

Empowering the Deaf and Hard of Hearing Community: Enhancing Video Captions Using Large Language Models

Nadeen Fathallah*
Analytic Computing, Institute for
Artificial Intelligence, University of
Stuttgart
Stuttgart, Germany
nadeen.fathallah@ki.uni-stuttgart.de

Monika Bhole
University of Stuttgart
Stuttgart, Germany
st185128@stud.uni-stuttgart.de

Steffen Staab
Analytic Computing, Institute for
Artificial Intelligence, University of
Stuttgart
Stuttgart, Germany
University of Southampton
Southampton, UK
steffen.staab@ki.uni-stuttgart.de

Abstract

In today's digital age, video content is prevalent, serving as a primary source of information, education, and entertainment. However, the Deaf and Hard of Hearing (DHH) community often faces significant challenges in accessing video content due to the inadequacy of automatic speech recognition (ASR) systems in providing accurate and reliable captions. This paper addresses the urgent need to improve video caption quality by leveraging Large Language Models (LLMs). We present a comprehensive study that explores the integration of LLMs to enhance the accuracy and context-awareness of captions generated by ASR systems. Our methodology involves a novel pipeline that corrects ASR-generated captions using advanced LLMs. It explicitly focuses on models like GPT-3.5 and Llama2-13B due to their robust performance in language comprehension and generation tasks. We introduce a dataset representative of real-world challenges the DHH community faces to evaluate our proposed pipeline. Our results indicate that LLM-enhanced captions significantly improve accuracy, as evidenced by a notably lower Word Error Rate (WER) achieved by ChatGPT-3.5 (WER: 9.75%) compared to the original ASR captions (WER: 23.07%). ChatGPT-3.5 shows an approximate 57.72% improvement in WER compared to the original ASR captions.

CCS Concepts

• **Human-centered computing**; • **Accessibility**; • **Accessibility design and evaluation methods**; • **Applied computing**; • **Document management and text processing**; • **Document management**; • **Text editing**;

Keywords

Video Captioning, Deaf and Hard of Hearing accessibility, Large Language Models, Automatic Speech Recognition

ACM Reference Format:

Nadeen Fathallah, Monika Bhole, and Steffen Staab. 2024. Empowering the Deaf and Hard of Hearing Community: Enhancing Video Captions Using Large Language Models. In *Proceedings of Software Development and Technologies for Enhancing Accessibility and Fighting Info-exclusion (DSAI '24)*. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/nnnnnnn>.

1 Introduction

The Deaf and Hard of Hearing (DHH) community represents a significant portion of the global population. According to the World Health Organization (WHO), over 5% of the world's population – or 430 million people – are estimated to have some degree of hearing loss and require rehabilitation services for disabling hearing loss (including 34 million children). The prevalence of hearing impairment is expected to rise due to factors such as population growth, aging demographics, and increased exposure to harmful noise levels. In particular, it is projected that by 2050, nearly 700 million people will require some form of hearing assistance [36].

The challenges faced by the DHH community in accessing multimedia content are substantial. Many individuals in this group rely on captions as a primary means of accessing auditory information. However, the quality of captions often falls short of the accuracy required for complete comprehension, significantly affecting the community's ability to fully participate in educational, social, and professional settings. The need to bridge this gap is critical, as video content is increasingly used for education, communication, and entertainment, making accessible captioning an essential aspect of ensuring inclusivity for all.

Although video captions have enhanced accessibility to some extent, there remains a considerable deficit in their precision, user-friendliness, and overall effectiveness. Our research initiative explores the challenges related to the DHH community's use of video captions and improving the quality of captions generated by assistive tools.

Video content is a primary education, communication, and entertainment medium, facilitating knowledge sharing and communication. Barriers to video content accessibility arise when it has inaccurate captions. Incomplete or inaccurate captions pose a significant challenge for the DHH community, who rely on these captions to access auditory information. Accurate captions are essential to ensure that video content is inclusive and accessible to all individuals, regardless of hearing ability. Traditional video captioning

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions to permissions@acm.org.

DSAI '24, November 13–15, 2024, Abu Dhabi, UAE

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
<https://doi.org/10.1145/nnnnnnn>

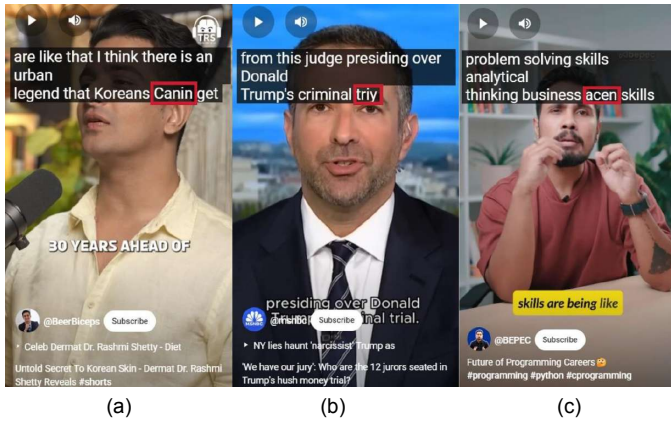


Figure 1: Example of traditional ASR systems (YouTube’s automatic video captioning feature) inaccurate captions. In figure (a) ASR system generates "Koreans canin" instead of "Koreans can", figure (b) "triy" instead of "trial", and figure (c) "business acen" instead of "business acumen".

techniques, such as manual captioning or automatic speech recognition (ASR) systems, often fail to capture the exact spoken words [16]. This inadequacy can lead to distorted meanings or omitting essential information, disrupting content accessibility for the DHH community. Our study aims to improve the quality of video captions by integrating Large Language Models (LLMs) into the video captioning pipeline.

Figure 1 shows inaccurate video captions produced by Youtube’s automatic video captioning feature that uses ASR to generate captions. Such errors can alter the content’s intended meaning, leading to viewers’ confusion. These incorrect captions serve as the input to our proposed pipeline, illustrated in figure 3 to improve the quality of video captions, and the corrected captions generated by LLMs are the expected output. The goal of this research is to correct these errors to ensure that video content is fully accessible. Our research question is:

RQ. Can Large Language Models (LLMs) be used to enhance the quality of video captions for the DHH community?

The primary contributions of this research project are:

- Investigating and identifying the issues and challenges the DHH community faces when using video captions.
- Using the identified challenges to curate a dataset specifically tailored to these issues.
- Leveraging LLMs to improve the quality of ASR-generated video captions produced by existing assistive technologies (e.g., YouTube’s automatic video captioning feature).
- Evaluating the performance of LLMs in correcting video captions by measuring their accuracy improvements through metrics such as Word Error Rate (WER).

2 Related work

The DHH community relies on captioning solutions such as ASR systems to access auditory information. This section presents current ASR systems for video captioning, challenges in ASR systems for video captioning, and why LLMs are suitable to address those challenges.

2.1 ASR Systems for Video Captioning

Automatic Speech Recognition (ASR) is a technology that uses Machine Learning/Artificial Intelligence to convert human speech into text. It is crucial for the DHH community as it provides a cost-effective, real-time solution for accessing spoken content, enhancing communication, and enabling participation in various activities. Captioning with ASR systems has seen significant advancements in generating high-quality captions, making them a cost-effective alternative to human captioning services [8], [24].

Deep learning, particularly end-to-end models like Connectionist Temporal Classification (CTC) and Sequence-to-Sequence (Seq2Seq) architectures, has notably increased transcription accuracy [10], [5]. Transformer-based models improve performance by handling long-range dependencies in audio sequences [34]. Notable implementations that provide a cost-efficient solution for the DHH community in educational settings include the tabletop tool [33] to facilitate DHH users’ communication, the APEINTA system [13] to generate captions on multiple platforms, and the E-Scribe’s web-based solution [4]. Despite these advancements, challenges in ASR, such as accuracy in diverse environments, ambient noise, and contextual understanding, remain [18].

2.2 Challenges in Video Captioning

Despite improvements in ASR technology, the suitability of automatic captions for the DHH community remains a contentious issue. Even minor inaccuracies can significantly impact comprehension, particularly for individuals who rely solely on captions to access auditory information. Many members of the DHH community consider automatic captions to be only a starting point rather than a complete solution. For instance, a seemingly small error rate—such as 0.1%—could lead to the misinterpretation of critical content, particularly when it involves homophones, specialized terminology, or idiomatic expressions. The reliance on ASR alone without further correction can, therefore, result in incomplete or confusing information. This concern highlights the importance of integrating more sophisticated correction mechanisms to bridge the gap between automatic captions and human-generated quality.

ASR systems still face challenges:

Accuracy Issues: Individuals have distinct accents and dialects, varying from standard American or Indian accents, and regional pronunciation differences. Consequently, video content may contain variations that affect captioning precision, leading to ASR errors [16], [15].

Quality of Captions: Captions often suffer from typos, incorrect grammar, and unnatural breaks. ASR systems may misinterpret domain-specific terminology, causing inaccuracies (e.g., confusing "SQL" with "sequel") [16].

Ambient Noise: Background noise significantly increases captioning errors by impacting speech recognition accuracy. Managing ambient noise is crucial for improving caption precision [16].

Homophones: ASR systems often confuse homophones (e.g., "to," "two," "too") due to contextual misunderstanding, leading to inaccurate captions [16].

Speaker Speed: Fast speaking can reduce ASR accuracy, causing captions to miss or misinterpret spoken words [16].

Code-Switching: Switching between languages or dialects within a conversation challenges ASR systems, requiring seamless handling of multiple linguistic inputs [19].

2.3 LLMs for improving Video Captions Quality

Large Language Models (LLMs) are machine learning models that can comprehend and generate human language text [14]. Earlier studies have typically utilized models like BERT to enhance ASR through methods like rescoring and distillation. Rescoring involves improving the quality of generated text by re-evaluating multiple possible outputs produced by an ASR model. For instance, ASR systems often generate an n-best list—a set of multiple possible transcriptions ranked by likelihood—and a rescoring model can be used to select the most accurate one. For example, [30] applied BERT for n-best list rescoring, while [37] used Minimum Word Error Rate (MWER) loss to train a BERT-based rescoring model that minimizes errors across the top-ranked transcriptions.

Distillation, on the other hand, is a training approach where a smaller, less complex model (the "student") is trained using the knowledge transferred from a larger, more complex model (the "teacher"). This is done by using the teacher's outputs as "soft labels" to provide more informative learning signals. In [9], BERT was employed in a distillation framework to create such soft labels for training ASR models, effectively teaching the student model nuanced semantic representations. Additionally, [17] explored how to transfer the semantic knowledge embedded in large models, improving ASR performance by enriching the distilled information.

Although these techniques demonstrated the effectiveness of LLMs in improving ASR outputs, they were limited to older LLMs with comparatively less advanced language capabilities.

With the rapid development of newer and significantly more powerful LLMs, such as ChatGPT and Llama2, the potential to enhance ASR systems has increased substantially. These newer models introduce in-context learning, a capability not present in earlier models, which allows them to adapt more flexibly to new prompts without retraining. For example, [31] uses pre-trained LLMs for transforming noisy ASR annotations into high-quality captions, demonstrating the potential of LLMs to generate "human-like" captions without the need for labeled training data.

Moreover, recent advancements in LLMs make them suitable for addressing the following challenges in video captioning.

Handling Accents and Dialects: LLMs, trained on diverse datasets, adapt to various accents and dialects, enhancing transcription accuracy. [27] highlight that models like GPT-3 can generalize across different linguistic inputs, reducing errors in captions caused by pronunciation variations.

Improving Caption Quality and Accuracy: LLMs enhance grammatical accuracy and coherence. [3] show that GPT-3 generates

syntactically correct sentences, which reduces typos and grammatical errors in captions.

Managing Ambient Noise: LLMs leverage contextual understanding to correct words distorted by background noise. [34] introduced the Transformer architecture, enabling models to maintain accuracy even in noisy environments by predicting word sequences based on context.

Differentiating Homophones: LLMs excel at using context to differentiate homophones, reducing errors. [35] demonstrate that models like GPT-3 can distinguish between homophones (e.g., "to," "two," and "too") accurately based on surrounding text.

Coping with Fast Speech: LLMs can handle rapid speech by maintaining contextual coherence, improving transcription accuracy for fast speakers. [28] show that context-aware models like BERT can effectively manage high-speed inputs.

Addressing Code-Switching: LLMs manage code-switching by handling multiple languages within a single context. [38] discuss how models like mT5 enable seamless transitions between languages in captions

3 Methodology

3.1 Dataset

3.1.1 Dataset Collection.

While existing datasets like LibriSpeech [23], TED-LIUM [11], and Common Voice [2] offer valuable resources for improving ASR system-generated captions, they are not entirely suitable for our task due to their limitations in domain diversity and capturing specific captioning challenges. These datasets often lack the varied contexts and errors in real-world video captions. To address this, we generated our open-domain dataset, comprising videos from diverse domains such as education, cooking, travel and tourism, entertainment, and news. The duration of collected videos ranges from 1-3 minutes. With accessibility often being an "afterthought", we curated our dataset with the DHH community in mind, ensuring it encompasses a wide range of caption challenges users frequently encounter. Our curated dataset includes multiple instances of each of the following common captioning challenges: accuracy and quality issues (e.g., typos, grammatical mistakes, domain-specific terminology), ambient noise, homophones, and speaker speed. Our data collection process was driven by the goal of gathering real-world data from a widely used platform that offers an ASR system for generating captions. Recognizing YouTube's extensive use by the DHH community and its ASR system for generating captions as an assistive tool for accessing auditory information, we chose it as our source. Our dataset is publicly available at <https://github.com/monikabh001/Improving-the-Quality-of-Video-Captions-for-the-DHH-Community-Using-LLM>, promoting transparency and enabling further research and development. Details about the dataset are provided in Table 1. This process resulted in the collection of 52 videos.

3.1.2 Dataset Annotation.

Numerous studies have considered ASR an assistive tool for DHH to access auditory information [16], [15]. However, inaccurate captions are a significant limitation of ASR. After collecting 52 videos

Table 1: Overview of our dataset collected. The table summarizes the number of videos, the average duration of videos, and the total duration of video content in each domain. "Number of Videos" represents the total count of videos collected in each category. "Average Duration" denotes the mean length of the videos within each category. "Total Duration" indicates the cumulative length of all videos in each domain.

	Education	Cooking	Travel and Tourism	Entertainment	News	Total
Number of Videos	10	10	10	10	12	52
Average Duration	1m 33s	1m 57s	1m 9s	1m 39s	1m	1m 27s
Total Duration	19m 34s	15m 34s	11m 31s	15m 48 s	11m 28s	73m 55s

from diverse domains, we utilized the YouTubeTranscriptApi [25] to retrieve the ASR-generated captions from YouTube. Thus, the assistive tool under investigation in this study is YouTube’s ASR automatic captioning feature, and the retrieved captions serve as the input to our proposed model. To evaluate our model’s caption correction performance, we manually generated the ground truth captions for these videos, enabling us to compare the corrected captions produced by our model to the actual words spoken in the videos. Details about the dataset variables are presented in Table 2

3.2 Model Selection

This study aims to integrate LLMs with assistive tools commonly used by DHH students to generate and correct automatic captions. The captions produced by these assistive tools serve as inputs for the LLMs, after that we prompt the LLM to correct any inaccuracies. This section details the various LLMs we experimented with to enhance video captioning quality:

- GPT-2: OpenAI’s second-generation generative pre-trained transformer model is known for its performance in language modeling and reading comprehension tasks. However, it has shown limitations in summarization and question-answering tasks [6]. Our implementation of GPT -2 for caption correction revealed no significant improvement in the quality of the captions, and it generated irrelevant context for the input caption.
- T5: As an encoder-decoder model, T5 excels in translation and summarization tasks. However, it performed poorly in the caption correction task, struggling with understanding context and preserving errors from the input. [12].
- Llama2-13B: Meta’s LLM, Llama-2, specifically the 7B model [32]. This model showed better handling contextual information and correct captions successfully.
- GPT-3.5: Representing a significant leap from GPT-2, GPT-3.5 demonstrated superior ability in generating coherent and contextually accurate text [3]. Its advanced processing capabilities resulted in more fluent and precise captions, making it highly effective for enhancing ASR-generated captions. We employed the ChatGPT-3.5 web interface for our corrections.

In our model selection process, we tested various sample captions with errors, one of which is illustrated in figure 2 'I was walking in the son. The day was brght and son y.'. GPT-2 generated irrelevant content, which can be primarily attributed to its nature as a text-generation model. Meanwhile, T5 failed to correct all the caption errors and generate the whole caption. Both Llama2-13B and ChatGPT-3.5 successfully corrected the caption,

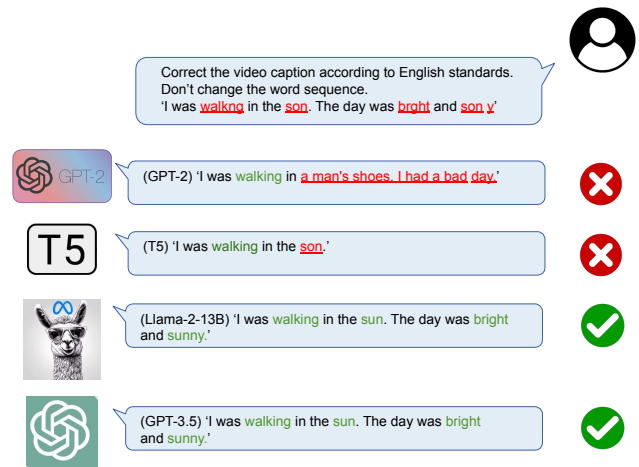


Figure 2: Selection process of Large Language Models (LLMs) for improving the quality of captions generated by ASR systems. The figure illustrates the performance of different LLMs (GPT-2, T5, Llama2-13B, GPT-3.5) in correcting a sample video caption while maintaining the original word sequence.

demonstrating their effectiveness. This comparison highlights the necessity of selecting robust LLMs to improve caption quality in assistive technologies. The notable performance of Llama2-13B and ChatGPT-3.5 led us to incorporate them into our pipeline for further comparative studies to determine the most suitable model for our task.

3.3 Pipeline for improving video caption quality using LLMs

To improve the quality of ASR system-generated captions, we propose a pipeline that leverages LLMs’ advanced capabilities, including contextual understanding, text generation, text correction, and semantic coherence. Our proposed pipeline is illustrated in figure 3.

Input: Our dataset of video content covers various domains like education, cooking, travel and tourism, entertainment (such as movies), and news, collected from YouTube. Videos range is between 1 and 3 minutes. We chose YouTube’s automatic caption

Table 2: Description of our dataset variables. The table includes three columns: "Name" lists the variable names, "Description" provides a brief explanation of each variable, and "Type" indicates the data type of each variable.

Name	Description	Type
VideoID	A unique identifier assigned to each video in the dataset	integer
URL	The web address where the video can be accessed on YouTube	text
Youtube_Caption	The captions generated automatically by YouTube's ASR system for the video	text
Ground Truth_Caption	The manually generated accurate transcription of the video's spoken content	text
Domain	The category or subject area to which the video belongs, such as education, cooking, or news	text

generation feature as the assistive tool under investigation to generate captions. The input to the LLM is the captions generated by YouTube's automatic caption generation feature as text.

Improving captions with LLMs: We prompt LLMs to correct input inaccurate captions without altering the word sequence. To construct the prompts, we use zero-shot prompting, which refers to the model's ability to perform a task without any prior examples or specific training on that task, based solely on the instructions given in the prompt [29], with their vast pre-existing knowledge. LLMs can generate accurate and contextually appropriate captions without needing additional task-specific information or training. We conduct experiments with two LLMs, ChatGPT-3.5 and Llama2-13B, based on their performance as discussed in 3.2.

Output: The output generated by LLM is the corrected captions. The output generated by LLM is then compared with manually created ground truth. The generated captions are more accurate and contextually relevant than the initial captions.

Implementation: We implemented our project using Python within a Jupyter Notebook environment.

Link to GitHub project repository: <https://github.com/monikabhohle001/Improving-the-Quality-of-Video-Captions-for-the-DHH-Community-Using-LLM>

4 Evaluation

The final step is the evaluation of our caption correction pipeline. We employ two approaches for evaluation: how the selected LLMs (Llama2-13B and ChatGPT-3.5) perform on different challenges in captions, such as homophones, inaccurate words, and domain-specific terminology, and the use of quantitative metrics for a more objective assessment of caption accuracy and quality.

4.1 LLM Performance on Captioning Challenges

Prompt: "Correct the caption according to English standards. Don't change the word sequence".

- Homophones and inaccurate words:
 - Input caption: "I was walkng in the son. The day was bright and pleasant because of son y weather."
 - ChatGPT output caption: "I was walking in the sun. The day was bright and pleasant because of sunny weather."
 - Llama2-13B output caption: "I was walking in the sun. The day was bright and pleasant because of sunny weather."
- Domain-specific terminology:

- Input caption: "Today, we'll explore advanced concepts in Sea Plus Plus programming, focusing on poly more fizz 'em."
- ChatGPT output caption: "Today, we will delve into advanced concepts in C++ programming, specifically focusing on polymorphism and encapsulation."
Llama2-13B output caption: "Today, we'll delve into advanced Sea Plus Plus programming concepts, with a focus on polymorphism."

The results demonstrate that both LLMs, ChatGPT-3.5 and Llama2-13B, effectively corrected homophones and inaccurate words in the input captions, resulting in coherent and contextually appropriate captions.

Regarding domain-specific terminology, ChatGPT-3.5 provided a more comprehensive correction by identifying and correcting "Sea Plus Plus" to "C++", whereas Llama2-13B retained "Sea Plus Plus.". Llama2-13B also changed "focusing on" to "with a focus on.", while this improves language use within the caption, it may not help DHH users if the video discusses bad writing styles. Additionally, if such changes by Llama2-13B change the length of the caption significantly, it will disrupt caption and video synchronization. These results indicate that while both models successfully correct general language errors, ChatGPT-3.5 shows a slightly better capability in handling domain-specific terminology. However, ChatGPT-3.5 also added "encapsulation" to "polymorphism". While this shows better domain-specific understanding, it may not help DHH users if the video doesn't discuss encapsulation.

4.2 Quantitative evaluation metrics

We employed three standard metrics to measure the performance of our caption correction pipeline: Word Error Rate, BLEU Score, and ROUGE Score. These metrics are essential for understanding how closely the generated captions match the ground truth captions.

To clarify the terms used in our evaluation, here are the definitions:

- Generated caption: Automatically generated caption by an assistive tool (e.g., YouTube's ASR system).
- Ground truth caption: The correct caption of the video, which is generated manually.
- Predicted caption: This is the caption generated by LLM (Corrected caption).

Word Error Rate (WER) [1]: Measures the percentage of errors in the predicted caption compared to the reference caption. It accounts for incorrect, omitted, or inserted words, calculated

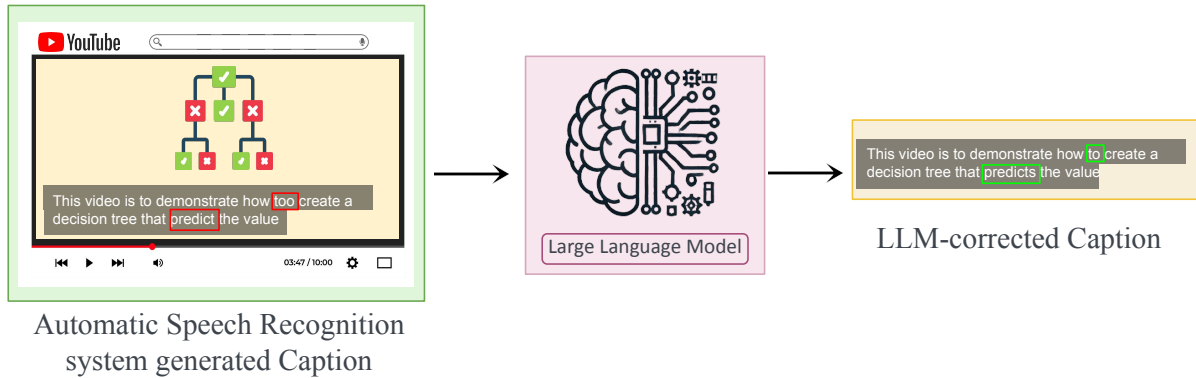


Figure 3: Pipeline - Our proposed caption correction pipeline leveraging LLMs. The input is the ASR system-generated caption (text), shown on the left, which includes errors highlighted in red. The output is the LLM-corrected caption (text), shown on the right, where the corrections are highlighted in green.

using the formula in Equation 1. A lower WER indicates a higher accuracy of the predicted captions, and a WER of 0% indicates a perfect match.

$$WER = \frac{S + D + I}{N} \quad (1)$$

- Substitutions (S): Incorrect words in the predicted caption.
- Deletions (D): Missing words in the predicted caption present in the ground truth caption.
- Insertions (I): Extra words in the predicted caption, not in the ground truth caption.
- N : Total number of words in the ground truth caption.

Bilingual Evaluation Understudy (BLEU) [20]: Calculates the precision of n-grams, n-grams are consecutive sequences of n words, measuring the percentage of n-grams in the predicted caption that matches the ground truth caption. It evaluates how well the predicted caption captures the exact word sequences of the ground truth caption. Commonly used n-grams include unigrams (single words), bigrams (two-word sequences), and trigrams (three-word sequences). The final BLEU score is calculated by aggregating the precision of unigrams, bigrams, trigrams, and four-grams, balancing the contributions of shorter and longer sequences. The BLEU score ranges from 0 to 1, with higher values indicating better quality.

Recall-Oriented Understudy for Gisting Evaluation (ROUGE) [20]: Measures overlapping n-grams between predicted caption and ground truth caption. The ROUGE score determines the recall of n-grams, ensuring the LLM-predicted caption includes essential content from the ground truth caption. The score ranges from 0 to 1, with higher values indicating better quality. ROUGE scores are categorized as follows:

- ROUGE-N: Measures overlap of n-grams.
- ROUGE-L: Measures the longest common subsequence (LCS).

WER measures word-level errors, which do not account for the overall meaning and context of the captions; it severely penalizes minor errors that do not change the sentence’s meaning. BLEU and ROUGE help address this limitation by evaluating the precision and

recall of word sequences. BLEU measures the precision of n-grams, capturing the overlap of words between the predicted and ground truth caption. However, it can struggle with longer sequences and recall. ROUGE emphasizes recall and the longest common subsequence, making it better suited for assessing the overall structure and meaning. Together, these metrics ensure that captions are not only error-free but also accurate and comprehensive, capturing both lexical and semantic similarities.

The following example illustrates how each metric evaluates different aspects of caption correction performance, as shown in Table 3:

Ground truth caption:
 "The quick brown fox jumps over the lazy dog."
 Predicted caption:
 "The quick brown fox leaps over the lazy dog."

5 Results

Leveraging LLMs to improve the quality of ASR video captioning tools has shown to be a promising approach. In our evaluation, we tested two different LLMs, ChatGPT-3.5 and Llama2-13B, using our dataset of 52 videos. The experiments were conducted using Google Colab, specifically utilizing its Python notebook environment. The models’ performances were assessed using standard natural language processing metrics: WER, BLEU, and ROUGE scores, as detailed in Table 4.

Our results demonstrate that ChatGPT-3.5 (WER: 9.75%) shows a significant improvement over the original Youtube-ASR-Caption system (WER: 23.07%), highlighting its superior accuracy in reducing errors per word. Additionally, the BLEU score for ChatGPT-3.5 is notably high at 0.85 compared to the original Youtube-ASR-Caption system BLEU score of 0.67, suggesting superior precision in n-gram matching and, thus, higher accuracy in caption generation. ChatGPT-3.5 achieved a high ROUGE-1 score of 0.98, similar to the Youtube-ASR-Caption system, indicating a strong recall of unigrams (single words). Moreover, ChatGPT-3.5 slightly outperforms the ROUGE-2 (0.97) Youtube-ASR-Caption system.

Table 3: Evaluation Metrics for caption correction performance

Metric	Score	Explanation
WER	11.11%	Measures the word-level errors, indicating an 11.11% error rate due to the single substitution of "jumps" with "leaps".
BLEU	0.66	Evaluates the precision of n-grams, resulting in a lower score due to the substitution affecting the overall n-gram match.
ROUGE-1	0.89	Measures unigram recall, showing high similarity by capturing the overlap of individual words between the captions.
ROUGE-2	0.75	Measures bigram recall, reflecting the impact of the verb substitution on the sequence of two-word combinations.
ROUGE-L	0.89	Measures the longest common subsequence, indicating strong structural similarity despite the verb change.

Llama2-13B, while performing well in certain aspects, did not achieve the same overall performance level as ChatGPT-3.5. Llama2-13B did not improve the WER compared to the Youtube-ASR-Caption system. This increase in WER can be attributed to Llama2-13B's tendency to improve the language of the text, which results in better readability and coherence but does not match the original captions exactly. For example, Llama2-13B might change "focusing on" in the original caption to "with a focus on" as shown in section 4.1, which improves the language but increases the WER because it deviates from the ground truth wording. The relatively lower BLEU score (0.68) and ROUGE scores (ROUGE-1: 0.90, ROUGE-2: 0.82, ROUGE-L: 0.88) for Llama2-13B indicate that, while it captures some aspects of the captions correctly, it struggles with precision and recall compared to ChatGPT-3.5.

6 Conclusion

In conclusion, integrating LLMs into caption correction systems can significantly enhance the accuracy and coherence of captions, improving accessibility for the DHH community. Our evaluation shows that ChatGPT-3.5 outperforms Llama2-13B regarding caption quality, capturing both individual words and longer sequences effectively and ensuring the generated captions are detailed and contextually accurate.

However, a limitation of this project is that LLMs might miss or misinterpret voice intonations, cultural references, and idioms that manual captioning can understand and convey accurately. For instance, a human captioner can appropriately interpret the emphasis in "Really?!" to express surprise, while an LLM might transcribe it as "Really". Similarly, cultural references like "Diwali" (the Hindu festival of lights) or phrases like "namaste" (a traditional Hindu greeting) could be misunderstood by LLMs but accurately captured by human captioners. To address these limitations, we propose employing multi-modal LLMs and incorporating ML models designed to understand complex human communication nuances.

7 Future Work

Future research and development efforts can enhance the effectiveness of video captioning through several key areas. Expanding our current dataset beyond YouTube ASR system-generated captions to include other widely used platforms like Microsoft Teams and Zoom would increase the applicability of our solutions to the DHH community.

In addition, considering the potential scalability of our approach is crucial for maximizing its impact. Investigating how well the LLM-based captioning system can handle larger datasets and diverse input types will be essential for scaling up the solution to different environments and applications. Exploring the deployment of this solution in low-resource settings or environments with limited computational capabilities is a necessary step to ensure accessibility across a broad range of users. Optimization of model efficiency, such as using quantization techniques or lighter model variants, could further facilitate the scalability of this approach to a production-level tool.

The applicability to other assistive technologies is another avenue worth exploring. Beyond video captioning, the methods developed here could extend to applications like real-time speech-to-text systems used in classrooms or workplaces, benefiting not only the DHH community but also those with auditory processing disorders or language learners. Integrating LLMs for enhanced subtitle generation in augmented reality (AR) or virtual reality (VR) environments could also pave the way for immersive and accessible experiences, improving real-time communication. Exploring these broader applications will further demonstrate the flexibility and impact of LLM-based solutions in assistive technology.

Addressing code-switching is another essential area of focus. YouTube's ASR system feature currently handles only one language at a time, which presents a significant limitation. Existing code-switching ASR datasets such as [7], [22], [21] handle switching between two languages only, which is not suitable for evaluating LLM performance on audio containing more than two languages. Curating a dataset to address this challenge requires using tools such as Whisper by OpenAI [26], designed to handle multiple languages within a single audio stream. We conducted some preliminary experiments to evaluate LLMs' performance for the code-switching challenge, a sample of the experiments is shown in figure 4. Our results show that ChatGPT-3.5 and Llama2-13B can successfully handle mixed-language inputs and produce accurate English captions.

Furthermore, experimenting with more advanced prompt engineering techniques, such as context-aware prompting. Context-aware prompting is a prompt engineering technique that includes relevant information in the prompt to improve LLM responses. Contextual information includes but is not limited to the domain of the video and its metadata.

Lastly, user experiments on the pipeline will be conducted with the DHH community to gather feedback and make necessary modifications. This feedback can also improve LLM performance through reinforcement learning, ensuring a user-centered solution that effectively meets their needs.

Table 4: Results – Comparison of Caption Correction Performance: YouTube’s ASR System vs. ChatGPT-3.5 and Llama2-13B

Metric	WER	BLEU	ROUGE-1	ROUGE-2	ROUGE-L
Youtube-ASR-Caption	23.07	0.67	0.98	0.96	0.98
Llama2-13B	24.42	0.68	0.90	0.82	0.88
ChatGPT-3.5	9.75	0.85	0.98	0.97	0.98

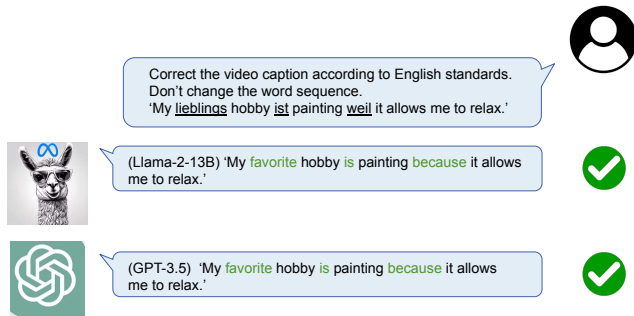


Figure 4: Preliminary experiments to evaluate LLMs’ performance on code-switching challenges in ASR video captions demonstrate that Llama-2-13B and ChatGPT-3.5 successfully corrected the caption, showcasing their ability to handle mixed-language inputs and produce accurate English captions.

Acknowledgements

This work was supported by the Integrated AI in Teaching at the University of Stuttgart project (IKLeUS), funded by the German Federal Ministry of Education and Research (BMBF).

References

[1] Ahmed Ali and Steve Renals. 2018. Word error rate estimation for speech recognition: e-wer. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers*. Iryna Gurevych and Yusuke Miyao, (Eds.) Association for Computational Linguistics, 20–24. doi: 10.18653/V1/P18-2004.

[2] Rosana Ardila et al. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*. Nicoletta Calzolari et al., (Eds.) European Language Resources Association, 4218–4222. <https://aclanthology.org/2020.lrec-1.520/>.

[3] Tom B. Brown et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*. Hugo Larochelle, Marc Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, (Eds.) <https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfc4967418bfb8ac142f64a-Abstract.html>.

[4] Zdenek Bumbalek, Jan Zelenka, and Lukas Kencl. 2010. E-scribe: ubiquitous real-time speech transcription for the hearing-impaired. In *Computers Helping People with Special Needs, 12th International Conference, ICCHP 2010, Vienna, Austria, July 14-16, 2010, Proceedings, Part II (Lecture Notes in Computer Science)*. Klaus Miesenberger, Joachim Klaus, Wolfgang L. Zagler, and Arthur I. Karshmer, (Eds.) Vol. 6180. Springer, 160–168. doi: 10.1007/978-3-642-14100-3_25.

[5] William Chan, Navdeep Jaitly, Quoc V. Le, and Oriol Vinyals. 2016. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2016, Shanghai, China, March 20-25, 2016*. IEEE, 4960–4964. doi: 10.1109/ICASSP.2016.7472621.

[6] David L. Chen and William B. Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *The 49th Annual Meeting of the Association for*

Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA. Dekang Lin, Yuji Matsumoto, and Rada Mihalcea, (Eds.) The Association for Computer Linguistics, 190–200. <https://aclanthology.org/P11-1020/>.

[7] Shammur Absar Chowdhury, Amir Hussein, Ahmed Abdelali, and Ahmed Ali. 2021. Towards one model to rule all: multilingual strategy for dialectal code-switching arabic ASR. In *22nd Annual Conference of the International Speech Communication Association, Interspeech 2021, Brno, Czechia, August 30-September 3, 2021*. Hynek Hermansky, Honza Cernocký, Lukás Burget, Lori Lamel, Odette Scharenborg, and Petr Motlicek, (Eds.) ISCA, 2466–2470. doi: 10.21437/INTERSPEECH.2021-1809.

[8] Jean Louis Fendji Kedieng Ebongue, Diane M. Tala, Blaise Omer Yenke, and Marcellin Atemkeng. 2022. Automatic speech recognition using limited vocabulary: A survey. *Appl. Artif. Intell.*, 36, 1. doi: 10.1080/08839514.2022.2095039.

[9] Hayato Futami, Hirofumi Inaguma, Sei Ueno, Masato Mimura, Shinsuke Sakai, and Tatsuya Kawahara. 2020. Distilling the knowledge of bert for sequence-to-sequence asr. *arXiv preprint arXiv:2008.03822*.

[10] Alex Graves, Abdel-rahman Mohamed, and Geoffrey E. Hinton. 2013. Speech recognition with deep recurrent neural networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2013, Vancouver, BC, Canada, May 26-31, 2013*. IEEE, 6645–6649. doi: 10.1109/ICASSP.2013.6638947.

[11] François Hernandez, Vincent Nguyen, Sahar Ghannay, Natalia A. Tomashenko, and Yannick Estève. 2018. TED-LIUM 3: twice as much data and corpus reparation for experiments on speaker adaptation. In *Speech and Computer - 20th International Conference, SPECOM 2018, Leipzig, Germany, September 18-22, 2018, Proceedings (Lecture Notes in Computer Science)*. Alexey Karpov, Oliver Jokisch, and Rodmonga Potapova, (Eds.) Vol. 11096. Springer, 198–208. doi: 10.1007/978-3-319-99579-3_21.

[12] huggingface.co. [n. d.] T5. Retrieved May 6, 2024 from https://huggingface.co/docs/transformers/en/model_doc/t5.

[13] Ana Iglesias, Javier Jiménez, Pablo Revuelta, and Lourdes Moreno. 2016. Avoiding communication barriers in the classroom: the APEINTA project. *Interact. Learn. Environ.*, 24, 4, 829–843. doi: 10.1080/10494820.2014.924533.

[14] Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know. *Trans. Assoc. Comput. Linguistics*, 8, 423–438. doi: 10.1162/TACL_A_00324.

[15] Sushant Kafle and Matt Huenerfauth. 2017. Evaluating the usability of automatically generated captions for people who are deaf or hard of hearing. In *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility, ASSETS 2017, Baltimore, MD, USA, October 29 - November 01, 2017*. Amy Hurst, Leah Findlater, and Meredith Ringel Morris, (Eds.) ACM, 165–174. doi: 10.1145/3132525.3132542.

[16] Saba Kawas, George Karalis, Tzu Wen, and Richard E. Ladner. 2016. Improving real-time captioning experiences for deaf and hard of hearing students. In *Proceedings of the 18th International ACM SIGACCESS Conference on Computers and Accessibility, ASSETS 2016, Reno, NV, USA, October 23-26, 2016*. Jinjuan Heidi Feng and Matt Huenerfauth, (Eds.) ACM, 15–23. doi: 10.1145/2982142.2982164.

[17] Yotaro Kubo, Shigeki Karita, and Michiel Bacchiani. 2022. Knowledge transfer from large-scale pretrained language models to end-to-end speech recognizers. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 8512–8516.

[18] Jinyu Li, Li Deng, Reinhold Haeb-Umbach, and Yifan Gong. 2015. Robust automatic speech recognition: a bridge to practical applications.

[19] Ke Li, Jinyu Li, Guoli Ye, Rui Zhao, and Yifan Gong. 2019. Towards code-switching ASR for end-to-end CTC models. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, United Kingdom, May 12-17, 2019*. IEEE, 6076–6080. doi: 10.1109/ICASSP.2019.8683223.

[20] Chin-Yew Lin. 2004. Rouge: a package for automatic evaluation of summaries. In *Text summarization branches out*, 74–81.

[21] Loren Lugosch, Mirco Ravanelli, Patrick Ignoto, Vikrant Singh Tomar, and Yoshua Bengio. 2019. Speech model pre-training for end-to-end spoken language understanding. In *20th Annual Conference of the International Speech Communication Association, Interspeech 2019, Graz, Austria, September 15-19, 2019*. Gernot Kubin and Zdravko Kacic, (Eds.) ISCA, 814–818. doi: 10.21437/INTERSPEECH.2019-2396.

- [22] Dau-Cheng Lyu, Tien Ping Tan, Engsiong Chng, and Haizhou Li. 2010. SEAME: a mandarin-english code-switching speech corpus in south-east asia. In *11th Annual Conference of the International Speech Communication Association, INTER-SPEECH 2010, Makuhari, Chiba, Japan, September 26-30, 2010*. Takao Kobayashi, Keikichi Hirose, and Satoshi Nakamura, (Eds.) ISCA, 1986–1989. doi: 10.21437/INTERSPEECH.2010-563.
- [23] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2015, South Brisbane, Queensland, Australia, April 19-24, 2015*. IEEE, 5206–5210. doi: 10.1109/ICASSP.2015.7178964.
- [24] Miltiades Papadopoulos and Elaine Pearson. 2008. Accessible lectures: moving towards automatic speech recognition models based on human methods. In *Proceedings of the 10th International ACM SIGACCESS Conference on Computers and Accessibility, ASSETS 2008, Halifax, Nova Scotia, Canada, October 13-15, 2008*. Simon Harper and Armando Barreto, (Eds.) ACM, 273–274. doi: 10.1145/1414471.1414534.
- [25] Przemysław Pietrzak. 2023. Youtube transcript api. Accessed: 2024-06-29. <https://pypi.org/project/youtube-transcript-api/>.
- [26] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA* (Proceedings of Machine Learning Research). Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, (Eds.) Vol. 202. PMLR, 28492–28518. <https://proceedings.mlr.press/v202/radford23a.html>.
- [27] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1, 8, 9.
- [28] Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2021. Fastspeech 2: fast and high-quality end-to-end text to speech. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net. <https://openreview.net/forum?id=piLPYqxtWuA>.
- [29] Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. 2024. A systematic survey of prompt engineering in large language models: techniques and applications. *CoRR*, abs/2402.07927. arXiv: 2402.07927. doi: 10.48550/ARXIV.2402.07927.
- [30] Joonbo Shin, Yoonhyung Lee, and Kyomin Jung. 2019. Effective sentence scoring method using bert for speech recognition. In *Asian Conference on Machine Learning*. PMLR, 1081–1093.
- [31] Nina Shvetsova, Anna Kukleva, Xudong Hong, Christian Rupprecht, Bernt Schiele, and Hilde Kuehne. 2023. Howtocation: prompting llms to transform video annotations at scale. *arXiv preprint arXiv:2310.04900*.
- [32] Hugo Touvron et al. 2023. Llama 2: open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288. arXiv: 2307.09288. doi: 10.48550/ARXIV.2307.09288.
- [33] Joris van Gelder, Irene van Peer, and Dzmitry Aliakseyeu. 2005. Transcription table: text support during meetings. In *Human-Computer Interaction - INTERACT 2005, IFIP TC13 International Conference, Rome, Italy, September 12-16, 2005, Proceedings* (Lecture Notes in Computer Science). Maria Francesca Costabile and Fabio Paternò, (Eds.) Vol. 3585. Springer, 1002–1005. doi: 10.1007/11555261_90.
- [34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.
- [35] Jason Wei et al. 2022. Emergent abilities of large language models. *Trans. Mach. Learn. Res.*, 2022. <https://openreview.net/forum?id=yzkSU5zdwD>.
- [36] World Health Organization. 2024. Deafness and hearing loss. Accessed: 2024-09-30. (2024). <https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss>.
- [37] Liyan Xu, Yile Gu, Jari Kolehmainen, Haidar Khan, Ankur Gandhe, Ariya Rastrow, Andreas Stolcke, and Ivan Bulyko. 2022. Rescorebert: discriminative speech recognition rescoring with bert. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6117–6121.
- [38] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. Mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*. Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tür, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, (Eds.) Association for Computational Linguistics, 483–498. doi: 10.18653/V1/2021.NAACL-MAIN.41.

Received 30 June 2024