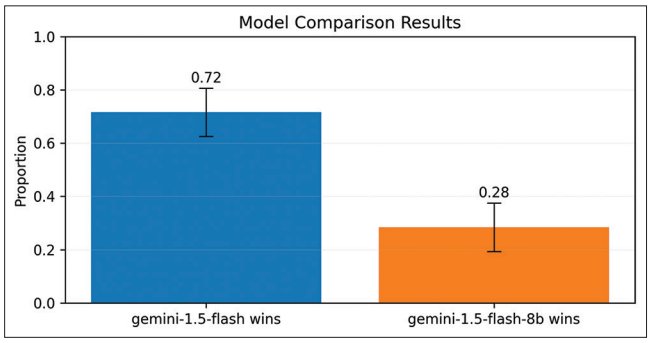


Fig. 10. Results of the toy A/B evaluation between the two LLM models on the task of zero-shot summarization of the transcripts.

6.1.2. Rapid and cheap proof-of-concepts

The dataset can be used for rapid prototyping and proof-of-concept implementations of new lecture processing techniques. Its modest yet statistically significant sample size gives it room for cheap experimentation and testing, especially in the emerging field of LLMs where API costs can add up quickly. All without the hassle of collecting and processing new data.

6.1.3. LLM post-training

The dataset can also be used with some LLM post-training techniques, such as fine-tuning, direct preference optimization, knowledge distillation, and prompt engineering research, to alter the behavior of the LLMs toward a certain task. For example, fine-tuning an LLM to generate realistic transcripts for classroom simulations or using segments of the transcript for few-shot prompts.

6.1.4. Research on educational dialogue analysis

The dataset can also be used for research on educational discourse analysis, such as analyzing the dynamics of teacher-student interactions, including questioning, prompting, and feedback. This can aid in understanding and improving instructional practices in higher education.

6.2. Dataset Limitations

6.2.1. Dataset size

While the current sample size of 95 transcripts is sufficient for most A/B evaluations with a decent effect size, it may not be enough for more nuanced analyses. Future work could benefit from a larger dataset (e.g., $n = 385$ for 10% effect size or even larger) to capture more subtle differences in performance, especially when comparing techniques with similar performance levels. All while ensuring diverse representation across different academic fields and universities. This would allow for more robust statistical analyses and generalizations.

6.2.2. Language and demographic diversity

The dataset currently only focuses on English lectures, primarily from Ivy League universities in North America, which may lead to applications with less generalizability to non-English speaking contexts or different demographics. Future work could benefit from expanding the dataset to include lectures in other languages and from a wider range of universities.

6.2.3. Detailed metadata

The dataset currently lacks detailed metadata about the lectures, such as the course name, professor's background, student demographics, and related attributes. This information could provide valuable context for any lecture processing techniques applied to the transcripts. For example, including the course level (e.g., undergraduate, graduate) would help a lecture rating model understand the expected complexity and depth of the content.

6.2.4. University bias

The dataset contains lectures from Ivy League universities such as Harvard and MIT, which may not always be representative of other universities or institutions. This may introduce a bias in the results of the evaluations. However, testing the techniques against one another will always yield the superior technique regardless of the university bias.

While the limitations might create obstacles for certain research undertakings, they reflect limitations either in terms of time or resources on our part, and deliberate choices to prioritize the A/B testing capabilities. Nonetheless, the dataset should provide a robust foundation for core lecture processing research, with identified gaps that can be addressed in future work by us or the community.

7. CONCLUSIONS AND FUTURE WORK

We present LectBench-95, a dataset of 95 university lecture transcripts from 3 different academic disciplines (STEM, Humanities, and Social Sciences) and 17 different subjects, with a total of 94 hours of transcribed lecture content. The dataset is curated with a focus on diversity (i.e., in the academic subjects, vocabulary used, and lecturer speech rates), quality, and student-lecturer interactivity as demonstrated. In addition, effort has been put into selecting a sample size that would always yield statistically significant results with a solid effect size. It is primarily intended for the evaluation and comparison of different lecture processing techniques (educational natural language processing) in A/B settings, as well as for rapid prototyping and proof-of-concept implementations. The dataset is released under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0) license, and users are encouraged to respect the license and the original content creators. As for future work, any of the gaps identified in the limitations subsection (6.2) suggest complementary datasets that can be built on top of LectBench-95.

REFERENCES

[1] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan … & D. Amodei. Language models are few-shot learners. In: "*Advances in Neural Information Processing Systems*". Curran Associates, Inc., United States, pp. 1877-1901, 2020.

[2] A. I. Open, J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. S. Altman, S. Anadkat, R. Avila, I. Babuschkin …& S. Balaji. "*GPT-4 Technical Report*". Cornell University, United States, 2023.

[3] Y. Hicke, A. Agarwal, Q. Ma and P. Denny. "*AI-TA: Towards an Intelligent Question-Answer Teaching Assistant using Open-Source LLMs*". Cornell University, United States, 2023.

[4] B. Oh, Y. Lee and Y. Kim. Applicability of pretrained language models: Automatic Screening for children's language development level. In: L. Biester, D. Demszky, Z. Jin, M. Sachan, J. Tetreault, S. Wilson, L. Xiao and J. Zhao, editors. "*Proceedings of the Second Workshop on NLP for Positive Impact (NLP4PI)*". Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid), pp. 149-156. 2022.

[5] A. Xenakis, I. Dimos, M. Feidakis, D. Sotiropoulos, K. Kalovrektis and G. Nikolaou. An LLM-based smart repository platform to support educators with computational thinking, AI, and STEM Activities. In: "*Empowering STEM Educators With Digital Tools*". IGI Global Scientific Publishing, United States, pp. 107-136. 2025.

[6] C. Cohn, N. Hutchins, T. Le and G. Biswas. "A chain-of-thought prompting approach with LLMs for evaluating students' formative assessment responses in science". *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 21, pp. 23182-23190, 2024.

[7] D. Demszky and H. Hill. "*The NCTE Transcripts: A Dataset of Elementary Math Classroom Transcripts*". Association for Computational Linguistics, USA, 2023.

[8] T. J. Kane, D. F. McCaffrey, T. Miller and D. O. Staiger. "*Have we Identified Effective Teachers? Validating Measures of Effective Teaching using Random Assignment*". Bill and Melinda Gates Foundation, MET Project, [Research Paper], 2013.

[9] A. Suresh, J. Jacobs, C. Harty, M. Perkoff, J. H. Martin and T. Sumner. "*The TalkMoves Dataset: K-12 Mathematics Lesson Transcripts Annotated for Teacher and Student Discursive Moves*". Cornell University, United States, 2022.

[10] K. Stasaski, K. Kao and M. A. Hearst. CIMA: A large open access dialogue dataset for tutoring. In: J. Burstein, E. Kochmar, C. Leacock, N. Madnani, I. Pilán, H. Yannakoudakis, and T. Zesch, editors. "*Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*". Association for Computational Linguistics, Seattle, WA, USA, pp. 52-64, 2020.

[11] A. K. Rai, S. D. Jaiswal and A. Mukherjee. "A deep dive into the disparity of word error rates across thousands of NPTEL MOOC videos". *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 18, pp. 1302-1314, 2024.

[12] A. Hernandez and S. Yang. "*Multimodal Corpus Analysis of Autoblog 2020: Lecture Videos in Machine Learning*". Springer, Berlin, pp. 262-270. 2021.

[13] A. Hernandez, P. Klumpp, B. Das, A. Maier and S. H. Yang. Autoblog 2021: The importance of language models for spontaneous lecture speech. In: "*Text, Speech, and Dialogue: 25th International Conference, TSD 2022, Brno, Czech Republic, September 6-9, 2022, Proceedings*". Springer-Verlag, Berlin, Heidelberg, pp. 291-300, 2022.

[14] M. Song, I. Aslan, E. Parada-Cabaleiro, Z. Yang, E. André, Y. Yamamoto and B. Schuller. Lecture Video Highlights Detection from Speech. In: "*2024 32nd European Signal Processing Conference (EUSIPCO)*". pp. 361-365. 2024.

[15] T. Lv, L. Cui, M. Vasilijevic and F. Wei. "*VT-SSum: A Benchmark Dataset for Video Transcript Segmentation and Summarization*". arXiv:2106.05606.

[16] J. Wright, M. Liberman, N. Ryant and J. Fiumara. "*Evaluating Speech-to-Text Systems with PennSound*". Cornell University, United States, 2025.

[17] D. S. Singh, A. Gupta, C. V. Jawahar and M. Tapaswi. Unsupervised Audio-Visual Lecture Segmentation. In: "*2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*". pp. 5221-5230, 2023.

[18] H. Wang, F. Yu, X. Shi, Y. Wang, S. Zhang and M. Li. "*SlideSpeech: A Large-Scale Slide-Enriched Audio-Visual Corpus*". Cornell University, United States, 2023.

[19] D. Ďurišková, D. Jurášová, M. Žilinec, E. Šubert and O. Bojar. Khan Academy Corpus: A Multilingual Corpus of Khan Academy Lectures. In: "N. Calzolari, M. Y. Kan, V. Hoste, A. Lenci, S. Sakti

and N. Xue, editors. "*Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*". ELRA and ICCL, Torino, Italia, pp. 9743-9752, 2024.

[20] A. Rousseau, P. Deléglise and Y. Estève. TED-LIUM: An Automatic Speech Recognition Dedicated Corpus. In: Calzolari, K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk and S. Piperidis, editors. "*Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, N". European Language Resources Association (ELRA), Istanbul, Turkey, 2012, pp. 125-129.

[21] F. Hernandez, V. Nguyen, S. Ghannay, N. Tomashenko and Y. Estève. "TED-LIUM 3: Twice as much data and corpus repartition for experiments on speaker adaptation. In: A. Karpov, O. Jokisch and R. Potapova, editors. "*Speech and Computer*". Springer International Publishing, Cham, pp. 198-208, 2018.

[22] D. W. Lee, C. Ahuja, P. P. Liang, S. Natu and L. P. Morency. "Lecture presentations multimodal dataset: Towards understanding multimodality in educational videos. In: "*2023 IEEE/CVF International Conference on Computer Vision (ICCV)*". IEEE, United States, pp. 20030-20041, 2023.

[23] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey and I. Sutskever. "*Robust Speech Recognition Via Large-Scale Weak Supervision*". Cornell University, United States, 2022.

[24] C. Graham and N. Roll. "Evaluating openAI's whisper ASR: Performance analysis across diverse accents and speaker traits". *JASA Express Letters*, vol. 4, no. 2, p. 025206, 2024.

[25] P. M. McCarthy and S. Jarvis. "MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment". *Behavior Research Methods*, vol. 42, no. 2, pp. 381-392, 2010.

[26] P. Wingrove. "How suitable are TED talks for academic listening?" *Journal of English for Academic Purposes*, vol. 30, pp. 79-95, 2017,

[27] W. Xiao and S. Sun. "Dynamic lexical features of PhD theses across disciplines: A text mining approach". *Journal of Quantitative Linguistics*, vol. 27, no. 2, pp. 114-133, 2020.

[28] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger and Y. Artzi. "*BERTScore: Evaluating Text Generation with BERT*". Cornell University, United States, 2019.

[29] M. Laouenan, P. Bhargava, J. B. Eyméoud, O. Gergaud, G. Plique and E. Wasmer. "A cross-verified database of notable people, 3500BC-2018AD". *Scientific Data*, vol. 9, no. 1, p. 290, 2022.

APPENDIX

A. Lexical Diversity Statistics

We report a more detailed lexical diversity descriptive statistics in Appendix Table 1. Including MTLD, TTR, and vocabulary statistics.

B. Speech and syllable rate statistics

We provide speech-rate (WPM) and syllable-rate (SPM) distributions in Appendix Table 2.

C. Transcript confidence statistics

Descriptive statistics for the Confidence scores (from Whisper's average token-level probabilities) are provided in Appendix Table 3.

D. Google Sheet Screenshot

Appendix Fig. 1. shows the Google sheet used to determine lecture interactivity via the storyboard images. As can be seen, 4 thumbnails are retrieved, each of which showcase different time points of the video lecture.

E. Anonymization and ethical considerations

Attempts were made to anonymize each segment by removing any Personally Identifiable Information (PII). The plan was to anonymize the following:

- Names: Replaced with the term "Person X" (e.g., "James" becomes "Person 1").
- Organizations: Replaced with the term "[REDACTED ORG]."
- Emails: Replaced with the term "[REDACTED EMAIL]."

This process was intended to ensure the privacy of any individuals mentioned in the transcripts. The Name and Organization replacements were done through Named Entity Recognition, specifically using the spaCy library[1]. The email replacement was done using a simple regex pattern that matches the standard email format.

The anonymization process, however, caused historical figures' names like "Henry Ford" to be replaced with "Person X." This is a problem as it removes crucial context from the transcripts. The magnitude of the impact this may have on the results is still not understood. To solve this problem, cross-checking with historical databases of notable people like the one by [29] was attempted, but was not feasible as sometimes only the last names were used when referencing such people (e.g., "James Madison" might be referred to by "Madison"), which could be common first names of other people. An attempt was also made to use the MediaWiki API to cross-check the names and used a scoring approach to only single out very historical and very popular people of a certain type but this method proved problematic as well. Another approach considered was to use the most popular baby names list of the United States for the last 5 years and use them as a roster list to anonymize the names, but to no avail. Thus, informed by the fact that all the lectures are public-domain YouTube videos, it was concluded that it may not be strictly necessary to anonymize the transcripts and therefore the idea of anonymization was abandoned altogether. Users are still encouraged to anonymize the dataset if their use case requires it.

1 https://spacy.io/—Specifically the en_core_web_lg model.

APPENDIX TABLE 1: Summary of measure of textual lexical diversity (MTLD), TTR, and vocabulary statistics

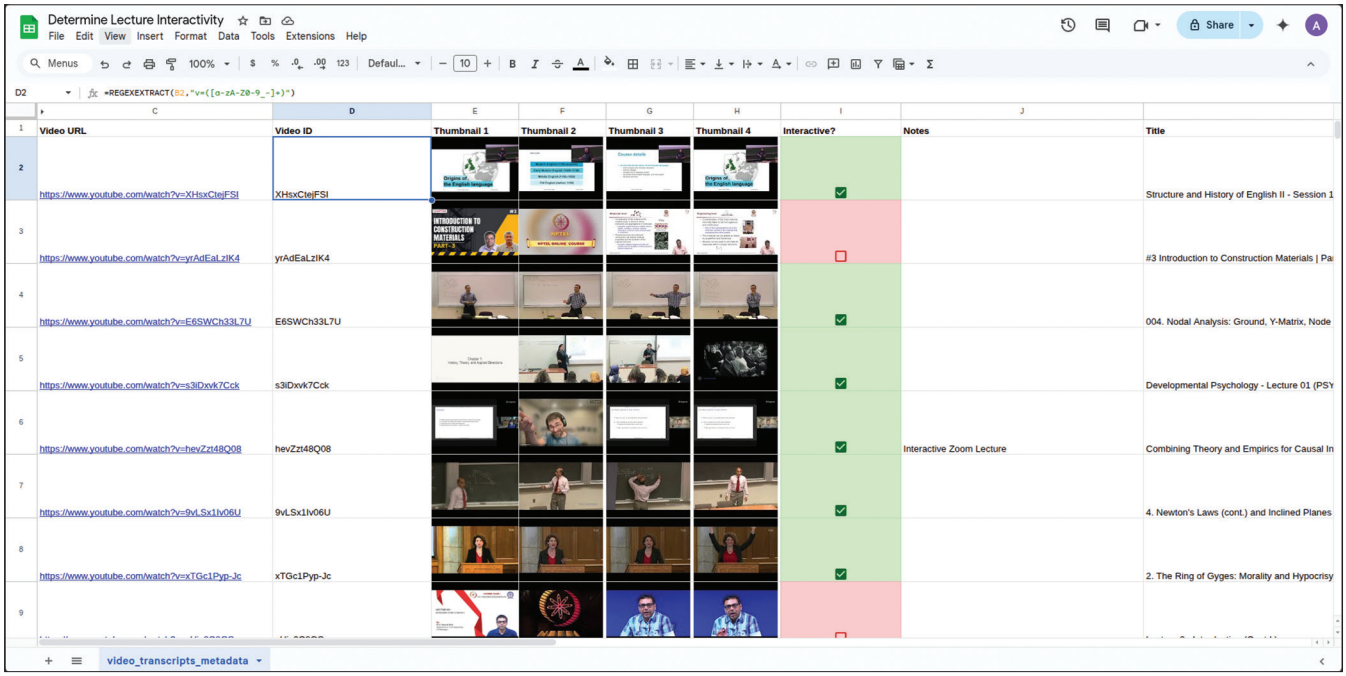
MTLD (Lexical diversity)	
Mean MTLD	54.49
Median MTLD	52.47
Standard deviation	11.56
Min MTLD	32.28
Max MTLD	90.35
25th percentile	46.42
75th percentile	60.34
Type-token ratio (TTR)	
Mean TTR	0.1578
Median TTR	0.1530
Range	0.0741–0.2822
Vocabulary statistics	
Average vocabulary size	1292 unique words
Vocabulary range	513–2381 unique words
Average hapax ratio	0.079
Average sophistication ratio	0.165

APPENDIX TABLE 2: Speech and syllable rate statistics

Speech rate (Words per minute [WPM])	
Mean speech rate	144.3 WPM
Median speech rate	141.7 WPM
Standard deviation	21.0 WPM
Min speech rate	96.9 WPM
Max speech rate	201.2 WPM
25th percentile	130.6 WPM
75th percentile	156.6 WPM
Syllable rate (Syllables per minute [SPM])	
Mean syllable rate	203.6 SPM
Median syllable rate	205.8 SPM
Range	127.6–266.6 SPM

APPENDIX TABLE 3: Transcript confidence statistics

Total segments	
Total segments across all transcripts	76,329
Overall confidence statistics	
Mean average confidence	0.8402
Median average confidence	0.8461
Standard deviation	0.0508
Min average confidence	0.7018
Max average confidence	0.9118
Segment-level statistics	
Mean segment confidence	0.8338
Median segment confidence	0.8508
Segments below 0.7	4,808 (6.30%)
Segments below 0.5	142 (0.19%)



Appendix Fig. 1. Screenshot of the Google Sheet interface for assessing video interactivity via storyboard images.