Advancing Learnersourced Caption Editing for Video-Based STEM Education

ANONYMOUS AUTHOR(S)

Captions play a major role in making educational videos accessible to all and are known to benefit a wide range of students. However, many educational videos, especially STEM-related videos, either do not have captions or have inaccurate captions. Prior work has shown the benefits of using crowdsourcing to obtain accurate captions in a cost-efficient way. However, there is a lack of understanding of how learners edit captions of STEM-related videos either individually or collaboratively. In this work, we conducted a user study where 58 learners (in a STEM course of 387 students) participated in the editing of captions in 89 lecture videos that were generated by Automatic Speech Recognition (ASR) technologies. More specifically, for each video, different learners conducted two rounds of editing. Eighteen students participated in follow-up interviews and shared their editing experiences. Based on editing logs, we created a taxonomy of errors in STEM-related captions (e.g., STEM, Non-STEM, Equations). From the interviews, we identified individual and collaborative error editing strategies during the different stages of the "Find-Fix-Verify" crowdsourcing model (e.g. prioritizing errors in STEM words). To better support the learnersourced caption editing for STEM-related videos, we gathered participants' suggestions for better system and policy supports (e.g., help identify Non-STEM errors) that may improve their caption editing experience. We then further evaluated the feasibility of applying machine learning models to provide assistance to editors. Our work provides theoretical implications towards the "Find-Fix-Verify" model for STEM video caption editing, and both system design suggestions and practical implications for advancing learnersourced STEM-related captioning.

CCS Concepts: • **Computer systems organization** → **Embedded systems**; *Redundancy*; Robotics; • **Networks** → Network reliability.

Additional Key Words and Phrases: learnersourcing, captions, STEM education, video-based learning

ACM Reference Format:
Anonymous Author(s). 2021. Advancing Learnersourced Caption Editing for Video-Based STEM Education . In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 29 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

As the popularity of video-based learning (e.g., MOOCs) continues to grow, it is crucial to provide accessible online videos to all, including but not limited to, users who are non-native speakers and students with disabilities. The COVID-19 pandemic has necessitated a rapid shift to the use of online educational videos, increasing the urgency of providing accessible online educational videos [34]. Prior research has shown that accurate captions play an important role in making lecture videos accessible [15], especially for students who are Deaf or Hard of Hearing (DHH). Further, following the principles of Universal Design Learning (UDL), captions are found to be beneficial to a wide range of students [12] that also includes students with other disabilities (e.g., ADHD, dyslexia) and students who prefer to learn by reading or searching transcriptions.

However, creating captions is challenging. Commercial captioning services have a slow turnaround (requiring a business day or longer) and are expensive (approximately $1 per audio-minute)[48], while centralized university services have capacity limits which can only take on a small fraction of the total number of content hours generated by a university each each day. Alternatively, captions generated by Automatic Speech Recognition (ASR) algorithms are cheaper (e.g., Microsoft and Google cloud ASR services are approximately $1 per audio-hour) but have been shown to be too inaccurate to be used exclusively for learning with educational videos[37]. Generating domain-specific captions, e.g. captions for Science, Technology, Engineering, and Math (STEM) courses, presents additional challenges due to the use of scientific nomenclature, slides with significant graphical content on slides, and diverse instructors with accents [1, 38].

Prior work has achieved significant success in using crowdsourcing and learnersourcing techniques to edit ASR generated captions to lower the cost of obtaining high-quality captions [16, 26]. However, learnersourcing for STEM-related videos is not yet investigated. For example, what are the challenges of captioning of STEM lecture videos for individual learners and for groups of learners when they collaboratively edit captions? Gaining an empirical understanding and addressing the above questions can advance algorithms that improve captioning services for STEM and suggest opportunities for human-in-the loop captioning.

In this paper, we deployed a system for learnersourced caption editing at a large enrollment ($N = 387$) course at a US university where 58 students edited 89 STEM lecture videos. The captions were edited by two rounds of editors. Follow-up interviews were conducted with 18 of those editors to further understand the challenges, strategies used, and need for better support in editing STEM captions. We conducted qualitative analysis on the interview results, performed quantitative system log analyses for understanding editing behavior specific to STEM captions, and examined the results using the widely-adopted "Find-Fix-Verify" crowdsourcing paradigm [9].

Our findings make the following contributions to the HCI and CSCW communities. First, we developed a taxonomy for STEM-based caption edits, which allowed us to gain new and empirical understandings of individual and collaborative editing behavior within different categories of STEM edits. Second, we identified various strategies for individual and collaborative caption editing specifically in the context of STEM video-based learning. These strategies informed future design of captioning systems that can improve the efficiency of captioning services. Third, inspired by learners' challenges and suggestions, we evaluated the feasibility of applying machine learning algorithms to better support STEM-based captioning editing, which further contributed to the literature of human-machine collaboration. Additionally, we propose new designs and discuss theoretical and practical implications of learner-sourced captioning service for STEM-based learning.

2 RELATED WORKS

2.1 Video Captions Video-based Learning

Research has shown that to make the content of educational videos accessible to the widest audience possible, it is important to improve the readability of the text and captions of the videos [15]. Students who are Deaf or Hard of Hearing rely on captions for access to video content. Further, following the principles of Universal Design Learning (UDL), captions are found to be beneficial to a wide range of students [12]. Yet many videos remain uncaptioned or have machine-generated captions with high error rates [43].

Compared to general videos, captioning educational videos correctly requires more domain knowledge. For example, automatically generating video captions for STEM courses presents unique challenges due to substantial scientific nomenclature, technical terminology, and slides with significant graphical content [38]. At the same time captions are important for understanding such technical terms in STEM videos. For example, without access to the exact symbol or equation mentioned by the instructor, the lecture content could be hard to understand. Therefore, it is necessary to especially understand how should domain-related terms be captioned for video-based learning. By gaining such understandings, implications could be proposed to improve caption quality for educational videos.

2.2 Crowd-sourcing Video Captions

Open source models (e.g. Mozilla DeepSpeech) and commercial cloud-based Automatic Speech Recognition services exist today including cloud services by Microsoft, Amazon and Google and start-up companies focused on automatic speech to text services (e.g., Otter.ai) can, with minimal work, be used to create captions from the video's audio stream. They provide free or low-cost captioning but do not meet the ADA accuracy goal. For example a 2020 analysis found that though Google's API outperformed IBM Watson and Facebook's Wit.ai ASR, but recorded an average word error rates (WER) of approximately 9% [23]. For comparison, professional human transcription with a word accuracy of 99% is considered the best-practice and sufficient to meet the compliance requirements of the American with Disabilities Act (ADA) for public-shared video content. Practically a WER of 9% is significant barrier to efficient and accurate learning; a student would be attempting to learn with mistakes and inaccurate statements that occur in most sentences. Further, speaker accents, sub-optimal audio, and mis-identification of domain-specific and topic-specific words by ASR (e.g. "bite" vs "byte"), suggests 9% is a *lower-bound* when ASR is used in educational settings.

Researchers have been studying crowd-sourcing video caption systems that harness both low-cost automatic generated caption and human intelligence in video caption creation. For example, Huang etc. leveraged complementary contributions of different workers to design, implement and evaluate an efficient crowd-sourcing system for video captions–BandCaption [26]. The system combines automatic speech recognition with input from crowd workers to provide a cost-efficient captioning solution for online videos. They considered four stakeholder groups as a source of crowd workers: individuals with hearing impairments, second-language speakers with low proficiency, second-language speakers with high proficiency, and native speakers. Such a system enables crowd workers who have different needs and strengths to accomplish micro-tasks and make complementary contributions.

Researchers found through case studies that editors correct errors to improve the readability of the lecture transcripts for student use and enhance the accuracy of future speech recognition models [38]. Nevertheless, error correction is still the most time-consuming task to create lecture transcripts [38].

Researchers also explored editing of auto-generated subtitles of online educational videos by instructors prior to publishing [47]. However, in our literature review we found a lack of quantitative understanding of the error correction behavior by humans in crowd-sourced context on automatically generated transcripts. By better understanding human editing of machine-generated information, we are able to propose design insights on new ways for future systems. This interactive process between machine learning and human is called 'Interactive Machine Learning' [4].

2.3 From Crowdsourcing to Learnersourced Captioning

Research shows collaboration between different crowd workers, either synchronously or asynchronously, is necessary to enable complex and sustainable crowdsourcing systems [28]. HCI and CSCW community also attempts to better address responsibility concern in cooperative Work when designing crowdsourcing systems. Retelny et al. proposed "flash teams", a framework for dynamically assembling and managing paid experts from the crowd by highlighting "who is working together and who is responsible for which tasks"[40]. They implemented such framework into an app, Foundry, an end-user authoring platform and runtime manager.

The "Find-Fix-Verify" model has been shown effective in various collaborative crowdsourcing systems, such as crowdsourced writing [9] and micro-task assignment [10]. The Find stage, asks workers to identify patches that need more attention; the Fix stage recruits workers to revise an identified patch; the Verify stage performs quality control on revisions by recruiting workers to vote on others' work. Researchers also proposed and implemented the "Mark-Edit-Approve" model that allows subsequent workers to edit earlier workers' edits to conduct crowd-captioning quality control [26]. Commercial online learning platforms such as Coursera has Coursera's Global Translator Community (GTC) program designed to greatly expand the number of courses offering high-quality subtitle translations through edit-review stages [13].

There exist educational systems, such as ClassTranscribe [39] and ICS Videos [18], that can generate captions for lecture videos in a learner-sourcing fashion. Previous studies [6, 14, 18] have shown that learner-sourcing tools are effective and efficient for captioning lecture videos and have considerable value in educational practice. Besides, the generated transcription can be used to search and index lecture videos. Researchers have found that searching transcriptions can predict improved exam performance that is statistically significant [6, 8]. Previous research has also shown that involving students in fixing captions in foreign language educational videos does not impair learning and also helps reduce errors in the captions [16]. However, despite the fact that these learnersourcing tools could be beneficial for educational purposes, none of these works studied how learners worked together in the learnersourcing experience. To the best of our knowledge, this is the first study to focus on students' behaviors and attitudes towards collaborating with each other to help with improving the captions of lecture videos in a STEM online class.

Specifically, in this paper, we addressed the following research questions:

RQ1: How do individual learners make edits to crowd-sourced captions?

RQ2: How do learners collaborate with other learners in crowd-sourced caption editing?

RQ3: How can the system better support learners to conduct STEM-caption edits?

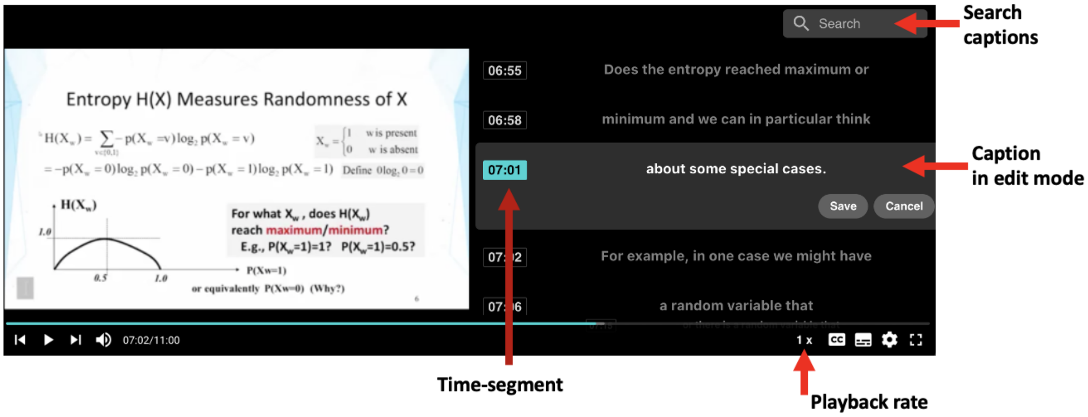
3 METHOD

In this section, we describe our study design and methodology. We first describe the system used for editing ASR generated captions. Next, we describe the caption editing activity where 58 learners in a STEM course participated in editing captions of 89 out of 93 lecture videos with two editing rounds. Follow-up interviews were conducted with 18 student editors to understand their editing experience and behavior. Each edit was logged for further analysis.

3.1 ClassTranscribe–a Video-based Learning System

ClassTranscribe [7, 8] is an open-source web platform for delivery of educational online lecture videos. In this system, the lecture audio is initially transcribed using an automatic speech-to-text cloud service (Microsoft Azure Cognitive Services Speech-To-Text) at a cost of approximately $1 per audio-hour. The captions are indexed to enable keyword-based search. The frontend technology of the ClassTranscribe system is comprised of javascript libraries (including react and redux) and json-based web API, media file server and task engine. The server is comprised of a Postgres SQL database, RabbitMQ message bus with application code written in python and C#. The system is built and deployed as a set of Docker containers on a dedicated Linux virtual machine. The database schema and design choices are published in [7]. Source code is available at github.com/ClassTranscribe.

Figure 1 shows the system user interface for editing captions. The lecture captions are displayed on the right side of the video. Each caption is annotated with the corresponding video time-segment. By clicking on the time-segment, users can jump to that video moment. Clicking on a caption opens up the caption in the "edit mode". After editing the caption, users can click on the "Save" button or hit "return/enter" on the keyboard to save their edits. Any edits made are instantaneously reflected on the interface. Users can also search for keywords within captions across all course lecture videos. Captions can also be turned off. Other useful features include adjusting the playback rate, pausing/playing video, adjusting the video progress bar (seeking) etc. All user activities on the system, including searching, watching a video, seeking, editing a caption are logged in the SQL database for later analysis.



Entropy H(X) Measures Randomness of X

$H(X_w) = \sum_{v\in\{0,1\}} -p(X_w = v)\log_2 p(X_w = v)$

$= -p(X_w = 0)\log_2 p(X_w = 0) - p(X_w = 1)\log_2 p(X_w = 1)$

$X_w = \begin{cases} 1 & w \text{ is present} \\ 0 & w \text{ is absent} \end{cases}$ Define $0\log_2 0 = 0$

For what X_w, does $H(X_w)$ reach maximum/minimum?
E.g., $P(X_w=1)=1$? $P(X_w=1)=0.5$?

06:55 Does the entropy reached maximum or

06:58 minimum and we can in particular think

07:01 about some special cases.

07:2 For example, in one case we might have

07:6 a random variable that

Fig. 1. Interface for Editing Captions

3.2 Caption Editing Activity Design

In Fall 2020, the lecture videos of an online senior-level computer science course on Text Mining and Analytics at a large public university in the US were uploaded onto the ClassTranscribe system. Students in the course were provided with an opportunity to participate in an extra-credit activity to fix errors in captions in lecture videos. There were a total of 93 lecture videos in the course. Each video was on average 12 minutes long. Each lecture video transcript has about 1500 words.

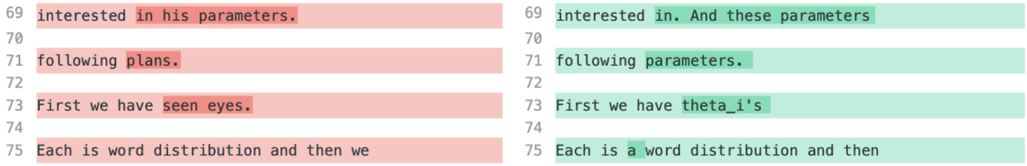
The caption editing activity was divided into two tasks. The task of *Editor One* was to take a first pass at correcting errors in captions. After *Editor One* completed their task, *Editor Two* then reviewed the captions to fix any remaining errors. In this way, *Editor One* and *Editor Two* sequentially edited the captions to fix the errors. Students knew about who signed up to edit the lecture videos as the two editor types, but they were unaware about the exact edits made on the videos by other editors.

The system described above was released to students towards the end of the course only for this activity. Prior to that, students used Coursera [1] to watch the online course lectures. Interested students were provided with an opportunity to sign up as *Editor One* and *Editor Two* for one lecture video each to get 1% extra-credit. To enable participation from many students, each student could sign up for a maximum two lecture videos each as *Editor One* and *Editor Two* for a total of 2% extra-credit. Students were given two weeks to complete the *Editor One* task first. The following two

[1] https://www.coursera.org/

weeks were for completing the *Editor Two* task. Other extra-edit activities were also released simultaneously to give all students a fair chance at receiving extra credits. Besides, we did not set any minimum number of edits to get the extra credit. Students were completely free on deciding what and where they wanted to edit and how many edits they wanted to do.

For every caption edit made on the system by editors, the log captured who made the edit, the time of edit, the caption *before* and *after* the edit, the corresponding lecture video time-segment, and the lecture name. Figure 2 shows four captions *before* (left) and *after* (right) edits. The edited words are highlighted. As can be seen, each caption is not necessarily a complete sentence because captions are segmented based on the corresponding video time-segments. Further, multiple words could be edited and logged within a single caption-level edit. 58 editors edited 89 out of 93 lecture videos. Further details about the log statistics are presented in the findings throughout the paper.



```
69 interested in his parameters.                 69 interested in. And these parameters
70                                               70
71 following plans.                              71 following parameters.
72                                               72
73 First we have seen eyes.                      73 First we have theta_i's
74                                               74
75 Each is word distribution and then we         75 Each is a word distribution and then
```

Fig. 2. Sample captions *before* and *after* an edit

3.3 Follow-up Interview

In Spring 2021, we recruited students who had previously edited captions to participate in an online semi-structured interview. All interviews were conducted on Zoom and we recorded the interview process with consent from participants. Our study was approved by the University IRB.

We were able to recruit 18 participants via email. Each participant was paid at $20/hour (pro-rated). and interviews were approximately 45 minutes on average. Table 1 summarizes the demographics of the interviewees. None of the interview participants reported having any physical or mental chronic conditions that would prevent them from understanding the speech in lecture videos.

The major questions asked in the interview are as follows: why and how participants use captions for learning with educational videos; provide examples of different errors they noticed in captions and explain how they impact their learning; explain how and why they edited specific errors in captions; if/why they chose to ignore certain errors; explain the difference between being *Editor One* and *Editor Two* in terms of their behavior, experience, effort or process of editing; explain the impact of the editing activity on their learning; explain their motivation to make edits; and provide suggestions to improve the system to better support caption editing and improve their overall experience. To facilitate the discussion related to their editing process and experience, we showed them their editing logs (2) comparing captions *before* and *after* their edits and probed them to explain any interesting or important edits.

We referred to the previous paper to first established properties of what participants said without relying on existing theories (open coding) and proceeded to identify relationships among the codes (axial coding) and conducted a thematic analysis on the interview data in a similar process described in [21, 22]. First, two authors familiarized themselves with the data by reading the transcripts carefully. Two authors then performed open coding with four participants' transcripts independently and then met to discuss and compare their codes. Their initial inter-rater reliability (observed proportionate agreement) was 78%. They then discussed discrepancies and revised and expanded the existing categories until they reached an agreement of 100%. After that, the first author coded the remaining data through an iterative process, in which she met with the second author regularly to discuss the codes and iterate on the findings.

Table 1. Interviewee demographics. There were eight Females (F), nine Males (M) and one who preferred not to disclose their gender. Fifteen participants identified as Asian or Asian American (A), two as White (W), and one as Black (B). Thirteen participants were from Computer Science (CS). Other majors included Computer Engineering (CE), Aerospace Engineering (AE), Cognitive Psychology (CP), Civil and Environmental Engineering (CEE), and Computer Science Statistics. Ten participants were working professionals in the Data Science graduate degree program (DSG), six other graduate students (G) and two undergraduates (UG). Thirteen participants were non-native English speakers but reported no problems in understanding or speaking English (NNV2) and five native English speakers (NV). Three participants performed some voluntary edits without officially signing up in the caption editing activity.

PID	Gender	Age	Race	Major	Program	English Proficiency	Voluntary Edits
P1	F	18-24	A	CS & Stats	UG	NNV2	No
P2	M	35-44	W	CS	DSG	NV	Yes
P3	F	18-24	A	CS	DSG	NNV2	No
P4	M	25-34	A	CS	DSG	NV	No
P5	F	18-24	A	CP	UG	NNV2	No
P6	F	18-24	A	CS	G	NNV2	No
P7	F	18-24	A	CS	G	NNV2	No
P8	M	35-44	A	CS	DSG	NNV2	Yes
P9	M	25-34	A	CS	DSG	NV	No
P10	M	18-24	A	CS	UG	NNV2	No
P11	F	18-24	A	CE	UG	NNV2	No
P12	M	18-24	A	CEE	UG	NNV2	No
P13	M	25-34	A	CS	DSG	NNV2	No
P14	F	18-24	A	CS	DSG	NNV2	No
P15	F	18-24	A	AE	UG	NV	No
P16	-	25-34	A	CS	DSG	NNV2	No
P17	M	35-44	W	CS	DSG	NNV2	Yes
P18	M	25-34	B	CS	DSG	NV	No

4 INDIVIDUAL LEARNERS' CAPTION EDITING BEHAVIOR (RQ1)

We analyzed 10, 378 rows of word-level edits generated by 58 learners (editors) with caption editing activity time period and 18 follow-up interviews to answer **RQ1: How do individual learners make edits to crowd-sourced captions?** By coding the editing log by edit type, we generated deeper insights on different types of edits made. As described in the section 3.3 section, during the interviews, interviewers went through each participants' editing log that highlighted captions *before* and *after* edits. This allowed interviewers to explore and understand the editors' editing goal, editing process and editing strategies using explicit editing examples. We found that the participants implicitly followed the three stages in the 'Find - Fix - Verify" model during the caption editing process. Thus, we present our findings based on those three stages.

4.1 A Taxonomy of STEM Caption Edits

To better understand what types of edits were made by the editors, two authors manually coded on word-level edit log by edit type. We identified two main categories of edits- **STEM Edits, Non-STEM Edits, Typos**; within STEM edits, we identified two subcategories - Symbol(Abbreviation) and Non-Symbol edits; within Symbol edits, we identified

Equation, Abbreviation and Other Edits. The frequency of different types of edits was determined using the server database log of user actions.

4.1.1 *Manual Coding Process on Different Types of Edits.*

Editing behaviors included removing, substituting, and adding words. Since the purpose of our study is to identify which words are often mis-transcribed, we only focus on coding added or substituted words. Multiple edits may occur in one caption, so we transformed our editing log from caption-level to word-level. The word-level edit log generated by 58 participants has 10, 378 word-level edits from 8, 826 caption-level edits, which includes 1, 904 unique words. To better understanding what kinds of edits were made by users, two researchers first drafted a code-book on edit type based on different types of errors observed. Edits: **STEM; Non-STEM; Typo**. **STEM Edits**: Symbol; Non-Symbol. Symbol Edits: Equation, Abbreviation, and Other edits. **Typos** here refers to the definition 'an unintentional error that happens when you accidentally hit the wrong key on a keyboard'. Typo was added due to obvious spelling errors noticed while reading captions; it does not indicate a mismatch between audio content and captions. The first-round inter-rater agreement was 76%(Edits)-94%(STEM Edits)-96%(Symbol Edits). The two authors discussed finalizing all disagreed cases one by one by closely reading "before" and "after" edited captions in sentences, referring to original video clips. Each types' examples were provided in the following section together with reasons for making such edits.

A word could be in multiple categories/subcategories, based on the context and whether it formed a STEM phrase. For example "Map" in the context of *Map Reduce algorithm* is labelled as STEM edit, whereas in the context of *Map out* is labelled as Non-STEM edit. There were also some edits (less than ten) where the new word was spelled correctly but were not what the lecturer said; these were still labelled in the same way. This is because we wish to study learner edited captions and not modify them based on our judgement. Punctuation added at the start or end of a word were counted as separate words and labelled based on their context e.g., "(" in the edit, *x+½ ⟶ (x+1)/2* was categorized as an Equation edit, whereas *TF term frequency ⟶ TF (term frequency)* is STEM.

4.1.2 *Reason of Utilizing Manual Coding.*

In previous work, researchers computed the domain specificity values for all the canonical words to identify domain-specific words [36]. They also removed filler utterances (e.g., "um, ahh, yeah"), conversational contractions (e.g., "wanna, gonna, gotta") functional words, and punctuation from the keyword list. Additionally, a large number of speech recognition algorithms were evaluated using word error rate (WER), which ignored punctuation errors [17]. In our editing log, we identified set of word-level edits features that automatic labeling fail to apply – punctuation, numbers and capitalization. Given that such edits took up 20% of total edits, we found is necessary to utilize manual coding.

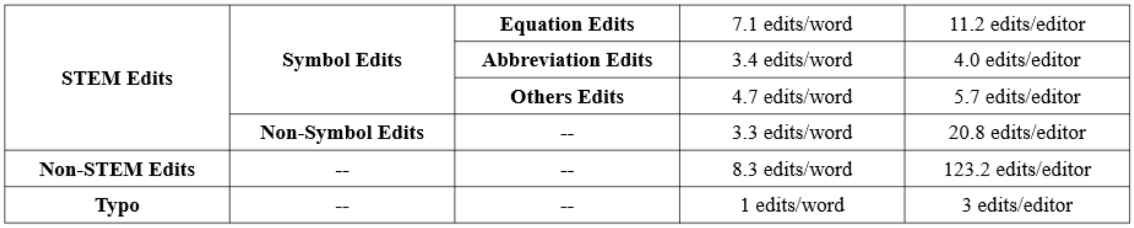
Punctuation. There were 984 punctuation edits (10% of total edits), such as "," was edited more than 500 times to indicate a pause with a sentence, "*"*" was edited more than 50 times to quote to certain information from slides (*general terms or functional terms, the ⟶ general terms or functional terms, like "the", "a"*). Some edits had specific mathematical meaning, "*)*" was edited 10 times; "-" was edited 9 times; "*(*" was edited 5 times; "*]*" was edited 3 times ; "*[*" was edited 3 times.

Numbers and Greek letters. There were 432 number edits (4% of total edits). Of these 111 edits contained Arabic numerals, such as *1*. In 92 cases, numbers were written in text format, e.g., *one*. 95% of these cases were direct conversion between numerical format "1" and textual format "one". There were 228 cases that contained Greek names spelled-out in text format, (e.g., *theta* 130 times). Other edits included *beta lambda alpha gamma delta mu pi*. There was only one example where the participant entered a Greek symbol character, "π".

Capitalization. There were 636 edits (6%of total edits) that included uppercase in edited words. Within all capitalization edits, there were 231 cases were primarily referring to algorithms. There were around 50% cases when before edits cases were misspelled, for example, *analysis, often called PLA.* → *analysis, often called the PLSA.*. For the other 50%, the before edits words were spelled correctly but miscommunicate information, for example, *So do you use map or G map?* → *So do you use MAP or gMAP?.*

4.1.3 Frequency of Different Types of Edits per Editor. In total, 58 participants revised in total 10, 378 edits (Mean = 178.9, SD = 131.2). In Figure 3, we printed frequency of edits per editor (learner) on the last column. As shown in Figure 4, word-level **STEM Edits** were of 30 − 40 percentage. Within **STEM Edits**, a large percentage were Non-Symbol Edits and Symbol Equation Edits. There were eight participants that made far fewer total edits, in which five of them were volunteers that did not sign-up for editing activity. More editing behavior of Volunteers is presented under RQ2.

There were 7, 143 **Non-STEM Edits** (Mean = 123.2, SD = 151.8), 3, 061 **STEM Edits**(Mean = 52.8, SD = 41.8), and 174 Typos. Using a Wilcoxon test, there were significant more **Non-STEM Edits** than **STEM Edits** (W = 2333.5, $p < 0.001$). With-in **STEM Edits**, there were 1, 853 Symbol Edits (Mean=31.9, SD = 22.1), 1, 209 Non-Symbol Edits. There were 647 Equation Edits, 232 Abbreviation Edits, and 330 Other Edits. Using a Wilcoxon test there were significantly more Equation Edits than Abbreviation Edits (W = 2,071, $p < 0.03$), Other Symbol Edits (W = 1,119.5, $p < 0.001$), and Non-Symbol Edits (W = 432.5, $p < 0.001$).



STEM Edits	**Symbol Edits**	**Equation Edits**	7.1 edits/word	11.2 edits/editor
		Abbreviation Edits	3.4 edits/word	4.0 edits/editor
		Others Edits	4.7 edits/word	5.7 edits/editor
	Non-Symbol Edits	--	3.3 edits/word	20.8 edits/editor
Non-STEM Edits	--	--	8.3 edits/word	123.2 edits/editor
Typo	--	--	1 edits/word	3 edits/editor

Fig. 3. Different Types of Edits and Edit Frequency (second last column: by unique word; last column by editor/learner)

4.1.4 Frequency of Making Different Types of Edits per Unique Word. The 10,378 edits include 1,904 unique words. Such results indicate that there were repetitive edited word. In Figure 3, we printed frequency of edits per unique word on the second last column. Non-STEM edits and Equation Edits within **STEM Edits** were most likely to include repetitive edits.

High frequent **STEM Edits** were mainly consisted of Symbol Edits. Symbol Edits included 207 unique words and were in total edited 1370 times (Mean = 6.8. SD = 13.2 per unique word). Within Symbol Edits, there were 27 unique words were edited more than ten times. A majority of them, 62.9%, were Equation Edits. The word *theta* was edited most frequently (130 times) , such as from *So P of Cedar one is the probability of* → *So P of theta one is the probability of.* Below we show more examples on high frequent Equation Edits: *d* was edited 65 times; *da* was edited 42 times; *i*was edited 37 times; *R* was edited 35 times; *B* was edited 33 times; *1* was edited 31 times; *alpha* was edited 29 times; *K* was edited 27 times; *Y* was edited 23 times; *beta* was edited 21 times; *n* was edited 20 times. Within Symbol Edits, there were a low percentage of repetitive Abbreviation Edits and Other Symbol Edits. Here we demonstrate several repetitive Abbreviation Edits: *PLSA* was edited 69 times; *IDF* was edited 44 times; *BM25* was edited 22 times. Additionally, we list several Other Symbol Edits: *D3* was edited 37 times; *D4* was edited 33 times; *d1* was edited 32 times; *D2* was edited 20 times.

Within high frequent **STEM Edits**, there were also a small percentage of Non-Symbol Edits. Non-Symbol Edits included 512 unique words and were in total edited 1692 times (Mean = 3.3. SD = 6.2). There were 34 unique words that were edited more than ten times; *sum* was edited most frequently (74 times), such as *to check what's inside this some then⟶ to check what's inside this sum then*. Here we show top Non-Symbol Edits examples. *sub* was edited 67 times; *topic* was edited 40 times; *term* was edited 36 times; *vector* was edited 29 times; *weighting* was edited 26 times; *NLP* was edited 25 times; *surfer* was edited 25 times; *weight* was edited 25 times; *representation* was edited 23 times; *estimate* was edited 23 times.

Non-STEM Edits included more repetitive edits than **STEM Edits**. 864 unique words was in total edited 7,142 times (Mean = 8.3. SD = 34.0). There were 90 unique words were edited more than ten times, 13 unique words were edited more than 100 times. The most frequent edits was “,” (547 times). Comma was used to indicate a pause between parts of a sentence. Below, we show more examples on high frequent Non-STEM Edits. *and* was edited 403 times; *word* was edited 323 times; *a* was edited 269 times; *the* was edited 267 times; “.” was edited 250 times; *to* was edited 253 times; *we* was edited 172 times; *because* was edited 152 times; *in* was edited 142 times; *were* was edited 139 times; *text* was edited 138 times; *is* was edited 128 times.

There were no repetitive **Typo**, which means there were 174 unique word-level typos. We observed two cases intentionally repetitive words. Learners intentionally meant *occurred*, but misspelled it as *occured* or *ocurr* for 13 times by four different participants. Another example repetitive Typo: learners intentionally meant *likihood*, but misspelled it as *likehood* or *liklihood* or *likelyhood* or *lihood* or *likehood* for five times by five different participants.

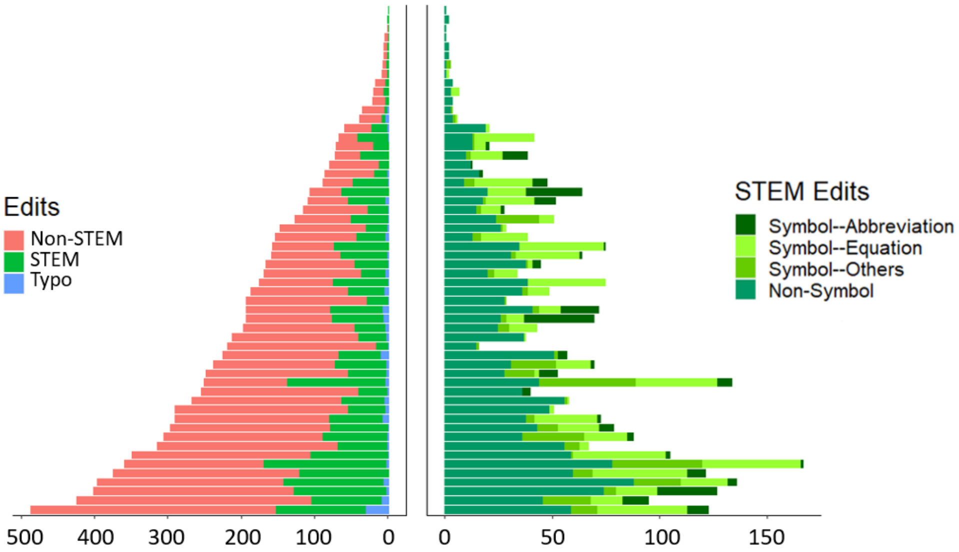


Fig. 4. Numbers of Different Types of Edits Made by 58 Editors. Left: Edits: **STEM Edits**, **Non-STEM Edits**, **Typo Edits**. Right: **STEM Edits**: Symbol Edits (Abbreviation, Equations, Others), Non-Symbol Edits. **STEM Edits** on the right side and on the right side is of same value.

In sum, students' caption editing log shows that there were significantly more **No-STEM Edits** than **STEM Edits**; **Non-STEM Edits** contained more repetitive word-level edits than **STEM Edits**. Within **STEM Edits** – Symbol Edits, there were significantly more Equation Edits than Abbreviation Edits, Other Edits, and Non-Symbol Edits; Equation Edits contained more repetitive word-level edits. To notice, repetitive edits mean after-edit word is the same; is does not mean before editing work is the same.

4.2 Individual Editing Strategies

Interview data revealed that the main goal for editing captions was to improve the accuracy of captions. Participants described three steps to improve accuracy – finding errors, fixing errors, and verifying errors – and the strategies employed at each step are discussed below.

4.2.1 Goal of Caption Editing. Improving caption accuracy through error correction was the main goal for caption editing task. They also considered efficiency while editing captions. Editors used different strategies across different steps to improve accuracy and increase efficiency. The agreed editing criteria was not to represent exactly what the lecturer spoke, but to seek a balance between accuracy and efficiency. Other secondary goals included to learn content better (e.g. better prepare for exams), and improve confidence in caption editing.

4.2.2 Identifying Errors. **Strategies: Guessing; Compare Audio, Slides, Captions** Our participants perceived that "noticing error" was a rather easy task than making actual edits and "providing the correct answer". To improve efficiency, some editors directly identify the error by directly looks at captions and identify words that "were not supposed to appear in the context, such as the word *Opera*" (P12) or "spelling itself is wrong" (P14). Another mainly used approach to improve accuracy is, when identify errors, learners watch the video and simultaneously skimming captions. Under such an approach, they can notice mismatch between audio and caption, slides and captions.

Perceived Likeliness in Noticing Different Types of Errors Different strategies were described in identifying errors in STEM and non-STEM content. For **STEM Edits** - Non-Symbol Edits were most likely to be noticed. One participant explained that such errors could be easily recognized by looking at slides and listening to the audio, "It does not require any prior understanding of what the last sentence was about, what term the professor is talking about." For example, *Now tax retrieval refers to finding* → *Now text retrieval refers to finding*. **Non-STEM Edits** were harder to notice and "sounded similar." For example, the ASR text, *the real world in some way it was* was similar to the final version, *the real world in some way it would.*

Prior Domain Knowledge Impacted Likelihood of Noticing Domain-related Errors A majority of participants agreed that editing captions towards the end of the semester allowed them to have a general understanding of the course and be more familiar with domains-related terms. P11 thought it might be harder to capture STEM errors during transcribing when watching for the first time because the participant would be less familiar with some domain-specific words e.g., might not recognize *organize* should be *tokenize.*

General Usage of Caption Impact Likeness of Noticing Errors Learning context, languages, and learning purposes. Context of learning impacts learners' reliance on captions to access knowledge. P17 said sometimes just read the caption without opening the audio since it is multi-tasking and working full-time, and the captioning accuracy impacts his understanding. P7 was learning on public context also had similar preferences. Additionally, proficiency of languages, in our case English, impacts their reliance on captions. P6 said as a non-native speaker, they found reading easier than listening. She could not stop herself from looking at captions while taking classes in English, which would never happen when consuming videos of native languages (Chinese). In such a case, it was more likely for her to notice errors in captions since the participant read them all the time. Learning purpose impacts people's attention paid to captions. participant P7 reported they noticed more errors in captions when performing the editing task than reading captions for learning; she was "actively hunting for errors." She tried to make sense of all captions and use captions as supplementary material to slides and audio while trying to learn.

Access to Visual Information Impact Likeness of Noticing Domains-related Errors Slides provided editors with information-rich references, and editors recognized errors in captions when there was a mismatch between caption terms and slide terms. Edits that required reference to slides were considered harder to notice. For example, P13 changed *with a regularizer function called were* to *with a regularizer function called R* to *with a regularizer function called r*. The participant found it easy to notice *were* did not read correctly within the textual context and replaced it with *R*, but the participant replayed the video multiple times and closely read slides to make the second edit. Another reason given by P13 was *"people might auto-correct that(errors) in mind when under certain context by looking on the slides, missing such errors, but it needs to be fixed."*

4.2.3 Editing Errors. **Strategies: Prioritize Edits; Guessing; Listen Carefully; Read Slides Carefully; Copy + Paste** A widely used strategy to improve efficiency was to prioritize edits based on errors' impact on understanding course content Many of our participants classified errors as major errors and minor errors based on perceived. *"it's not important to understand every single word that instructor says – only major concepts"* (P11) They prioritized errors that had a major impact on understanding. Another strategy to improve accuracy was to replay video clip multiple times and read slides closely. Another interesting way to improve the efficiency was to guess the correct word based on context of sentence and slides. To further improve the readability of captions, several participants moved sentences that were scattered in different lines into one line. They explained such behavior as "making arguments easier to understand by putting a full sentence together instead of one line having only two words and need to read between multiple lines to get the meaning." Under such cases, they mentioned they prioritized readability over precise match between caption and video timestamp.

Editing Effort of Different Edit Types Major errors were considered as errors changing the meaning or context of a sentence. Domain-related words were perceived as most impacting learners' understanding, *virtual* ⟶ *visual* is an example for change of context (P11); *parody* ⟶ *paradigmatically* is an example for change of meaning (P4). Participants observe such errors often happen between similar sounding words (P2, P5, P17). Nouns were more misleading when misspelled (P7), especially symbols in equations numbers such as *theta* (P11), *one* (P4). *"... these errors can cause confusion even punctuation can sometimes cause confusion e.g. 'KL, divergence' vs. 'KL-divergence'."* (P3) Major errors were likely to generate *"unnecessary cognitive load"* for learners (P2).

Minor errors were defined as errors that were unlikely to change meaning of a sentence and doesn't impact the context of a sentence. Such errors were either ignored, or sometimes revised sometimes ignored, or left for a second pass (P9, P10, P11). An example is capitalization indicating a start of sentence, such as *then* (P8, P13, P14, P15). And misspelling of non-domain related words, such as *vector* ⟶ *the vector*, *becausw* ⟶ *because*, *word* ⟶ *words*. Such errors were also called *"grammatical errors"* by some participants while providing similar explanation and examples (P4, P5, P7, P8).

According to interview results, **STEM Edits** were considered to be correcting major errors and of top priority to edit. Within **STEM Edits**, *Equation Edits* have a major impact to understanding course content and would to best extend make such edits. In total 1219 Equation edits were made. Participants understand that auto-generated captions were especially challenging in generating equation captions correctly. P12 said *"they would need to use the correct* *presentation for numbers, especially Greek alphabet and punctuation that I don't think they even have."* Example: *H(X|Y)*; *I(X;Y)*; *theta*; *sub-j*; *n(s)*. *Abbreviation Edits*, P7 also described it as *"jargon"*, also has a major impact on understanding. It was especially confusing when misspelled and various ways, such as *PLSA* misspelled as *PLA*, *SL*, *PLC*, etc. The same edited word (punctuation, numbers and capitalization and) and could be either having 'major impact' or 'minor impact' on learners' understandings based on the context it appears in (e.g. *one* → *1*) and capitalization (e.g. *map* → *MAP*). Another widely reported case was when system wrongly transcribed STEM word into simple words and changing the context (e.g. *And complete this magic.* → *and completely symmetric.*)

General Usage of Caption Impact Perceived Error Importance in Captions. Note-taking and caption searching. In many interviews, participants mentioned using captions for note-taking required higher caption quality and they had less tolerance in caption errors (P5, P6, P7). These participants copy-pasted captions into their notebook for further learning, and caption correction impacted the notes' quality. Some participants mentioned their usage of "caption searching" feature to better capture the concepts they were interested in reviewing. Some Participants used the browser-supported "Ctrl+F" to search captions. If captions were incorrect they were less likely to find the desired content. For example, *"searching captions allowed me to quickly skim through parts that I am interested"* (P15).

4.2.4 Verifying Edits. **Strategies: Proof-read; Refer to Outside Sources; Skip Errors; Follow "Transcription Norms"** During the interview, half participants felt confused or were uncertain while making edits and trying to provide the correct text. The other half were sure with all the edits they made. Most edited all errors that they noticed; a few ignored unsure errors and "left it for others." To address uncertainty and improve accuracy, two participants confirmed with external trusted sources the correct spelling and formatting of domain-related terms and symbols (P2, P3). To increase confidence and accuracy, participants read through whole caption again and verified if each sentence was readable and well connected.

Confidence in Making Different Edit Types Confident participants were more certain in editing **STEM Edits** compared to **Non-STEM Edits**. Participants reported that the corrected text of STEM errors could often be found be on the slides, and they only needed to type the corrected word manually, though it could be tedious. For **Non-STEM Edits**, participants needed to closely re-listen to video clips multiple times and play at a slower speed (e.g., 0.75-times original speed) to identify the correct word; they made sure they were spelling it correctly. In some cases they made spelling mistakes e.g., *lenngth*, *skiped*, *semtence* (classified as Typo Edits in our study). They also found such edits were challenging for non-native English speakers (P3, P12, P16, P17). Some participants mentioned they used "guessing" to propose a correct word. Within **STEM Edits**, participants expressed less confidence in editing Equations. Interviewees considered manually transcribing equations required greater *"cognitive overload"* and they *"struggle with formatting correctly"* even though the original Equation could be viewed on the slides. By formatting, multiple participants mentioned they were unsure how to transcribe numbers, especially Greek numerals and punctuation.

Prior Domain Knowledge Impact Editing Confidence and Confusions A majority of participants agreed that editing captions towards the end of the semester made them more confident when editing all kinds of errors. Five participants thoughts such domain knowledge could be especially helpful in correcting domains-related terms. As mentioned in prior sections, it was easy to identify errors when noticing words that don't belong to current content; however, providing a correct answer required domain knowledge; an example is *cinematica relations.* → *syntagmatic relations.* P11 felt confident editing captions as she'd already watched the video before and had domain knowledge. P11 such domain knowledge unconsciously used in "guessing" answers and improving editing efficiency. Learners(Editors) will be less likely to "be confused about whether the word is something they don't know, causing unnecessary doubts for new learners, or it needs to be corrected" (P12).

Familiarity with Transcription Task Impact Editing Confidence and Confusions An important factor that impacted editors' confidence and confusion with caption editing was familiarity with the transcription task. The participants described the transcription task as "subjective" and were unsure of the extent to which the transcription should record exactly what the lecturer said. For example, P9 said *"I don't know if I should be 100% accurate, should I be typing down every word the professor say, or should I rephrase it in some way."* P15 was confused by a similar overall task, *"I don't know what's the correct way, so I just take down whatever the professor says, also those filler words and when the professor changes him mind in the middle of a sentence."* Other participants reported hesitation on whether to include filler words and repeated words (e.g., "um") (P2, P3, P7). P2 eventually gave up editing task dues such reasons after making less than 20 edits, far more minor than average edits per person (he did not receive extra credit for caption editing activity).

Participants encountered several challenges in manually transcribing visual information on slides into captions. Several participants said they must have equations in captions, but they were not sure if their edits were correct or could cause more confusion. Two were unclear how to refer to information from slides, P7 used quotation markers to separate information obtained from the slides to the rest of the sentences, *"the 'the' was actually part of an example sentence presented on the slides, we were talking about how the word 'the' impacts sentence understanding, I am not sure if this is the best way."*

Two participants found their previous experience on transcription helpful for the caption editing activity. P11 compared the transcription task to previous and more challenging transcription tasks of classroom recordings of children's speech that was harder to transcribe because of the greater background noise, multiple speakers and use of non-standard words (e.g., a child referred to a "copper mine" as a "park"). Such transcribing experience made her perceive editing online video caption easier and possible without training.

4.2.5 Motivation and Outcome for Editing Captions. It is not surprising that most participants thought "extra credit" was the main motivation to participate in the caption editing activity. They thought adding "2 %" to their final grade was acceptable considering the time spent finishing such a task. Besides extra credit, participants enjoyed the process of *"making things perfect"*, and meet their habits of *"being perfectionism"*. They felt satisfied upon editing as they improved the validity of caption quality. Three participants(P9, P10, P16) considered their effort to be altruistic behavior that helps peers while having accessibility constraints and highly rely on captions. They also found their work sustainable for future learners to motivate them to make edits (P9, P10, P15). Participants also found that conducting caption editing activity before the final exam can help them better prepared for it. They benefit from closely reading and listening to each line, dictating equations and abbreviations (P1 - P5, P9). Besides, such editing activity helps them feel *"more prepared"* for final exams. However, some thought the editing motivation was to discourage because other learners were not recognizing editing efforts. For example, P8 :*"I don't know who is going to use this caption in the future, you know I don't know who I am helping, the future users won't know my name as well, then why should I do it so carefully?"*

Summary: We first provided different types of edits made by examining editing log. We pinned down the decision-making process on making edits through interview data – identifying errors, editing errors, and verifying edits in improve caption editing accuracy and efficiency. We encapsulate the editing strategies that were used, identified factors that impact such process — e.g., general usage of the caption, prior domain knowledge, familiarity with transcription task. Our findings show that **STEM Edits** take up 35% of total edits, they were easier to notice, have a major impact on understanding class content, and editors were confident with their editing except Equation Edits. **Non-STEM Edits** take up 65% to total edits, were harder to notice, and editors less confident with their editing. Editors' domain knowledge and general use of captions also have a great impact on caption editing behavior. Reading slides and re-listening to video clips were two widely used strategies improve accuracy. Guessing and prioritizing edit type was used to improve efficiency.

5 COLLABORATIVE LEARNERS' CAPTION EDITING BEHAVIOR (RQ2)

As we introduced in the Method section, *Editor One* and *Editor Two* were the two roles participants signed-up for to participate in caption editing activity. By investigating system using log, we found five participants edited captions while they were taking online class using current system as *Volunteers*. To answer **RQ2:How do learners collaborate with other learners in crowd-sourced caption editing?**, we first analyze on how *Editor One Editor Two*, and *Volunteers* collaborate to edit caption using log analysis and interview analysis. Then we report on how the editing strategies used in collaborative editing through interview analysis.

5.1 Shared Responsibility between *Editor One* and *Editor Two*-Log Analysis

Below, we report on how *Editor One* and *Editor Two* share responsibility in caption editing by investigating their editing counts, editing type and editing error in editing log. We triangulate our findings by interview questions regarding collaborative editing experience (e.g. *When being Editor Two, what do you think of quality of Editor One's work?*, *Did your editing standard change while being Editor One and Editor Two?*) and their perceptions of human-created captions for the same educational content (e.g. *What do you think of the quality of captions on Coursera for the same content?*). We removed all edits under videos where participants felt confused about editing task (such as editing video with similar ID and not viewing video with correct ID) and where perceived editor sequence differ from actual editor sequence. After data cleaning, we have 8623 edits (91.0% of total edits) made by 54 participants from 89 videos (95.7% of total videos). Among the 54 participants, 26 served twice as *Editor One* and twice as *Editor Two*, 20 served as least once as *Editor One* or *Editor Two*. Limited by the task available, eight participants who signed-up late were only able to sign-up as *Editor One* or *Editor Two* for available tasks. In Figure 5, we show *Editor Ones* and *Editor Twos* edits (total edits, **STEM Edits**, Symbol Edits, and Equation Edits) for 58 videos.

5.1.1 Editor One Makes more Edits than Editor Two per Video. *Editor Ones* in total made 7467 edits (Mean= 80.2, SD = 62.5, per video) , while *Editor Twos* in total made 1021 edits (Mean= 11.0, SD = 19.9, per video). *Editor One* made significantly more edits then *Editor Two* per video (W = 6247.5, $p < 0.001$). Same results for STEM Edits (W = 7130.5, $p < 0.001$), Non-STEM Edits (W = 7181, $p < 0.001$), Symbol Edits (W = 6846.5, $p < 0.001$), Non-Symbol Edits (W = 7176, $p < 0.001$). Our analysis also shows *Editor One* made significantly more Typos than *Editor Two*. Within 7467 edits made by *Editor One* (W = 6247.5, $p < 0.001$), there were 149 Typos; within 1021 edits made by *Editor Two*, there were 24 Typos. This result alone is not that surprising given that there were likely to be less errors after prior editors has made edits. There were videos where participants make zero edits. To be more specific, there were three video whose *Editor One* that made zero edits, 26 videos whose *Editor Two* made zero edits. Interview results below reveals *Editor Two's* explanations for such behavior.

5.1.2 Editor One and Editor Two Focus on Different Types of Edits. *Editor One* made 2112 STEM Edits within which 926 were Symbols and 1186 were Non-Symbols. *Editor Two* made 307 STEM Edits within which 154 were Symbols

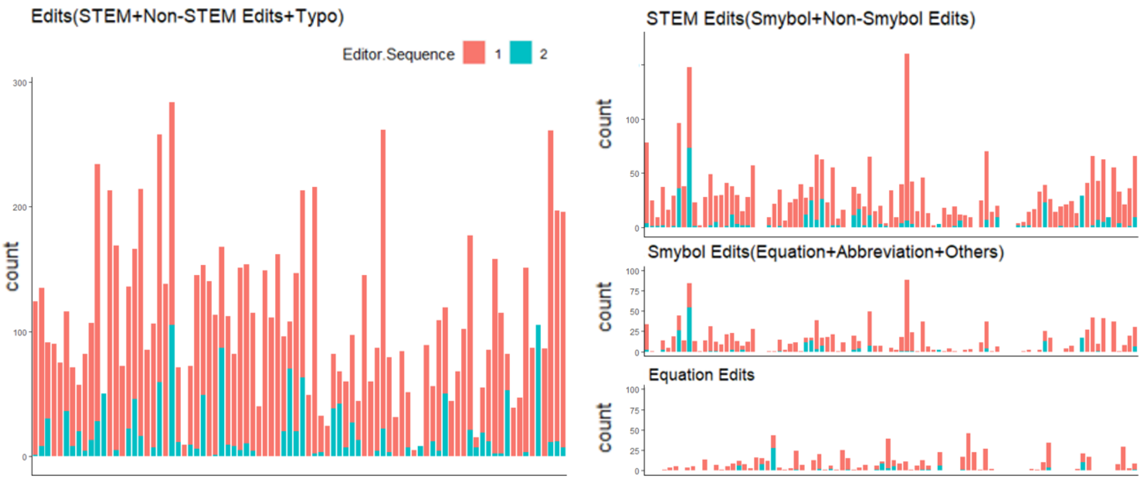


Fig. 5. Numbers of Different Types of Edits Made from 89 Videos

and 153 were Non-Symbols. *Editor One* and *Editor Two* doesn't show significant difference in STEM/Non-STEM and Symbol/Non-Symbol Edits.

Within Symbol Edits (Abbreviation, Equation and Others Symbols) , *Editor One* and *Editor Two* made significantly different kinds of edits according to Chi-Square testing (X-Square=7.13, p=0.03). Post-hoc results show that *Editor Two* were more likely to make Others Symbols edits than *Editor One* (Residuals=2.26, p = 0.05). *Editor One* made 249 Others Symbols edits (24.1%) compared to 61 Others Symbols edits (33.0%) made by *Editor Two*. Within the 61 Other Symbol Edits, a large percentage were editing presentation format of numbers, for example changing *.9* to *9*, changing *011* to *0,1. 1*, changing *V3* to *view three*, changing *zero one* to *01*. Another editing on numbers is to correct the is misspelled numbers, for example, changing *for* to 4, changing *tool* to *2*. Numbers edits were one of the special and challenging cases under section 4.1.4 which requires close attention paid to context.

5.1.3 Intrinsically Motivated Volunteers (Did not Receive Extra-Credit for Making Edits). There were five participants that made edits 164 edits for in total two videos they did not signed-up as either *Editor One* or *Editor Two*. The Within the five *Volunteers*, three edited videos they did not sign-up for. The other two did not sign-up for any videos. One of the four participated in the interview and said wasn't aware of the caption editing activity and was purely motivated by interest and enjoyed the *"hunting process of making edits"*(P17). The participant made six edits for two videos, four Non-STEM and two STEM edits. It noticed such errors while watching the video on * system and corrected them upon noticing them, for example *water* to *world*, and *engine* to *engines*. But it would not make edits if it was *"not very confident"* about his edit. It said if * system caption editing feature was introduced earlier, the instructor may not need to give extra credit to receive as much quantity and quality of edits.

It is unsurprising to see that *Volunteers* (Mean= 82. SD = 109.1) made comparatively similar amount of total edits as *Editor One* (Mean = 80.2) , and more than*Editor Two* (Mean = 11.0) (W = 1130225, p < 0.001). Among five Volunteers, only one *Volunteers* edited prior to *Editor One* and four *Volunteers* edited after *Editor One*. Therefore, *Volunteers* were not likely to impact *Editor One* and *Editor Two* caption editing task by greatly reducing before-edit error counts.

Volunteers made significantly different types of Symbol Edits than non-Volunteers (editors with committed tasks) (chi-squared = 50.2, df = 2, p < 0.001). Volunteers made more Abbreviation edits that Non-Volunteers (Residual = 6.8, p < 0.001). Among all volunteers, three participated in the interview and agreed on they tended to revise errors that were *“easy to tell and provide the most accurate edit"*. P17 :*“...error such as PLA to PLSA, which is really impacting the understanding, a lot of times it was mistaken by of several simple words that were obviously not supposed to appear together. "* Another example was *particularly Sky our papers.* → *particularly SIGIR papers.*, *So in this approach Contacts, you’re* → *So in this approach contextual* .

5.2 Collaborative Editing Strategies

Interviews allowed us to understand better the different strategies used in Identifying Errors, Editing Errors, and Verifying Edits steps for the collaboration caption editing task. Additionally, we identified factors impacting caption editing. For example *Editor Two* knew that they were the second editor but do not know the existence of other *Volunteers*.

5.2.1 Goal of Caption Editing. Interviewees reported that they valued both accuracy and efficiency when editing captions as *Editor Two*. Another goal was to improve further captions’ readability, such as consistency of punctuation and capitalization used. In general, participants perceived themselves to be *“spending less time”* when being *Editor Two* compared to being *Editor One*. Participants agreed that *Editor Two* required more patience than *Editor One*. Efficiency was also considered more important while being *Editor Two* compared to being *Editor One*.

5.2.2 Error Identification. **Strategies: Over-Trust in *Editor One*; Speed-up Video; Compare Slides and captions** A majority of *Editor Twos* trusted *Editor Ones’* work, say they find captions edited by *Editor One* is of acceptable accuracy for accessing learning contents. They perceived only minor edits were needed for them to revise. Thus, they speed up videos (1.5 or 2 times original speed) and rarely pause to re-listen. Within the process, they highly relied on visual information to compare slides and captions for better consistency on word formatting on Capitalization and the Number in STEM words. P7 thought as *Editor Two* it focused on errors that were less noticeable than when they were *Editor One*. Many participants explained that they need to refer closely to slides to identify edits, requiring understanding of class content. At the same time, they admitted they perceived themselves being “less responsible" , “paying less attention", “feel relaxed" when being the *Editor One* than *Editor Two*.

Our participants provided interesting comparisons between caption quality on the Coursera system and our system, indicating preference in human involvement over purely machine-generated captions. Ten out of 18 participants said our systems’ caption had much more errors than Coursera, but the caption quality reached equal quality after *Editor One* finished editing. Five out of 18 initially thought Coursera and our system’s caption (before *Editor One* edited) quality was equal, but after knowing Coursera is created by human immediately changed perception to ’Coursera overperforms our system.’

5.2.3 Verifying Edits. **Strategies: Communicate Uncertainty through Special Notations** As we reported in findings of RQ1, editors’ confidence in certain edits impact editors’ editing strategies. Editors were more likely to make edits for errors when they could hear the words clearly, had confidence in their English proficiency and domain knowledge. We found that one strategy editors used to denote words that they were not confident about was to replace them with a special notation, “[INAUDIBLE].” Four editors utilized this strategy.

While we were not sure if the editors used the “[INAUDIBLE]” notation as an approach to request help from subsequent editors, we found two examples in which the second editor helped complete the part denoted as “[INAUDIBLE]” by the first editor. In one example, the caption prior to any edit was *"as much crew as possible for prediction."* The first editor revised it to *"as much [INAUDIBLE] as possible for prediction."* Editor two then fixed the inaudible part, *"much clues as possible for prediction."*

Participant P2, a *Volunteer* and a native speaker who replaced an "[INAUDIBLE]" notation, commented that *"seeing [INAUDIBLE] in others' caption could encourage other editors to make edits."* The participant was aware that most editors and learners were non-native speakers, and as a native speaker themselves, P2 was more confident in making such edits. Even though a few participants agreed that using special notations such as "[INAUDIBLE]" helped communicate confusion and draw the attention of subsequent editors, three interviewees (P4,P7,P9) thought that using notations such as "[INAUDIBLE]" may create confusion for general learners.

Editing Sequence Impact Confidence Even thought participants reported applying the same editing standard regardless of their editor roles, most of them said that they need more confidence for fixing the same error when being an *Editor Two* compared to being an *Editor One*. One reason for the need of additional confidence, as put succinctly by P*, is that as a second editor, "I don't want to step on someone else's toes." Another concern of second editors is adding personal styles that are different from those of the first editors, which may cause inconsistency in formatting. In particular, inconsistent formatting of critical concepts may cause disconnection between different occurrences of the same concept.

Summary Our findings show that *Editor One*, *Editor Two* and *Volunteers* contributed to caption editing in a complementary manner. *Editor One* made more edits and spent more time. *Editor Two* made fewer edits, made significantly different types of **STEM Edits** than *Editor One*. *Editor Two* were more likely to make Other Symbol edits, mainly edits on numbers. *Editor Two* considered efficiency to be more important than *Editor One*. *Editor Two* often over-trusted the work of *Editor One* and had lower self-efficacy than *Editor One*. Editors used special notations, e.g., "[INAUDIBLE]", to communicate uncertainty in their edits, and such notations resulted in further updates from subsequent editors.

6 SUGGESTIONS FOR BETTER SYSTEM SUPPORT AND FEASIBILITY EVALUATION OF MACHINE LEARNING-BASED SOLUTIONS

To answer **RQ3: How can the system better support learners to conduct STEM-caption edits?**, we investigate two perspectives. First, we asked interviewees for their suggestions of what additional support they would like during the caption editing task; the results are presented in Section 6.1. Second, based on our findings from RQ1 and participants' suggestions, we developed two machine learning models that help with error verification and identification respectively, and evaluated their feasibility.

6.1 Interviewee Suggestions to Improve Caption Editing Experience

In the interview, we asked participants for their suggestions on how to improve the crowd-sourced caption editing task, including the interface, the process, and the policies for the task. In this section, we present interviewees' suggestions on individual edits following the "Find - Fix - Verify" model.

6.1.1 Assist Errors Identification. Among our findings of how students identified caption errors (Section 4.2.2), we have seen that students sometimes had trouble identifying non-STEM errors. At the same time, students perceived many non-STEM errors as minor and less important, as discussed in the error editing section (Section 4.2.3). Therefore, minimizing the cognitive load of identifying non-STEM errors could be valuable. Three interviewees (P9, P17, P20) proposed that incorporating grammar checks or other auto-suggestion features into the system and highlighting likely errors could help them identify potential edits more easily.

Another factor that impacted students' ability to identify errors was their prior domain knowledge (Section 4.2.2). P11 suggested that editors could benefit from an overview of the main concepts mentioned in a particular video before the editing task, e.g., the explanation of a keyword.

6.1.2 Support Caption Error Editing. As shown in our findings of caption error editing (Section 4.2.3), interviewees dealt with a wide array of edit types. Some of them expressed the needs for more support when editing certain types of errors. We have seen that some terminologies were often misspelled repeatedly in various ways, e.g., *PLSA* as *PLA*, *SL*, *PLC*. Three interviewees (P1, P5, P15) asked for features to help avoid repetitive editing on the same errors, such as automatic correction. Equation edits was another major edit type. Two interviewees (P9, P11) wanted more support for equation edits, such as enabling Greek alphabet and special symbols. Individual interviewees requested improvement for less common but more complicated edits. P6 would like the interface to minimize the overhead for fixing timestamp mismatches between the video and the caption. P15 felt that the option to mark up fragmented sentences spoken by the instructor was missing.

Besides support for specific types of edits, two interviewees suggested ways to reduce the general workload for all edits. P20 thought auto-completion suggestions could be helpful in reducing editing efforts. P16 suggested to improve the user interface to minimize the number of clicks required for each edit.

6.1.3 Increase Editing Confidence and Coordinate Collaborative Verification Process. In the section of individual edit verification strategy (Section 4.2.4), we found that some participants lacked confidence in what to edit and how to edit. The same theme emerged in the suggestions they proposed. Three interviewees explicitly mentioned that providing editing guidelines would be very helpful when they evaluate whether they are doing the right thing (P1, P2, P16). Example guidelines range from defining the balance between transcribing everything (including speaker errors) versus optimizing the readability of the caption, to specifying transcription styles such as when to capitalize a word.

Prior domain knowledge was another factor that impacted participants' confidence in their edits (Section 4.2.4). Along the same line, P3 suggested that they would be more confident with the caption quality if their own edits could be checked by reviewers who have good domain knowledge. P11's suggestion of providing an overview of concepts mentioned in a video could also build confidence in an editor.

In the editing task, the second editor had a role to verify the captions edited by the first student. Among the findings on collaborative editing strategies (Section 5.2.3), while we didn't find any direct or intentional communication among fellow editors in the interviews, we observed unintentional communication between editors via special notations. One interviewee (P16), as a second editor, hoped that they had a channel to raise issues with the first editor anonymously since they thought the first editor left many errors unfixed. On the other hand, another interviewee thought that knowing what mistakes they made during caption editing would be interesting (P6). These ideas hinted at a potential need for a communication channel between the different editors of the same video in the verification process.

6.1.4 Motivate Editing Efforts. In the course being studied, students were asked to edit the captions before the final exam. While some students found the task helpful for preparing for the final exam as discussed in Section 4.2.5, seven participants preferred editing the captions at their own time while watching the videos. They thought that editing captions before an exam felt like extra labor, while editing when they proactively watch a video any time in a semester would feel more organic and hence more motivating. More specifically, P17 proposed that showing auto-suggestion of what to fix while a user was watching a video may encourage more edits. To motivate more organic editing efforts, P18 suggested to award extra credits to voluntary edits based on the amount and quality of contributions.

In addition to changing the timing of the task, three participants (P5, P7, P8) thought that visualizing editors' contributions could motivate editing behavior as well. Examples include showing celebratory animations or messages for every editing milestone, and visualizing a viewer's own edits in the video streaming mode so that editors can see their own contributions.

As a collaborative task, peer motivation could be a valuable tool as well. One interviewee suggested that promoting healthy competition among peers may further motivate students (P8). The same interviewee mentioned that recognizing one's contribution publicly, such as showing one's name as a contributor, would motivate them as well.

6.2 Machine Learning Algorithms for Efficient Edit Verification and Error Identification

Student suggestions in previous section and findings in RQ1 indicate that there is a need to reduce and optimize student effort during manual editing, esp. during identifying errors and verifying edits. The rich editing log data collected from students could potentially be used to train machine learning models to provide such needed assistance. Thus, we trained machine learning models leveraging our manually coded data described in section 4.1. In this section, we discuss the feasibility of such models.

6.2.1 Two Classifiers. From the findings on strategies used while editing errors individually (section 4.2.3), we found that STEM Edits are major edits and learners feel that it is critical to correct STEM errors. At the same time, from Section individual strategies used for verifying edits (section 4.2.4), *Editor One* lacked confidence in edits made especially in the case of Non-STEM edits emphasising the need for verification especially for Non-STEM edits. Overall, while STEM errors are more critical to be fixed, students also needed assistance with verifying Non-STEM Edits. Thus, verifying both Non-STEM and STEM edits is important for different reasons. From Section 5.2.3, students perceived themselves as paying less attention when they serve as *Editor Two*. So, it is possible they might miss some edits during verification. Our goal is to build a classifier to help **optimize efforts** during verification by prioritizing either STEM or non-STEM edits. We refer to this classifier as Classifier-Verify (CV).

From student suggestions on assistance with identifying errors (section 6.1.1), students suggested an edit "auto-suggestion" feature that highlighted likely errors to help identify potential non-STEM errors easily. Our goal is to build a classifier that detects likely non-STEM errors in captions to **reduce efforts** during identifying errors. We refer to this classifier as Classifier-Identify (CI).

6.2.2 Problem Formulation. We first introduce some terminology to facilitate the discussion of the machine learning models. Our log data captures caption-level edits. Consider the unedited caption c_{bef} "This lecture will look" that was edited to c_{aft} "this lecture, we will". There are two new words inserted or substituted in c_{aft}, i.e., "this", "we". We refer to them as w_{aft}. Each w_{aft} is manually coded as described in section 4.1.

To identify which words in c_{bef} were edited (deleted, substituted), we first performed word phonetic alignment [41]. Without such alignment, it is not easy to identify edited words. For example, the immediate next word after "lecture" in c_{bef} is "will", which does not match with the word at corresponding location in c_{aft}, "," (comma). However, this does not mean that "will" is an error word. Alternatively, simply identifying the unique words from c_{bef} that are not present in c_{aft} also does not work because there might be multiple occurrences of a word in c_{bef}, only some of which may be errors edited by students.

Additionally, for new words that are inserted in c_{aft}, there is no corresponding word that was edited in c_{bef}. So, we augment c_{bef} with a special blank token, “BLK", at all inter-word positions. If there any insertions at an inter-word position, we mark the corresponding “BLK" token as an error. A similar strategy was used to handle insertion errors in [35], although they do not use an explicit blank token.

Table 2 shows a sample edited caption aligned with the Augmented caption, *Aug.* c_{bef}. The *incorrect* (I) and *correct* (C) labels indicate whether the word in c_{bef} is an error edited by students. Each word in *Aug.* c_{bef} is called w_{bef}. The *Incorrect* words that were substituted are assigned further error categories based on our taxonomy.

Using this terminology we now formally define the tasks of the two classifiers. The task of Classifier-Verify is posed as a binary classification problem, where given a word w_{aft} and c_{aft}, the classifier predicts whether w_{aft} is a *STEM Edit* or *Non-STEM Edit*. The task of Classifier-Identify is posed as a binary classification problem, to predict whether a given word w_{bef} and *Aug* c_{bef}, the classifier predicts whether w_{bef} is an *Non-STEM* word edited by students or *Other*. The Other category includes other types of errors (e.g., STEM) and Correct words.

Table 2. Sample caption edits alignment

Aug. c_{bef}	BLK	This	BLK	lecture	BLK	will	BLK	look	BLK	
c_{aft}	BLK	this	BLK	lecture	, we	will	BLK		BLK	
Label	C	I	C	I	I	C	C	I	C	

6.2.3 Implementation. We chronologically split the entire manually coded data into train-test splits based on the timestamp of edits. We chose this way of constructing the train and test sets to mimic a real scenario in a course where the initial edits were made by students could be used for training. Prior text classification work has also constructed the train/test datasets in a similar chronological fashion, e.g., [29].

For CV, we used a 80 − 20% train-test data split. After removing Typos, there were 8k and 2k samples (w_{aft}) in the train and test sets respectively. For CI, we used a 90 − 10% train-test data split because 1) it is a more challenging task, and 2) the number of samples was much larger for this task because we were categorizing every word in *Aug*; c_{bef}. So, 10% of the whole dataset should be a sufficiently large test set. There were 65k and 2k samples in the train set and test set for CI, respectively.

As discussed in the section 4.1.2 , Capitalization, punctuation, etc. are important for this task, so lower-casing and other standard text pre-processing steps were not applied and the raw data was used instead.

The baselines include two standard baselines often used in classification tasks. The *Majority* baseline classifies every word as the majority class. The *Random* baseline uniformly randomly selects a category to assign to each words. As our main objective is to test the feasibility of using the collected data and not necessarily find or design the best classifier for these tasks, we employed standard text classification algorithms Logistic Regression and Random Forest with standard bag-of-words features using unigram tf-idf word vectors [2]. We also used class-weighting [30] to handle imbalanced class distribution that is discussed below in the Results section.

Hyperparameters were tuned using grid search with a 5-fold cross validation on the training set.

6.2.4 Results. Table 3 shows the results (overall accuracy, Precision (P), Recall (R), and F1-measure (F1) for both STEM and *Non-STEM* class of CV the test set. As can be seen from the Majority baseline accuracy, Non-STEM class constitutes about 68% of the test set. Logistic Regression (LR) achieves the best overall performance for both STEM and Non-STEM classes. The high performance could be mainly due to many w_{aft} words (86%) and c_{aft} (<1%) in the test set that overlap with the training set.

Table 3. Performance of Classifier-Verify on test set

			STEM			Non-STEM		
Classifier	Method	Accuracy (%)	P	R	F1	P	R	F1
Baselines	Majority	67.6	0.0	0.0	0.0	0.68	1.0	0.81
	Random	52.1	0.35	0.53	0.42	0.68	0.48	0.56
CV	RF	85.4	0.92	0.6	0.73	0.83	0.99	0.90
	LR	**93.8**	**0.92**	**0.89**	**0.90**	**0.94**	**0.95**	**0.94**

We now investigate how much training data is required to achieve a comparable performance as obtained by using 80% of the dataset. Figure 6 shows the variance of test set performance on the STEM class of *CV* trained on the earliest (based on timestamps of edits) x% of the word-level edits, where x ranges from 1 to 80. Performance on the Non-STEM class follows similar patterns only with higher scores. Here, we used the best performing method, Logistic Regression with hyperparameters tuned on the whole training set. We can see that performance rises sharply and then almost plateaus. The precision of a model trained on about 10% (1k edits or equivalently edits made on about 6 − 7 lecture videos) of the total edits, is quite close in performance (within 10%) to that trained on the whole training set (80% of the dataset). As more data is added, especially the recall and F1 get a further boost. This is expected because with more data, the model sees more STEM (and non-STEM) words and thus, gets better at identifying them.

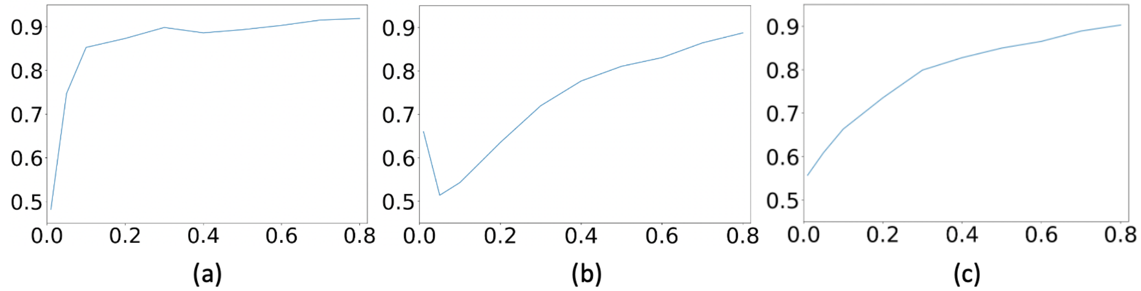


Fig. 6. Variance of (a) Precision (b) Recall and (c) F1 scores with training set size for the STEM class on the test set using CI

Table 4 shows the results of CI on the test set. Here, we show the overall accuracy, and Precision (P), Recall (R) and F1-measures (F1) on the Non-STEM class. Although the classifier performance is better than the baselines, this is clearly a more challenging task as can be seen from the much lower overall performance. This could be because 1) the same word can be mis-transcribed in many different ways 2) different instances of a word in the same caption (i,e., samples with same bag-of-words input features) could sometimes be correct and sometimes incorrect (i.e., have different labels). This is also an imbalanced problem as *Other* class constitutes about 93% of the test set. Also, between the two methods, i.e. Logistic Regression (LR) and Random Forest (RF), one has a better precision and another has a better recall. There is often a tradeoff between precision and recall. We investigate this further below.

Precision-Recall (PR) plots show how the precision and recall vary with different classification thresholds. These are especially informative for evaluating model performance on imbalanced classes [42]. Figure 7 shows the the PR plots

Table 4. Performance of Classifier-Identify on test set

Classifier	Method	Accuracy (%)	P	R	F1
Baselines	Majority	93.1	0.0	0.0	0.0
	Random	50	0.06	0.5	0.12
CI	RF	**94**	**0.87**	0.17	0.28
	LR	82.8	0.26	**0.57**	**0.36**

for logistic regression and random forest. We can clearly observe the precision-recall trade-off in the plots, i.e. high precision generally means lower recall (equivalently less false positives but low coverage). Overall, logistic regression had a higher average precision (0.45) than random forest (0.39).

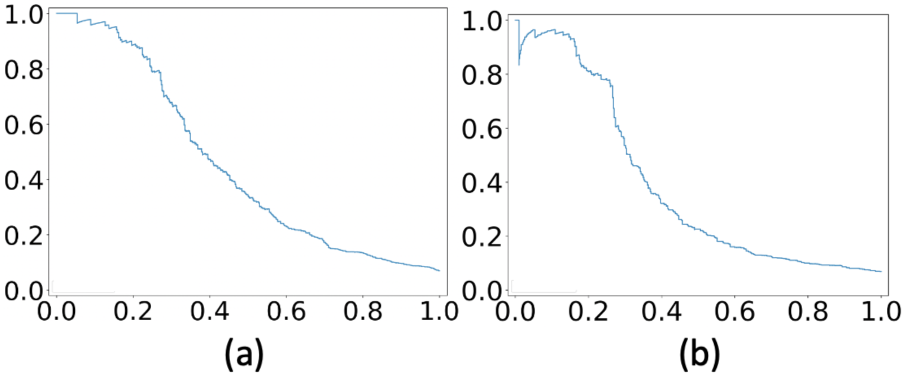


Fig. 7. Precision-Recall plot on the test set using (a) Logistic Regression and (Avg. P = 0.45) (b) Random Forest (Avg. P = 0.39) for CI. y-axis represents Precision and x-axis represents Recall

Summary Participants suggested several ways to better support both individual and collaborative STEM caption editing during the Find (Identify), Fix and Verify stages. Inspired by those suggestions and findings in RQ1, we designed two classifiers. Our model evaluation findings show that it is feasible to develop machine learning models to optimize and reduce effort during error identification and verification. The CV model to prioritize errors based on their types, during the Verify stage, achieves a high performance and could be tested for use in real-world systems. Using only about 1k labelled edits helps achieve a high-precision model and adding more training data could further boost coverage of detected STEM/Non-STEM edits. The CI model likely needs more advanced algorithms and features to accurately detect Non-STEM errors to help students reduce their efforts during the Identify stage. A trade-off between precision and recall could be made while deploying the system in practice.

7 DISCUSSION

In the section, we revisit the FFV model in the context of STEM learning videos and discuss how our findings contribute to it. Then we propose design implications for STEM learning videos caption editing systems.

7.1 "Find-Fix-Verify"FFV for STEM Learning via Videos

The "Find-Fix-Verify"(FFV) model has been shown effective in various crowd-sourcing systems, such as crowd-sourced writing [9] and micro-task assignment [10]. Our study provides empirical results on how FFV models apply to online lecture crowd-sourced caption system. Our finding align with previous three-stage model through identify errors, edit errors and verify edits errors. *Editor Two* and *Volunteers* builds on *Editor One's* work to further improve caption accuracy and readability.

From RQ1 and RQ2, we identified a suite of captioning needs that are unique to STEM learning. Our findings show that STEM errors greatly impact the understanding; editors make Symbol Edits (Abbreviation, Equation, Others) and Non-Symbol Edits to correct STEM errors. Overall, STEM errors are easier to be identified than Non-STEM error. Within STEM Edits, Equation Edits is most essential for learners to correctly understand learning content, however there are multiple challenges for either ASR or human to correctly transcribe Equations. For ASR, Equations require correctly transcribing number, punctuation and capitalization. For human, formatting of language (e.g. taking down every word to pronounce an equation or typing mathematics equation), and making sure formatting consistency between slides, audio stream and caption both raises challenges in editing errors and verifying errors. Lack of domain knowledge further decrease editors' providing the correct edit. As for Abbreviation and Other Symbol Edits, Volunteers were more likely to make Abbreviation Edits and later editors were more likely to make Other Symbol Edits. Moreover, we found there are more Non-STEM errors than STEM errors, and more repetitive errors in Non-STEM errors than STEM errors. Additionally, Non- STEM errors does not impact the learning much and does not providing editors learning opportunities by identifying/editing/verifying it, which could further decrease learners' motivation to edit.

Additionally, our RQ2 findings reveal interesting mental moods (e.g. trust and competition) when collaborating with other users on FFV crowd-sourcing writing system. Different from previous collaborating crowd-sourced editing system designed using the FFV model that divide tasks into smaller tasks, e.g., [49], our study does not perform task division, similar to Huang's work that allows subsequent workers to edit earlier workers' edits [26]. Our study gave different workers completely the same task performed sequentially by different editor. By doing so, we further reflect on task division and motivating collaboration in crowd-sourcing collaborative writing. Firstly, knowing the existence of other editors working on same task motivates some of the editors to be more careful in identifying errors and fixing errors, at the same time, could have conversing impact on some others. Editors who known that they are at later sequence felt less responsible and struggle to be more patient in editing task. Such effects could be medicated by *"Territoriality–unwillingness or inability to contribute when working with others"* [20] and *"Social loafing– Simultaneous workers may put forth less effort because they believe other group members will pick up the slack, or they feel their ideas are dispensable"* [27]. Existing research has shown that a sequential work structure in collaborative poem writing was more effective than a simultaneous work structure as the size of the group increased; territoriality partially accounts for these results [5]. Depending on the nature of the task, division into simultaneous tasks may be appropriate, or coordinated sequential might be a better fit [32]. To further improve effective collaborating in crowd-sourcing writing system, research has shown the giving workers specific roles (writer and an editor) mitigated the detrimental effects of the simultaneous work structure in collaborative writing [5]. Further work may also investigate whether such roles improves sequential editing. Additionally, empirical studies on *"Team Scaffolds"* shows that when team membership stability is not feasible, other dimensions of traditional team structures, such as boundedness and collective responsibility, can be adapted to facilitate group coordination. Future research should understand the conditions and practices that facilitate effective coordination and teamwork [46].

7.2 Design Implication for Online Lecture Videos Caption Editing

In this section, we propose design implications for improving captions in educational videos. We suggest opportunities for developing and improving machine learning algorithms for automated and machine-assisted caption generation, and recommendations for suitably assigning caption-edtiting tasks to crowdworkers and sustaining learnersourced caption editing.

7.2.1 *Machine learning algorithms for improving captions.*

Our taxonomy of the STEM edits and learners' strategies of caption edits showed some major limitations of captioning services, e.g., poor accuracy for different types of STEM words, poorly segmented captions. For improving captions for educational videos, we identify new opportunities to developing better machine learning algorithms. Firstly, the collected editing data could be used to improve the performance of automatic speech recognition algorithms. Using end-user input to improve the performance of machine learning models has been shown to help in building intelligent systems [4]. Our findings show several interesting editing behaviors on algorithmically generated captions that could be leveraged to improve such algorithms (in our case, Azure). For example, our studies showed that learners correct repetitive errors. So, future algorithms could learn from learner edits and auto-correct similar errors. Moreover, our interview participants perceived repeated errors could be instructor-related and course-related, therefore models could be updated based on course and instructor. Another important finding is that punctuation-related learner edits are important for improving STEM videos for two reasons: 1) they indicate sentence breaks (caption segments) and learners perceive that presenting a complete statement or argument within one caption improves perceived readability, and 2) accurate punctuations are important for understanding STEM terms (e.g., equations). Future algorithms could also learn from sufficient learners' punctuation edits to improve caption segmentation performance which is challenging when purely decoding audio files [3].

Secondly, in addition to using the editing log data to directly improve ASR algorithms, it could also be used to train ML models to further assist editors in editing. Such human-in-the-loop models provide more control to human editors, thus ensuring higher accuracy, while reducing and optimizing their efforts. Researchers also argue that machine learning systems should continuously *learn from users* by involving users at all stages of algorithms development, such as explorations that reveal human interaction patterns, refinement stages to tune details of the interface [4]. From RQ1 findings, we identified that editors perceive STEM errors to be more critical errors that need to be fixed accurately. On the other hand, editors need more assistance with verifying Non-STEM edits. Thus, it may be useful to provide users (e.g., course instructors or learners) *control* over the type of edits to be prioritized for verification. From the feasibility evaluation conducted in section 6.2.4, we noticed that that a high-accuracy classifier (Classifier-Verify) can be developed for classifying words inserted (or substituted) by editors. Such a classifier could be used for providing the aforementioned control. We note that the editing log data does not provide implicit training data. Edited words need to be manually labelled as STEM/Non-STEM. However, from the analysis done on the impact of training set size on CV performance section 6.2.4, we observed that it might be possible to deploy a high-precision classifier trained on a reasonably small sized dataset early during the course. Thus, minimal additional effort would be required to construct the initial training data. Adding more training data could further help boost the recall (i.e. increase coverage of identfied STEM/Non-STEM edits).

While the performance of Classifier-Identify (CI) aimed to assist editors in identifying Non-STEM errors is lower, the Precision-Recall plots in section 6.2.4 show how the classification threshold could be another "knob" to control in a practical system. A high precision model would detect few Non-STEM errors but the detected errors would indeed be actual Non-STEM errors (true positives). On the other hand, a high recall system would detect more Non-STEM errors but also generally have more inaccurate Non-STEM error detections (false positives). Additionally, there is also an opportunity to leverage and develop more advanced models esp. for sequence prediction (e.g. using RNN, BERT [19], GPT-3 [11]). If speech of the original video and ASR algorithm outputs are available, those could also be explored as is often done in standard ASR error detection and classification[35, 45]. Overall, our findings from section 6.2.4 showcase a promising future research direction for machine-assisted STEM (and other domain-specific) caption editing that has not been explored before.

7.2.2 Assigning caption-editing tasks based on crowdworkers' knowledge. From the individual editing strategies, we found that proficiency in languages impact likeliness of noticing errors and confidences in making edits. More importantly, we found other crowd-workers' previous experience, domain knowledge, and familiarity with transcription task also impact their editing behavior. For lecture videos caption editing, we highlight the necessity of matching crowd-working editors with tasks that they have sufficient knowledge in. Our findings show that editors without sufficient knowledge could be confused ("is this an error or am I not understanding this content well") and less confident in making edits. Previous worked also talked about how crowd-workers of different language background (native speaker/non-native speaker) can provide complementary contributions to improve quality caption quality [26]. Our research supports Huang's findings that proficiency in languages impact crowd-workers' caption editing behaviors. Pre-screening methods have also been utilized to select appropriate crowd workers for a given task, such as self-assessment for sentiment analysis task [25], pre-selection using machine learning models on crowd worker behavior data for image transcription task [24]. Besides self-assessment and behavior modeling, further crowd-working systems for online video lecture caption editing could also use algorithms, e.g. automatic question generation (AQG) techniques [31] to generate lecture-related question to pre-screen workers.

7.2.3 Sustaining learnersourced editing. Although our study used extra credit to motivate learners to participate, we identified 15 % participants that performed editing voluntarily with no reward. However, we also have several points to reflect on regarding how to sustain such editing behavior to other videos when there is no longer extra credit available. Firstly, to encourage learner-sourced edits, we should make the editing task better serve the purpose "to learn". For example, it should not overload and increase unnecessary cognitive load. There could be a learning mode and a editing mode; in learning mode fewer errors are highlighted for editing and avoid highlighting (changing color) distracts from reading for learning. Secondly, our findings under RQ1-3 explicitly suggested such editing task should be "gamified" and "engaging", e.g., showing number of edits made, showing edit "Completion" of a video, showing animations to create error "hunting" environment. Previous research found that some gamification elements, such as badges and leaderboards, can lead crowdworkers to do more work than they are paid for [33]. Thirdly, under RQ2-3, our findings suggest that participants are motivated when 'healthily competing with others' and 'be recognized by others'. Further research could investigate how to utilize counseling mechanisms to increase editing performance, such as 'Strength-Based Counseling Mode' motivating individuals to embrace strengths they may have when encountering adversity in the pursuit of higher goals [44].

7.3 Limitations and Future Work

This study contains several limitations that can benefit from future research. First, the study was conducted with only one course in the Computer Science department on the topic of text mining, taught by an instructor who is a non-native English speaker. Hence, the caption error taxonomy we presented in this study and the machine learning models we trained were specific to captions in the topic of text mining that were transcribed from audio spoken by a non-native speaker. Future studies on courses across different topics and disciplines and taught by instructors with different accents are needed to further expand the taxonomy of the STEM caption errors and evaluate the generalizability of the machine learning models.

Second, our interview sample biased towards non-native English speaker. The sample didn't include any student with chronic physical or mental health conditions that would prevent them from understanding the lecture video content. Both of these characteristics impact students' ability to edit captions and how much they valued accurate captions. Future work can further investigate such impacts by hearing more from native speakers and students with special needs.

Lastly, we developed the machine learning models with the goal of evaluating the feasibility of the machine learning-based solutions to support crowdsourced caption editing. Therefore, the models were more of a proof of concept and need further tuning and testing before they can be deployed in a real world scenario. Future work can improve the performance of our models and test their effectiveness in various use cases.

8 CONCLUSION

In this paper, we present our study on learner-sourcing to edit STEM-related video captions. Our study deployed a system for editing lecture video captions in a large (N=387) text mining course where 58 learners participated in editing captions of 89 lecture videos. Each lecture video was edited by two editors sequentially. Eighteen editors participated in follow-up interviews. From analysing system edit logs and qualitative analyses, we found that there is a taxonomy of errors in STEM captions. Moreover, participants used varied individual and collaborative strategies while editing the different types of errors. Inspired by the findings and students suggestions for better system support, we evaluated the feasibility of two proof-of-concept machine learning models to assist students identify and prioritize STEM and Non-STEM errors during the "Find" and "Verify" stages. We discuss the implications of our findings towards the "Find-Fix-Verify" paradigm for STEM caption editing. We also discuss the practical implications and system design suggestions based on our findings.

REFERENCES

[1] [n.d.].

[2] Charu C Aggarwal and ChengXiang Zhai. 2012. A survey of text classification algorithms. In *Mining text data*. Springer, 163–222.

[3] Aitor Alvarez, Carlos-D Martínez-Hinarejos, Haritz Arzelus, Marina Balenciaga, and Arantza del Pozo. 2017. Improving the automatic segmentation of subtitles through conditional random field. *Speech Communication* 88 (2017), 83–95.

[4] Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. 2014. Power to the people: The role of humans in interactive machine learning. *Ai Magazine* 35, 4 (2014), 105–120.

[5] Paul André, Robert E Kraut, and Aniket Kittur. 2014. Effects of simultaneous and sequential work structures on distributed collaborative interdependent tasks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 139–148.

[6] Lawrence Angrave, Karin Jensen, Zhilin Zhang, Chirantan Mahipal, David Mussulman, Christopher D Schmitz, Robert Thomas Baird, Hongye Liu, MS Wu, and R Kooper. 2020. Improving student accessibility, equity, course performance, and lab skills: how introduction of ClassTranscribe is changing engineering education at the University of Illinois. In *ASEE annual conference & exposition*.

[7] Anonymous. 2019. "What did I just miss?!" Presenting ClassTranscribe, an automated live-captioning and text-searchable lecture video system, and related pedagogical best-practices. In *ASEE Annual Conference and Exposition, Conference Proceedings*.

[8] Anonymous. 2020. Who Benefits? Positive Learner Outcomes from Behavioral Analytics of Online Lecture Video Viewing Using ClassTranscribe. In *Proceedings of the 51st ACM Technical Symposium on Computer Science Education*. 1193–1199.

[9] Michael S Bernstein, Greg Little, Robert C Miller, Björn Hartmann, Mark S Ackerman, David R Karger, David Crowell, and Katrina Panovich. 2010. Soylent: a word processor with a crowd inside. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*. 313–322.

[10] Alessandro Bozzon, Marco Brambilla, and Andrea Mauri. 2012. A model-driven approach for crowdsourcing search. In *CrowdSearch*.

[11] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165* (2020).

[12] Amanda S Clossen. 2014. Beyond the letter of the law: Accessibility, universal design, and human-centered design in video tutorials. *Pennsylvania Libraries: Research & Practice* 2, 1 (2014), 27–37.

[13] Coursera. [n.d.]. *Introducing Coursera's New Global Translator Community*. Coursera. https://coursera.tumblr.com/post/84088014661/introducing-courseras-global-translator-community

[14] Andrew Cross, Mydhili Bayyapunedi, Dilip Ravindran, Edward Cutrell, and William Thies. 2014. VidWiki: Enabling the crowd to improve the legibility of online educational videos. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. 1167–1175.

[15] Jeffrey S Cross, Nopphon Keerativoranan, May Kristine Jonson Carlon, Yong Hong Tan, Zarina Rakhimberdina, and Hideki Mori. 2019. Improving MOOC quality using learning analytics and tools. In *2019 IEEE Learning With MOOCS (LWMOOCS)*. IEEE, 174–179.

[16] Gabriel Culbertson, Solace Shen, Erik Andersen, and Malte Jung. 2017. Have your cake and eat it too: Foreign language learning with a crowdsourced video captioning system. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 286–296.

[17] Christopher Day. 2021. How We Discuss Errors and Automatic Speech Recognition. Retrieved April 15, 2021 from https://stenonymous.com/2021/04/12/how-we-discuss-errors-and-automatic-speech-recognition/

[18] Rucha Deshpande, Tayfun Tuna, Jaspal Subhlok, and Lecia Barker. 2014. A crowdsourcing caption editor for educational videos. In *2014 IEEE Frontiers in Education Conference (FIE) Proceedings*. IEEE, 1–8.

[19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[20] Michael Diehl and Wolfgang Stroebe. 1987. Productivity loss in brainstorming groups: Toward the solution of a riddle. *Journal of personality and social psychology* 53, 3 (1987), 497.

[21] Michaelanne Dye, David Nemer, Neha Kumar, and Amy S Bruckman. 2019. If it Rains, Ask Grandma to Disconnect the Nano: Maintenance & Care in Havana's StreetNet. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–27.

[22] Brianna Dym, Jed R Brubaker, Casey Fiesler, and Bryan Semaan. 2019. " Coming Out Okay" Community Narratives for LGBTQ Identity Recovery Work. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–28.

[23] Foteini Filippidou and Lefteris Moussiades. 2020. A Benchmarking of IBM, Google and Wit Automatic Speech Recognition Systems. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*. Springer, 73–82.

[24] Ujwal Gadiraju, Gianluca Demartini, Ricardo Kawase, and Stefan Dietze. 2019. Crowd anatomy beyond the good and bad: Behavioral traces for crowd worker modeling and pre-selection. *Computer Supported Cooperative Work (CSCW)* 28, 5 (2019), 815–841.

[25] Ujwal Gadiraju, Besnik Fetahu, Ricardo Kawase, Patrick Siehndel, and Stefan Dietze. 2017. Using worker self-assessments for competence-based pre-selection in crowdsourcing microtasks. *ACM Transactions on Computer-Human Interaction (TOCHI)* 24, 4 (2017), 1–26.

[26] Yun Huang, Yifeng Huang, Na Xue, and Jeffrey P Bigham. 2017. Leveraging complementary contributions of different workers for efficient crowdsourcing of video captions. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 4617–4626.

[27] Norbert L Kerr and Steven E Bruun. 1983. Dispensability of member effort and group motivation losses: Free-rider effects. *Journal of Personality and social Psychology* 44, 1 (1983), 78.

[28] Aniket Kittur, Jeffrey V Nickerson, Michael Bernstein, Elizabeth Gerber, Aaron Shaw, John Zimmerman, Matt Lease, and John Horton. 2013. The future of crowd work. In *Proceedings of the 2013 conference on Computer supported cooperative work*. 1301–1318.

[29] Bryan Klimt and Yiming Yang. 2004. Introducing the Enron corpus.. In *CEAS*.

[30] Sotiris Kotsiantis, Dimitris Kanellopoulos, Panayiotis Pintelas, et al. 2006. Handling imbalanced datasets: A review. *GESTS International Transactions on Computer Science and Engineering* 30, 1 (2006), 25–36.

[31] Ghader Kurdi, Jared Leo, Bijan Parsia, Uli Sattler, and Salam Al-Emari. 2020. A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education* 30, 1 (2020), 121–204.

[32] Walter S Lasecki, Christopher Homan, and Jeffrey P Bigham. 2014. Architecting real-time crowd-powered systems. *Human Computation* 1, 1 (2014).

[33] Sascha Lichtenberg, T Lembcke, M Brenig, AB Brendel, and S Trang. 2020. Can Gamification Lead to Increase Paid Crowdworkers Output? *15. Internationale Tagung Wirtschaftsinformatik* (2020).

[34] Liz McCarron. 2021. Creating Accessible Videos: Captions and Transcripts. *Communications of the Association for Information Systems* 48, 1 (2021), 19.

[35] Atsunori Ogawa and Takaaki Hori. 2017. Error detection and accuracy estimation in automatic speech recognition using deep bidirectional recurrent neural networks. *Speech Communication* 89 (2017), 70–83.

[36] Youngja Park, Siddharth Patwardhan, Karthik Visweswariah, and Stephen C Gates. 2008. An empirical analysis of word error rate and keyword error rate. In *Ninth Annual Conference of the International Speech Communication Association*.

[37] Becky Parton. 2016. Video captions for online courses: Do YouTube's auto-generated captions meet deaf students' needs? *Journal of Open, Flexible, and Distance Learning* 20, 1 (2016), 8–18.

[38] Rohit Ranchal, Teresa Taber-Doughty, Yiren Guo, Keith Bain, Heather Martin, J Paul Robinson, and Bradley S Duerstock. 2013. Using speech recognition for real-time captioning and lecture transcription in the classroom. *IEEE Transactions on Learning Technologies* 6, 4 (2013), 299–311.

[39] Jia Chen Ren, Mark Hasegawa-Johnson, and Lawrence Angrave. 2015. Classtranscribe: a new tool with new educational opportunities for student crowdsourced college lecture transcription.. In *SLaTE*. 179–180.

[40] Daniela Retelny, Sébastien Robaszkiewicz, Alexandra To, Walter S Lasecki, Jay Patel, Negar Rahmati, Tulsee Doshi, Melissa Valentine, and Michael S Bernstein. 2014. Expert crowdsourcing with flash teams. In *Proceedings of the 27th annual ACM symposium on User interface software and technology*. 75–85.

[41] Nicholas Ruiz and Marcello Federico. 2015. Phonetically-oriented word error alignment for speech recognition error analysis in speech translation. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. 296–302. https://doi.org/10.1109/ASRU.2015.7404808

[42] Takaya Saito and Marc Rehmsmeier. 2015. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PloS one* 10, 3 (2015), e0118432.

[43] Brent N Shiver and Rosalee J Wolfe. 2015. Evaluating alternatives for better deaf accessibility to selected web-based multimedia. In *Proceedings of the 17th international ACM SIGACCESS conference on computers & accessibility*. 231–238.

[44] Elsie J Smith. 2006. The strength-based counseling model. *The counseling psychologist* 34, 1 (2006), 13–79.

[45] Yik-Cheung Tam, Yun Lei, Jing Zheng, and Wen Wang. 2014. ASR error detection using recurrent neural network language model and complementary ASR. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2312–2316.

[46] Melissa A Valentine and Amy C Edmondson. 2015. Team scaffolds: How mesolevel structures enable role-based coordination in temporary groups. *Organization Science* 26, 2 (2015), 405–422.

[47] Juan Daniel Valor Miró, Rachel Nadine Spencer, A Pérez González de Martos, G Garcés Díaz-Munío, CIVERA Turró, J Civera, and A Juan. 2014. Evaluating intelligent interfaces for post-editing automatic transcriptions of online video lectures. *Open Learning: The Journal of Open, Distance and e-Learning* 29, 1 (2014), 72–85.

[48] Mike Wald. 2013. Concurrent collaborative captioning. (2013).

[49] Dong Wei, Senjuti Basu Roy, and Sihem Amer-Yahia. 2020. Recommending Deployment Strategies for Collaborative Tasks. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*. 3–17.