MUSCAT: MUltilingual, SCientific ConversATion Benchmark

Supriti Sinhamahapatra[1], **Thai-Binh Nguyen**[1], **Yiğit Oğuz**[1], **Enes Ugan**[1], **Jan Niehues**[1], **Alexander Waibel**[1,2]
[1]Karlsruhe Institute of Technology , [2]Carnegie Mellon University
{supriti.sinhamahapatra, thai-binh.nguyen, enes.ugan, jan.niehues}@kit.edu
yigit.oguz@student.kit.edu, alexander.waibel@cmu.edu

Abstract

The goal of multilingual speech technology is to facilitate seamless communication between individuals speaking different languages, creating the experience as though everyone were a multilingual speaker. To create this experience, speech technology needs to address several challenges: Handling mixed multilingual input, specific vocabulary, and code-switching. However, there is currently no dataset benchmarking this situation. We propose a new benchmark to evaluate current Automatic Speech Recognition (ASR) systems, whether they are able to handle these challenges. The benchmark consists of bilingual discussions on scientific papers between multiple speakers, each conversing in a different language. We provide a standard evaluation framework, beyond Word Error Rate (WER) enabling consistent comparison of ASR performance across languages. Experimental results demonstrate that the proposed dataset is still an open challenge for state-of-the-art ASR systems. The dataset is available in https://huggingface.co/datasets/goodpiku/muscat-eval

Keywords: multilingual, speech recognition, audio segmentation, speaker diarization

1. Introduction

Seamless communication across language boundaries is a long-term dream of mankind. The ultimate goal is to have a natural, multilingual conversation where each participant talks in their favorite language and is able to understand all the other languages. While significant progress has been made in terms of multilingual speech recognition in high resource (Barrault et al., 2023; Liu et al., 2023), as well as low resource settings (Robinson et al., 2025; Li et al., 2025), to the best of our knowledge, currently there is no realistic benchmark to evaluate systems in multilingual dialogue scenario.

To address the growing need for realistic and high-quality multilingual datasets, we present an unique collection of audio recordings designed as a benchmark for automatic speech recognition. In order to build strong systems for this benchmark, several challenges need to be addressed: multilingual speech input, speaker segmentation, audio condition, domain-specific vocabulary, and code-switching.

To collect the data, we setup a discussion between two bi-lingual speakers about scientific publications. Each speaker only speaks one language but understands both languages. While English is the dominant language for scientific communication globally, we aim for AI-based solutions that can facilitate scenarios where researchers can communicate scientific content in their native language without any compromise. To study this, we created a controlled simulation using bilingual speakers.

Our evaluation show limitation in current AI systems. The technology is not yet robust to support multilingual communication in scientific domains. In Figure 1, the upper part illustrates an example scenario of the MUSCAT dataset creation process, in which spontaneous conversations are recorded between two speakers, one speaking English and the other German. The lower part highlights a key challenge faced by state-of-the-art (SOTA) ASR models: their difficulty in accurately detecting language switches in spontaneous multilingual speech. In such cases, the model often either translates the utterance into the language of the preceding context or fails to transcribe certain segments altogether. This leads to an oracle setup for speech translation technology.

The main contributions of this paper are:

- A new benchmark [1] for multilingual conversations.
- A detailed analysis of the challenges of the proposed benchmark.
- State-of-the-art baseline results that highlight the difficulties of the task.

2. Data Collection

We aim to build a high-quality multilingual dataset. In order to achieve this, we first create a conversation setup where the challenges of multilingual, scientific conversations are highlighted. Next, we

[1]https://huggingface.co/datasets/goodpiku/muscat-eval

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 5926–5937
11–16 May 2026.

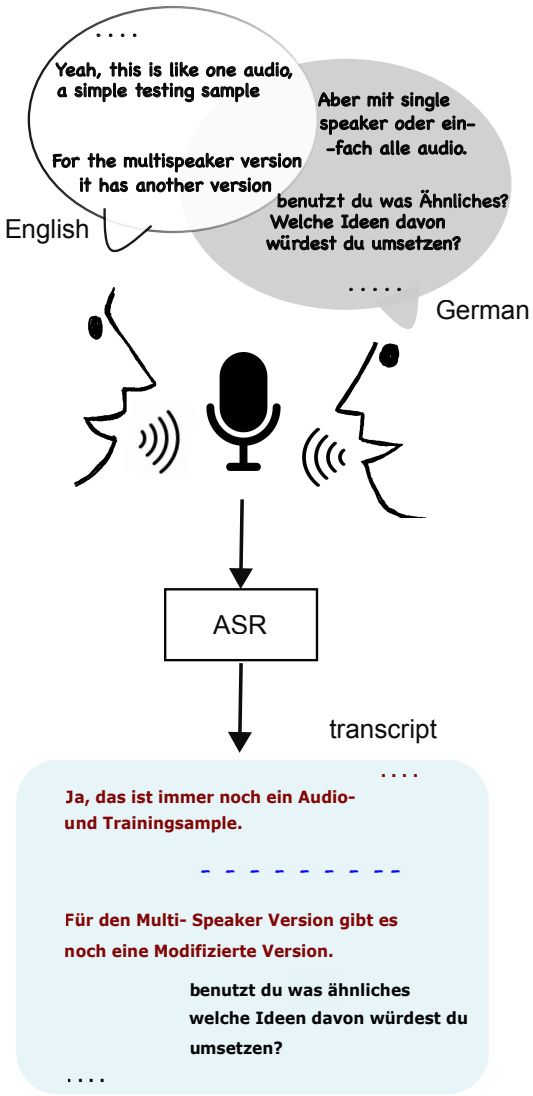

Figure 1: An example illustrating the creation of MUSCAT (upper part of the figure) and the challenges its multilingual diversity poses for state-of-the-art ASR systems (lower part of the figure). The ASR is unable to accurately detect the language switches in a spontaneous conversation denoted by red in the transcript. The blue dashed lines (– – –) represent the part of the conversation that ASR fails to transcribe.

design a recording setup that allows us to investigate the different challenges individually.

2.1. Conversation Setup

Each instance in the dataset contains a conversation between a pair of speakers over a scientific paper. The speaker possess prior knowledge of the paper being discussed. We create the oracle situation for speech translation in this scenario by having the speakers converse in two different languages. To carry out the conversation, the speakers need to be fluent in both languages. For our case, the speakers engaging in natural conversations are each fluent in English at a C1 level and are native speakers of the other language.

This unique setup allows for meaningful exchanges where speakers fully comprehend one another but respond solely in one of the two languages. These conversations offer paired speech[2] and transcripts for language pairs like English-German, English-Vietnamese, English-Chinese, and English-Turkish.

2.2. Recording Setup

The challenges associated with ASR vary depending on the recording environment. In order to evaluate different conditions jointly, we synchronously record the conversations with three different devices. The first device is Meeting Owl 3, which is a popular video conferencing system that captures 360° video and audio. The second device, the ReSpeaker USB Microphone, is a compact array microphone designed for high-fidelity, multi-directional audio capture. The third, Aria smart glasses by Meta, function as a wearable device that records first-person audio along with audio from the speaker's environment.

The Meeting Owl 3, referred to as simply OWL in the rest of the paper, was connected to a laptop via USB, and recordings were made using OBS Studio 3. The ReSpeaker USB Microphone Array was paired with a Raspberry Pi 3, also connected via USB, to provide an additional audio source. For brevity, we refer to this setup as Pi for the remaining of the paper. Finally, the Aria glasses by Meta were worn by one of the speakers and used to record the entire conversation from their perspective. Since Aria can be worn by only one person during recording, we randomly selected one speaker. This results in three German, one Chinese, one Vietnamese, and one English speaker wearing Aria. The other audio recording devices were kept approximately at the middle and equidistant from the speakers.

The recordings are made at a sampling rate of 44.1kHz using multiple microphones as mentioned above. After data collection, the recordings from all devices were manually aligned using Audacity to ensure they were perfectly synchronized. This combination of devices and meticulous alignment ensures that the dataset captures a wide range of audio perspectives, adding depth and variety. All software used were the latest versions available at the time of recording. To ensure minimum possible interference from external sound, an appropriately secluded room is used for the recordings.

[2]All participants gave their consent for their voices to be recorded and used for research purposes

3. Human Annotation

We annotate the collected data to be used as a benchmark for state-of-the-art ASR systems. In a first step, we perform a manual segmentation of the audio recordings which serves as the oracle to evaluate and compare two automatic segmentation approaches. Next, we create the multilingual transcripts of the audio.

3.1. Manual Segmentation

We perform the manual segmentation guided by two constraints. First, since each of our recordings consists of conversations in two different languages, we prioritize language-specific segmentation. This ensures that each segment comprises recordings in a single language. Second, to support optimal model performance, we limit each segment to a maximum duration of 30 seconds. The manual segmentation process is conducted using Label Studio (Studio, 2023), an open-source data annotation tool.

3.2. Human post-editing

Starting from the derived manual segments, we follow a two-step process to obtain the human transcript of our dataset. Firstly, to ease manual transcription of the recordings, we use a state-of-the-art ASR model, Whisper (Radford et al., 2023) for automatic transcription of the language specific segments. As a second step, the respective speaker manually corrects any mistakes made by ASR model, ensuring high quality transcription. This strategy was adopted to overcome the challenge of finding external annotators who possessed both fluency in the specific languages and familiarity with the complex scientific discourse discussed in the papers. Using the speakers for such task ensured that technical terms and domain-specific context were annotated accurately.

During the annotation process, we frequently observe instances of code-switching within the recorded speech. Since mixed-language utterances are known to pose particular challenges for current speech recognition systems (Klejch et al., 2021; Hamed et al., 2022; Ugan et al., 2025b), annotators were additionally instructed to mark all words belonging to the embedded language whenever code-switching occurred.

4. MUSCAT Dataset

The MUSCAT dataset consists of multilingual conversations of six recordings across eleven speakers. Each recording is between a pair of speakers, and there exists one speaker who is present in two recordings. All six recordings have at least one English speaker, while the other speaks one of the languages from German, Turkish, Chinese, and Vietnamese. Of the six recordings, half of the conversations are between a pair of English and German speakers, while the other half is between English and the remaining languages. We maintain gender diversity among the speakers in the dataset. To this end, among the eleven speakers of the MUSCAT dataset, six speakers are male and the remaining five speakers are female.

In order to evaluate different challenges related to this benchmark, we provided 6 different variations of the dataset. First, three different recordings with the different devices per speaker. Secondly, for each conversation, segmented and unsegmented version of the audio recording is available.

Table 1 summarizes the main aspects of the dataset. The total duration of our dataset is approximately 65 minutes, of which English-Vietnamese conversation comprises of 17 minutes, whereas English-Chinese, English-Turkish, and English-German conversations comprise 15, 12, and 21 minutes, respectively. Words spoken by each speaker are attributed to the word count of their respective language, which in total is 9,066 words.

Table 1: Dataset Statistics

Recordings	Languages	Total Duration	Total Word Counts
Recording 1	English	4.69 mins	463
	German	1.92 mins	288
Recording 2	English	1.39 mins	162
	German	2.74 mins	427
Recording 3	English	7.51 mins	1344
	Turkish	3.94 mins	447
Recording 4	English	11.90 mins	1362
	Chinese	2.79 mins	623
Recording 5	English	7.47 mins	972
	German	3.00 mins	426
Recording 6	English	10.04 mins	1489
	Vietnamese	6.83 mins	1063
Total		64.22 mins	9,066

5. Baseline

This section outlines the baseline configuration adopted in our experiments, detailing the ASR models used and the segmentation strategies applied during pre-processing.

5.1. ASR Models

Our goal is to evaluate the performance of SOTA ASR models on the MUSCAT dataset. To this end, we employ four SOTA models, Whisper, SALMONN,Phi-4 Multimodal and Wav2Vec2, and asses thequality of their generated transcriptions.These models represent diverse ASR

paradigms, including encoder–decoder architectures (Whisper), multimodal large language models (SALMONN and Phi-4 Multimodal), and CTC-based systems (Wav2Vec2).

Whisper Whisper is a transformer-based encoder-decoder model developed by OpenAI, primarily designed for automatic speech recognition (ASR) and speech translation tasks (Radford et al., 2023). It has been trained on approximately 680k hours of speech data collected from the internet. The model's encoder processes the input speech to generate audio features, which are then passed to the decoder. The decoder, using these audio features along with positional encodings, produces the corresponding transcription. Whisper also incorporates a set of context tokens that guide the model by specifying the language, the task to be performed, and the start and end points of the transcription.

SALMONN The SALMONN model, developed by Tsinghua University and ByteDance (Tang et al., 2023), extends the capabilities of Large Language Models (LLMs), such as Vicuna (Chiang et al., 2023), to directly perceive and interpret general audio inputs. This enhancement allows LLMs to perform competitively across a range of speech and audio processing tasks. SALMONN integrates information from two specialized encoders, Whisper (Radford et al., 2023) for speech and BEATs (Chen et al., 2022) for general audio using a window-level Q-Former module (Zhang et al., 2024). The resulting augmented audio tokens are aligned with the LLM's internal representations, enabling seamless multi-modal understanding.

Phi-4-multimodal Phi-4-multimodal (referred to as Phi) is a 5.6B-parameter instruction-tuned multimodal transformer developed by Microsoft. It is designed for unified processing of text, image, and audio inputs, enabling it to handle tasks across vision-language, vision-speech, and speech-language domains. The model supports a context length of up to 128K tokens and utilizes 32 transformer layers equipped with Grouped Query Attention (GQA) (Ainslie et al., 2023) for efficient long-context processing. Vision and audio modalities are mapped into the text embedding space via two-layer multilayer perceptrons (MLPs). Phi demonstrates strong performance on a wide range of multilingual and multi-modal benchmarks.

wav2vec2 wav2vec2 (Baevski et al., 2020) is a self-supervised learning framework designed to learn speech representations directly from raw audio. The model includes a convolutional feature encoder that converts audio into latent representations, followed by a transformer network that captures context over time.

We use the wav2vec2-large-960h-lv60-self model from Facebook, which is trained on 960 hours of audio for performing experiments with the english audios. For other languages, including German, Turkish, Chinese and Vietnamese we employ the wav2vec2-large-xlsr-53 model fine-tuned in a supervised manner on the respective Common Voice datasets (Ardila et al., 2019). The model is trained separately for each language using a Connectionist Temporal Classification (CTC) loss (Baevski et al., 2020; Graves et al., 2006) to perform ASR (Grosman, 2021c; ozcangundes, 2021; Grosman, 2021a; von Platen, 2021).

5.2. Segmentation

For the condition using unsegmented audio, segmentation is necessary because some of the SOTA ASR models cannot handle long audios. Feeding them longer recordings, such as 10-minute audio files, would likely degrade transcription accuracy. By breaking the recordings into shorter segments, we aim to align the input format with model training conditions, improving overall transcription quality.

Therefore, we process the data using the following three segmentation approaches among which two are automatic segmentation: *Segmented Hybrid Audio Segmentation (SHAS)* (Tsiamas et al., 2022), a commonly used segmenter for live transcriptions; *PyanNet segmentation* (Bredin and Laurent, 2021; Bain et al., 2023), a segmentation trained on voice activity detection and further fine-tuned for speaker diarization; *Human segmentation* serves as a ground truth segmentation. We provide description on the process of human annotation in Section 3.1.

SHAS detects pauses and other acoustic cues to identify natural breakpoints in speech, ensuring that segments correspond to meaningful units of conversation. This approach helps preserve the conversational structure while generating smaller, more manageable audio segments for further processing.

In order to achieve a segmentation more aligned with our ground truth as alternative, we use PyanNet segmentation (Bredin and Laurent, 2021). To further improve alignment to our scenario, we use a fine-tuned version, of the model, which was trained to track up to three speakers simultaneously in noisy scenarios. Inspired by WhisperX (Bain et al., 2023), we enforce length constraints through post-processing, where overly long segments are split at their lowest-confidence point, while overly short ones are merged with neighboring segments until the desired segment's duration is achieved.

6. Evaluation

We evaluate SOTA ASR systems to establish a baseline performance on this dataset. Through this analysis under varying segmentation and transcription conditions, we identify key challenges that the dataset presents for current ASR technology.

Metrics Word Error Rate (WER) is a common metric used to evaluate the accuracy of ASR systems. It measures how much the transcribed text deviates from the ground truth by computing the number of errors made during transcription, giving equal importance to every word in the transcript. Unlike the other languages in our dataset, Chinese is not a whitespace-separated language. We use jieba[3], a Python Chinese word segmentation tool for segmenting the Chinese text into words.

Our interest also lies in investigating model performances on special words. Additionally, we observe frequent occurrences of code-switching within the recorded speech and evaluate the model performance. The following provides details of this evaluation:

- *Domain-specific WER*: Our dataset comprises scientific conversations in which speakers frequently use domain-specific words. We measure the quality of the domain-specific words with respect to the reference and the hypothesis similar to recall and precision (Sinhamahapatra and Niehues, 2025). First, we investigate how many domain-specific words in the reference are missed or wrongly transcribed by the model, by aggregating the deletion and the substitution counts, and dividing it by the total occurrences of domain-specific words in the manual transcript. In this paper, we calculate a reference-centric WER metric $\text{WER}_{t_{ref}}$.

  $$\text{WER}_{t_{ref}} = \frac{|\text{substituted}+\text{deleted}|}{|\text{recognized} +\text{substituted}+\text{deleted}|}$$

  Next, we calculate the $\text{WER}_{t_{hyp}}$ to evaluate how many domain-specific words in the model's output are incorrectly transcribed.

  $$\text{WER}_{t_{hyp}} = \frac{|\text{substituted}+\text{inserted}|}{|\text{recognized} +\text{substituted}+\text{inserted}|}$$

- *Code-Switching Performance*: In our multilingual dataset, comprising spontaneous discussions on scientific topics, we observe code-switching behavior of speakers. Non-English speakers often incorporate English words. One reason for this tendency is the absence or limited familiarity of equivalent terminology in their native language. To assess code-switching performance, we employ the recently proposed Point-of-Interest Error Rate (PIER) metric (Ugan et al., 2025a). PIER is a variant of the traditional Word Error Rate (WER), designed specifically to measure ASR performance on code-switched segments by focusing on errors aligned with points of interest that is, embedded-language words. In our case, these annotated English words served as points of interest for PIER computation.

[3]https://github.com/fxsjy/jieba

6.1. Evaluation of Different Conditions

As an initial step, we compute the multilingual WER across the entire benchmark under nine distinct recording conditions (Table 2). These conditions are derived from the combination of three recording devices and three segmentation strategies: two automatic methods, PyanNet-based diarization and SHAS, and one manual segmentation approach, which serves as the oracle. This experiment is conducted using the SOTA ASR model Whisper.

The results in Table 2 demonstrates that our benchmark is challenging, with WERs reaching up to 31%. We observe that one of the main challenges is the segmentation of multilingual audio.

When using SHAS segments, WERs range between 27% and 31%. In contrast, applying the PyanNet speaker diarization method reduces WERs to approximately 21–23%. The lowest error rates, between 12% and approximately 18%, are achieved by Whisper using our oracle setups. Finally, we observe notable performance variations across the different audio recording conditions, underscoring the impact of recording quality on ASR accuracy.

6.2. Evaluation of Models for Multilingual Transcription

Table 3 summarizes our findings of this experiment. The main challenge of the MUSCAT dataset for the models is its multilingual composition. As a result, we evaluate the ASR performance on our dataset. Since it is only possible to separate the languages in the manual segmentation, we use these segments for the experiment. Furthermore, to ensure comparable conditions for both speakers, we focus on segments from the OWL recordings.

Since SALMONN is trained only on English, the scores for rest of the languages are not considered in this paper. The Phi model supports English, German, and Chinese, but does not handle the other languages in our dataset effectively. Among all tested models, Whisper achieves the lowest WER for each language, showing strong multilingual capabilities. Table 3 also demonstrates results for the wav2vec2 model which comparatively has higher WERs than the other models considered for this experiment.

Table 2: Overview table with WER on manual (oracle), PyanNet and SHAS segmented audio recordings across the three devices on Whisper

	Aria			OWL			Pi		
segmentation	**manual**	**PyanNet**	**SHAS**	**manual**	**PyanNet**	**SHAS**	**manual**	**PyanNet**	**SHAS**
all recordings	12.12	23.19	27.46	12.98	22.78	31.16	18.65	21.89	28.16

Table 3: Evaluation of models for multilingual transcription using manual segments of the OWL recordings. A dash (-) represents scores not considered for the languages, as the corresponding models were not trained on them.

Language	Whisper	SALMONN	Phi	wav2vec2
English	10.32	17.17	16.34	31.74
German	12.22	-	15.72	27.93
Turkish	15.96	-	-	71.24
Chinese	14.95	-	14.11	53.26
Vietnamese	24.18	-	-	81.84

Table 4: Whisper WER across different devices on the manually segmented audio. We mark the languages recorded using Aria glasses in the table with *(Aria)*.

Language	Aria	OWL	Pi
English	9.68	8.15	12.19
English *(Aria)*	15.06	21.21	39.06
German *(Aria)*	8.71	12.22	14.97
Turkish	16.63	15.96	23.50
Chinese *(Aria)*	9.26	14.95	18.74
Vietnamese *(Aria)*	26.25	24.18	22.95

Although Whisper is performing best on English, we still observe strong performance on all languages with Vietnamese being the most difficult. Similarly, wav2vec2 struggles the most for Vietnamese, Turkish, and Chinese among the five languages. This highlights the importance of having a diverse set of languages in MUSCAT to test the robustness of ASR systems.

6.3. Evaluation of Recording Devices

In this section we analyse the challenges concerning different recoding quality as described in Section 2.2. For this experiment, we use the SOTA model Whisper and summarize the results in Table 4. The table shows the WER scores of the transcripts produced by the model across different recording devices. Of three devices, the quality of recordings by Aria may be impacted if the speaker is not wearing it. As a result, we check the model performance separately with and without wearing Aria. To make the distinction explicit, we mark the languages those were recorded with Aria glass in table 4 with *(Aria)*. For example, the first row of the table shows scores of the recordings when a speaker is not wearing Aria and is speaking English. In contrast, the second row of the table contains scores for recordings where a speaker is speaking English while wearing Aria.

The table highlights that the challenges in the different audio conditions vary. If the speaker wears the glass (second row), the Aria microphone clearly leads to the best performance, with quality gains up to 29% for English (Aria) when compared to the OWL score. In contrast, for English (first row) where the Aria is not worn by the speaker, the WER is relatively higher than OWL. This indicates that additional research is needed to also perform good quality ASR without close-source microphones. In addition, Aria microphones perform surprisingly well also for the speaker not wearing the Aria glass (first and fourth row of Table 4). For these cases we see minor decrease in performance compared to the OWL recordings.

Finally, although both microphones (OWL and Pi) are positioned similarly, there is a clear performance gap. This also motivates additional research on high-quality ASR with lower-quality microphones.

6.4. Evaluation of Segmentation Approaches

We also investigate the impact of unsegmented audio, which is a more realistic condition, on the final ASR performance. Table 6 presents results obtained with two automatic segmentation methods and a manual segmentation, which serves as the oracle (described in Section 5.2). The evaluation is carried out using the Whisper model, which exhibits the strongest multilingual performance across all five dataset languages (Section 6.2).

We find that using SHAS segmentation, the model often fails to separate languages properly, resulting in mixed-language segments. This leads to almost three times higher WER for all language pairs, compared to manual segmentation, as the Whisper model struggles with detecting language to transcribe properly within a single segment. For instance, the SHAS score for English-Turkish recordings is 57.41 which is three times more the manual score 19.89.

In contrast, using the PyanNet segments which are trained for speaker diarization, leads to fewer segments with speaker overlap. As a result, the

Table 5: An example of ASR performance using two types of automatic segmentation. With the red color we indicate the substitutions, words omitted from the transcript are shown with gray, words in blue are inserted texts and, marked with green are the parts where the model is unable to switch to the respective language.

Reference transcript	ASR on SHAS segments	ASR on PaynNet segments
Okay, I have another question. Is this model have the similar architecture as the chatGPT model? Mehr oder weniger. Es ist ein Transformer, aber es ist so ein bisschen wie bei PaLM, dass die MLP-Schichten und die Attention-Schichten parallel zueinander sind statt sequenziell. So it's not autoregressive. It's a parallel structure? No, no, this is, das ist das ist nur innerhalb von der von einem Transformer-Block.	Okay, Ich habe noch eine Frage. Ist dieses Modell mit der gleichen Architektur wie das HHGPD Modell? Mehr oder weniger. Es ist ein Transformer, aber es ist so ein bisschen wie bei Plum, dass die MLP Schichten und die Attention Schichten parallel zueinander sind, statt sequenziert. So it's not autoregressive, it's a parallel structure? No, no, this is, das ist only inside of one transformer block.	Okay, I have another question. Does this model have the similar architecture as the chatGPT model? mehr oder weniger. Es ist ein Transformer, aber es ist so ein bisschen wie bei Plum, dass die MLP-Schichten und die Attention-Schichten parallel zueinander sind statt sequenziell. So it's not autoregressive. It's a parallel structure? Nein, nein, nein, this is, das ist nur innerhalb von der von einem Transformer-Block.

Table 6: Model WER across all languages pairs using OWL SHAS automatic segments and the manual segments

	Whisper		
Language pairs	**Manual WER**	**PyanNet WER**	**SHAS WER**
English-German	10.88	20.57	23.93
English-Turkish	19.89	32.53	57.41
English-Chinese	8.16	12.89	19.29
English-Vietnamese	12.89	24.10	31.19

segments are more likely to be language homogeneous, aligning better with the pre-trained distribution of single-language utterances. Consequently, this segmentation approach yields improved ASR performance to SHAS-based automatic segmentation. The second row of Table 6 shows one such example where PyanNet score is lower compared to the SHAS score. This highlights the potential of more advanced segmentation techniques to enhance transcription quality in multilingual settings. The findings emphasize the importance of multilingual datasets such as MUSCAT to benchmark segmentation techniques in ASR.

Table 5 presents an excerpt of a conversation in German and English from MUSCAT, in which two speakers discuss a scientific paper (Gunasekar et al., 2023). The first column contains the manually transcribed reference, while the second and third columns show automatic transcriptions generated by Whisper, based on SHAS and PyanNet segmentation, respectively. The example demonstrates that the model often fails to detect language switches in the SHAS-based segments, often producing translations instead of transcriptions. In contrast, the PyanNet-based segmentation partially mitigates this issue, though it occasionally omits parts of the conversation.

Table 7: WER and WER-Term on domain-specific words. Whisper, SALMONN, Phi and wav2vec2 are evaluated using OWL English manual segments.

	Whisper	**SALMONN**	**Phi**	**wav2vec2**
Total Counts	55	55	55	55
Recognized	33	24	19	4
Non Recognized	22	31	36	51
WER	10.32	17.17	16.34	31.74
$WER_{t_{ref}}$	35.08	46.87	59.67	77.99
$WER_{t_{hyp}}$	28.33	46.87	59.67	77.46

6.5. Model Performance on Special Words

Our dataset contains multilingual recordings of discussions over scientific papers. Such papers contain technical terms often not found is general discourse, referred to as special words in this paper. We want to measure the model performance in transcribing such special words, as previous research (Wang et al., 2024) (Yang et al., 2024) has highlighted that such words pose particular challenges for ASR systems. Since the introduced special words in scientific papers are often in English, we focus this experiment only on the English recordings using OWL.

For each recording, we extract domain-specific words from the scientific paper that the speakers use to discuss. To identify these words, we exclude all words found in a general-purpose dataset (Di Gangi et al., 2019), from all the words of the paper. The remaining words are considered as the special words associated with the paper. The *Total Counts* score in Table 7 represents the number of these special words occurring in the English portion of the recordings. The second and third rows of the table indicate the number of instances where the

models successfully recognize or failed to recognize these words. Finally, we compute the WER on the full vocabulary and the domain-specific WER on the special terms following the evaluation metric outlined in Section 6.

We evaluate the performance of four ASR models Whisper, SALMONN, Phi, and wav2vec2 on domain-specific words. For SALMONN, Phi, and wav2vec2, the $\text{WER}_{t_{ref}}$ and $\text{WER}_{t_{hyp}}$ are approximately 2.3 to 2.7 times higher than the overall WER across all words. In contrast, Whisper exhibits $\text{WER}_{t_{ref}}$, approximately 3.5 times higher than its overall WER, while its $\text{WER}_{t_{hyp}}$ scores are approximately 19% lower than the corresponding $\text{WER}_{t_{ref}}$ scores. These results illustrate the difficulty current ASR models face when transcribing scientific and technical terms, underscoring the value of the MUSCAT dataset for benchmarking scientific transcription performance.

Table 8: PIER ↓ on code switched tokens. Whisper, SALMONN, Phi and wav2vec are evaluated using OWL English manual segments.

Language	Whisper	SALMONN	Phi	wav2vec2
German	39.29	57.14	64.29	116.1
Turkish	38.46	100.0	100.0	53.85
Chinese	77.8	66.7	77.8	88.9
Vietnamese	44.76	124.76	262.86	102.91

6.6. Model Performance on Code-Switched Words

In the MUSCAT dataset, we observe conversational code-switching, where non-English speakers frequently incorporate English words into their speech. Code-switching ASR remains a well-known challenge for current models; therefore, we evaluate the code-switching capabilities of the aforementioned state-of-the-art ASR systems on our dataset using the OWL recording setup.

Our analysis reveals varying degrees of code-switching across languages: Chinese contains the fewest English insertions (9), while Vietnamese exhibits the most (103). Turkish and German speech include 17 and 54 instances, respectively.

As shown in Table 8, all models perform significantly worse on code-switched words than on general speech (see Table 3). Although the Phi-4 model achieves a WER comparable to Whisper, its PIER score more than doubles, indicating increased difficulty in recognizing embedded English words. The Wav2Vec2 model, which already exhibited the weakest overall performance, remains unsatisfactory under code-switching conditions as well. Overall, we observe that code-switching, while still challenging, appears to be handled slightly better in German speech, possibly due to the linguistic relatedness between German and English. In contrast, Whisper struggles most with Chinese code-switching, whereas multimodally pre-trained models such as SALMONN and Phi-4 perform relatively better on this type of data, likely benefiting from the broader linguistic and acoustic diversity encountered during large-scale pre-training.

These findings reinforce that code-switching remains a major limitation for current ASR systems and underscore the importance of multilingual, domain-specific datasets such as MUSCAT in advancing research on this challenging phenomenon.

7. Related Work

Our work presents a novel dataset that bridges the gap between conversational, multilingual, and academic domains. Existing general-purpose conversational datasets such as MultiWOZ (Budzianowski et al., 2018), DialoGPT(Zhang et al., 2019), (Li et al., 2017) and ConvAI2(Dinan et al., 2020), primarily focus on casual dialogue, including structured interactions or discussions extracted from platforms like Reddit. Other speech datasets include the AMI Meeting Corpus(Kraaij et al., 2005) which consists of meeting recordings and DIPCO(Van Segbroeck et al., 2019), a dataset with natural conversation around a dinner table. Both of these speech corpora consist of speech in English. With respect to all these datasets, our work contributes to the field of academic conversational datasets, specifically including one-to-one discussions about scientific papers from known conferences.

Each of the platforms like arXiv, PubMed, and Semantic Scholar primarily contain scientific papers, articles in the form of written text, while our focus is on multilingual speech data. Similarly, multilingual dialogue datasets like FLoRes-101(Goyal et al., 2022) and CoVoST(Wang et al., 2020), lack domain-specificity and are mostly text-based. There exist code-switching datasets where multiple languages can be present in the audio such as Arzen(Hamed et al., 2020), DECM(Ugan et al., 2024), SEAME(Lyu et al., 2010).

Recent advancements in multilingual speech processing have shifted toward spontaneous, multi-party conversational environments. A notable development is the DISPLACE 2024 and 2025 challenge corpora (Kalluri et al., 2024), which provide over 150 hours of multi-speaker, multi-lingual data specifically annotated for speaker and language diarization in overlapping speech scenarios. Furthermore, the MLC-SLM (Multilingual Conversational Speech Language Model) corpus (Mu et al., 2025) introduces a large-scale, 1600-hour benchmark that addresses the complexities of turn-taking and code-switching across 11 languages. These

datasets complement established benchmarks like ML-SUPERB 2.0 (Shi et al., 2024), which expands evaluation to 142 languages to test the cross-lingual generalization of foundational speech models. In addition to the above mentioned challenge-driven datasets, recent large-scale efforts such as SwitchLingua (Xie et al., 2025) have significantly expanded the diversity of conversational corpora. Introduced as a comprehensive multi-ethnic benchmark, SwitchLingua provides over 80 hours of audio from 174 bilingual speakers across 12 languages. In contrast, our dataset expands the coverage of multilingual conversational data in speech where each instance includes dialogues from a discussion on a scientific paper between two speakers speaking separate languages.

8. Conclusion

This paper proposes a novel multilingual dataset to evaluate current ASR systems. Our dataset encompasses scientific conversations in five languages, including English, German, Chinese, Turkish, and Vietnamese. Each conversation consists of a paired speech in two languages, one of which is always English, while the other is one of the four remaining languages.

We perform detailed evaluations on several key aspects of speech recognition using various ASR models, including analysis on different recording devices, evaluating across languages, verifying model capabilities on segmentation, and investigating model performances on domain-specific and code-switched words.

Experimental results from the MUSCAT dataset show that current SOTA ASR systems still face major challenges in handling natural, multilingual scientific discussions. Specifically, our evaluation indicates these models difficulty in accurately detecting when a speaker switches languages during a conversation. When these switches happen, the systems often make mistakes by either translating the speech into the previous language used or completely failing to transcribe the audio segments. Furthermore, the results show that current technology is not yet robust enough to handle the combination of specialized scientific vocabulary, code-switched words or phrases, and different audio conditions found in real-world expert dialogues.

9. Limitation

While the MUSCAT dataset provides a novel benchmark for evaluating multilingual scientific conversations, several limitations must be acknowledged. First, the overall scale of the corpus is relatively small, comprising approximately 65 minutes of audio and 9,066 words. Second, although the dataset encompasses five distinct languages, there is an imbalance in language distribution, with English dominating the conversations. Every recorded interaction involves at least one English speaker, resulting in a word count that is significantly skewed toward English. This asymmetry naturally stems from the domain of the dataset, since the dialogues are entirely based on scientific papers, the speakers often find it more comfortable and precise to articulate complex technical concepts in English. Finally, our baseline evaluation was inherently constrained by the language capabilities of certain state-of-the-art models; for instance, SALMONN is trained only on English , and Phi-4-multimodal primarily supports English, German, and Chinese.

10. Acknowledgement

This work was supported by the European Union's Horizon Europe Framework Programme under grant agreement No. 101213369, project DVPS (Diversibus Viis Plurima Solvo).

Additional support was provided by KiKIT (Pilot Program for Core-Informatics at KIT) of the Helmholtz Association.

We also acknowledge the use of the HoreKa supercomputer, funded by the Ministry of Science, Research, and the Arts of Baden-Württemberg, and by the Federal Ministry of Education and Research.

11. Bibliographical References

Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, et al. 2025. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*.

Joshua Ainslie, James Lee-Thorp, Michiel De Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. 2023. Gqa: Training generalized multi-query transformer models from multi-head checkpoints. *arXiv preprint arXiv:2305.13245*.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. 2019. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech

representations. *Advances in neural information processing systems*, 33:12449–12460.

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. Whisperx: Time-accurate speech transcription of long-form audio. *INTERSPEECH 2023*.

Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim, et al. 2023. Seamless: Multilingual expressive and streaming speech translation. *arXiv preprint arXiv:2312.05187*.

Hervé Bredin and Antoine Laurent. 2021. End-to-end speaker segmentation for overlap-aware resegmentation. In *Interspeech 2021*, pages 3111–3115.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Inigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. Multiwoz–a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. *arXiv preprint arXiv:1810.00278*.

Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, and Furu Wei. 2022. Beats: Audio pre-training with acoustic tokenizers. *arXiv preprint arXiv:2212.09058*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6.

Steven B. Davis and Paul Mermelstein. 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 28(4):357–366.

Mattia A. Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019. MuST-C: a Multilingual Speech Translation Corpus. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2012–2017, Minneapolis, Minnesota. Association for Computational Linguistics.

Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, et al. 2020. The second conversational intelligence challenge (convai2). In *The NeurIPS'18 Competition: From Machine Learning to Intelligent Conversations*, pages 187–208. Springer.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376.

Jonatas Grosman. 2021a. Fine-tuned XLSR-53 large model for speech recognition in Chinese. https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-chinese-zh-cn.

Jonatas Grosman. 2021b. Fine-tuned XLSR-53 large model for speech recognition in English. https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-english.

Jonatas Grosman. 2021c. Fine-tuned XLSR-53 large model for speech recognition in German. https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-german.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, et al. 2023. Textbooks are all you need. *arXiv preprint arXiv:2306.11644*.

Injy Hamed, Pavel Denisov, Chia-Yu Li, Mohamed Elmahdy, Slim Abdennadher, and Ngoc Thang Vu. 2022. Investigations on speech recognition systems for low-resource dialectal arabic–english code-switching speech. *Computer Speech & Language*, 72:101278.

Injy Hamed, Ngoc Thang Vu, and Slim Abdennadher. 2020. Arzen: A speech corpus for code-switched egyptian arabic-english. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4237–4246.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning –*

Data Mining, Inference, and Prediction. Springer, New York.

Robert Jones, Firstname2 Lastname2, and Firstname3 Lastname3. 2022. An excellent paper introducing the ABC toolkit. In (Smith et al., 2022), pages 105–109.

Shareef Babu Kalluri, Prachi Singh, Pratik Roy Chowdhuri, Apoorva Kulkarni, Shikha Baghel, Pradyoth Hegde, Swapnil Sontakke, SR Prasanna, Deepu Vijayasenan, Sriram Ganapathy, et al. 2024. The second displace challenge: Diarization of speaker and language in conversational environments. *arXiv preprint arXiv:2406.09494*.

Ondřej Klejch, Electra Wallington, and Peter Bell. 2021. The cstr system for multilingual and code-switching asr challenges for low resource indian languages. In *Interspeech 2021: The 22nd Annual Conference of the International Speech Communication Association*, pages 2881–2885. International Speech Communication Association.

Wessel Kraaij, Thomas Hain, Mike Lincoln, and Wilfried Post. 2005. The ami meeting corpus. In *Proc. International Conference on Methods and Techniques in Behavioral Research*, pages 1–4.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957*.

Zhaolin Li, Yining Liu, Danni Liu, Tuan Nam Nguyen, Enes Yavuz Ugan, Tu Anh Dinh, Carlos Mullov, Alexander Waibel, and Jan Niehues. 2025. Kit's low-resource speech translation systems for iwslt2025: System enhancement with synthetic data and model regularization. *arXiv preprint arXiv:2505.19679*.

Danni Liu, Thai Binh Nguyen, Sai Koneru, Enes Yavuz Ugan, Ngoc-Quan Pham, Tuan-Nam Nguyen, Tu Anh Dinh, Carlos Mullov, Alexander Waibel, and Jan Niehues. 2023. Kit's multilingual speech translation system for iwslt 2023. *arXiv preprint arXiv:2306.05320*.

Dau-Cheng Lyu, Tien Ping Tan, Engsiong Chng, and Haizhou Li. 2010. Seame: a mandarin-english code-switching speech corpus in south-east asia. *Interspeech*, pages 1986–1989.

Roger K. Moore and Lucy Skidmore. 2019. On the use/misuse of the term 'Phoneme'. In *Proc. INTERSPEECH 2019 – 20th Annual Conference of the International Speech Communication Association*, pages 2340–2344, Graz, Austria.

Bingshen Mu, Pengcheng Guo, Zhaokai Sun, Shuai Wang, Hexin Liu, Mingchen Shao, Lei Xie, Eng Siong Chng, Longshuai Xiao, Qiangze Feng, et al. 2025. Summary on the multilingual conversational speech language model challenge: Datasets, tasks, baselines, and methods. *arXiv preprint arXiv:2509.13785*.

ozcangundes. 2021. Fine-tuned XLSR-53 large model for speech recognition in turkish. https://huggingface.co/ozcangundes/wav2vec2-large-xlsr-53-turkish.

Lawrence R. Rabiner. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

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*, pages 28492–28518. PMLR.

Nathaniel Romney Robinson, Niyati Bafna, Xiluo He, Tom Lupicki, Lavanya Shankar, Cihan Xiao, Qi Sun, Kenton Murray, and David Yarowsky. 2025. Jhu iwslt 2025 low-resource system description. In *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, pages 315–323.

Jiatong Shi, Shih-Heng Wang, William Chen, Martijn Bartelds, Vanya Bannihatti Kumar, Jinchuan Tian, Xuankai Chang, Dan Jurafsky, Karen Livescu, Hung-yi Lee, et al. 2024. Ml-superb 2.0: Benchmarking multilingual speech models across modeling constraints, languages, and datasets. *arXiv preprint arXiv:2406.08641*.

Supriti Sinhamahapatra and Jan Niehues. 2025. Do slides help? multi-modal context for automatic transcription of conference talks.

Jane Smith, Firstname2 Lastname2, and Firstname3 Lastname3. 2022. A really good paper about Dynamic Time Warping. In *Proc. INTERSPEECH 2022 – 23rd Annual Conference of the International Speech Communication Association*, pages 100–104, Incheon, Korea.

Label Studio. 2023. Label studio: Open source data labeling platform.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2023. Salmonn: Towards generic hearing abilities for large language models. *arXiv preprint arXiv:2310.13289*.

Ioannis Tsiamas, Gerard I. Gállego, José A. R. Fonollosa, and Marta R. Costa-jussà. 2022.

Shas: Approaching optimal segmentation for end-to-end speech translation.

Enes Yavuz Ugan, Christian Huber, Juan Hussain, and Alexander Waibel. 2022. Language-agnostic code-switching in sequence-to-sequence speech recognition. *arXiv preprint arXiv:2210.08992*.

Enes Yavuz Ugan, Ngoc-Quan Pham, Leonard Bärmann, and Alex Waibel. 2025a. Pier: A novel metric for evaluating what matters in code-switching. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Enes Yavuz Ugan, Ngoc-Quan Pham, and Alexander Waibel. 2024. DECM: Evaluating bilingual ASR performance on a code-switching/mixing benchmark. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4468–4475, Torino, Italia. ELRA and ICCL.

Enes Yavuz Ugan, Ngoc-Quan Pham, and Alexander Waibel. 2025b. Weight factorization and centralization for continual learning in speech recognition. *arXiv preprint arXiv:2506.16574*.

Maarten Van Segbroeck, Ahmed Zaid, Ksenia Kutsenko, Cirenia Huerta, Tinh Nguyen, Xuewen Luo, Björn Hoffmeister, Jan Trmal, Maurizio Omologo, and Roland Maas. 2019. Dipco–dinner party corpus. *arXiv preprint arXiv:1909.13447*.

Patrick von Platen. 2021. Fine-tuned XLSR-53 large model for speech recognition in vietnamese. https://huggingface.co/not-tanh/wav2vec2-large-xlsr-53-vietnamese.

Changhan Wang, Anne Wu, and Juan Pino. 2020. Covost 2: A massively multilingual speech-to-text translation corpus.

Hao Wang, Shuhei Kurita, Shuichiro Shimizu, and Daisuke Kawahara. 2024. Slideavsr: A dataset of paper explanation videos for audio-visual speech recognition. *arXiv preprint arXiv:2401.09759*.

Peng Xie, Xingyuan Liu, Tsz Wai Chan, Yequan Bie, Yangqiu Song, Yang Wang, Hao Chen, and Kani Chen. 2025. Switchlingua: The first large-scale multilingual and multi-ethnic code-switching dataset. *arXiv preprint arXiv:2506.00087*.

Guanrou Yang, Ziyang Ma, Fan Yu, Zhifu Gao, Shiliang Zhang, and Xie Chen. 2024. Mala-asr: Multimedia-assisted llm-based asr. *arXiv preprint arXiv:2406.05839*.

Qiming Zhang, Jing Zhang, Yufei Xu, and Dacheng Tao. 2024. Vision transformer with quadrangle attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019. Dialogpt: Large-scale generative pre-training for conversational response generation. *CoRR*, abs/1911.00536.