# ‘Beach’ to ‘Bitch’: Inadvertent Unsafe Transcription of Kids’ Content on YouTube

**Krithika Ramesh [1]*, Ashiqur R. Khudabukhsh [2]†, Sumeet Kumar [3]†**

[1] Manipal Institute of Technology, Manipal
[2] Rochester Institute of Technology
[3] Indian School of Business, Hyderabad
kramesh.tlw@gmail.com, axkvse@rit.edu, sumeet_kumar@isb.edu

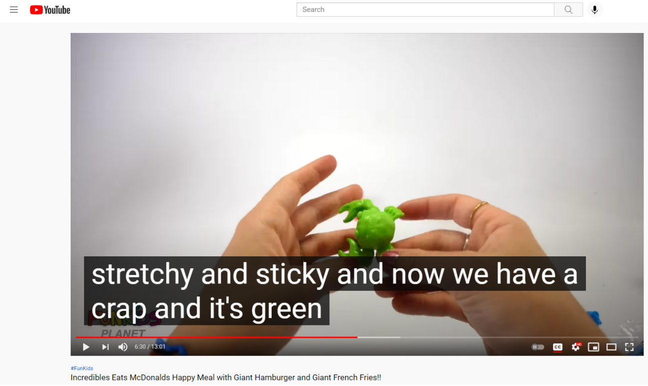
## Abstract

Over the last few years, YouTube Kids has emerged as one of the highly competitive alternatives to television for children’s entertainment. Consequently, YouTube Kids’ content should receive an additional level of scrutiny to ensure children’s safety. While research on detecting offensive or inappropriate content for kids is gaining momentum, little or no current work exists that investigates to what extent AI applications can (accidentally) introduce content that is inappropriate for kids.

In this paper, we present a novel (and troubling) finding that well-known automatic speech recognition (ASR) systems may produce text content highly inappropriate for kids while transcribing YouTube Kids’ videos. We dub this phenomenon as *inappropriate content hallucination*. Our analyses suggest that such hallucinations are far from occasional, and the ASR systems often produce them with high confidence. We release a first-of-its-kind data set of audios for which the existing state-of-the-art ASR systems hallucinate inappropriate content for kids. In addition, we demonstrate that some of these errors can be fixed using language models.

## Introduction

Over the last few years, YouTube Kids has emerged as one of the highly competitive alternatives to television for children’s entertainment. For example, between 2015 to 2021, the subscriber count of Ryan’s World, a highly popular YouTube channel for kids, rose from 32K to 30.3 millions[1]. With this steep viewership growth, content hosted in these channels has received escalated scrutiny. Following recent reports indicating occasional slip-ups in YouTube Kids’ content moderation systems, recent works have explored automatic detection of inappropriate content from videos to aid human moderation (Papadamou et al. 2020; Han and Ansingkar 2020; Alghowinem 2018).



stretchy and sticky and now we have a
crap and it's green

Incredibles Eats McDonalds Happy Meal with Giant Hamburger and Giant French Fries!!

Figure 1: Example of *inappropriate content hallucination*[2]. A screenshot of a YouTube video taken from a popular channel *Fun Kids Planet* with 1.34m subscribers. The ASR (automatic speech transcription) system incorrectly transcribes crab as crap.

While detecting offensive or inappropriate content for specific demographics is a well-studied problem, such studies typically focus on detecting offensive content present in the source, not how objectionable content can be (accidentally) introduced by a downstream AI application. With the growing trend of complex AI pipelines where one subsystem’s output is another sub-system’s input, it is increasingly becoming more important to acknowledge that inappropriate content may not always be present in the source, it can inadvertently creep in through a downstream AI application. Following existing literature in computer vision (Kayhan, Vredebregt, and van Gemert 2021) and natural language processing (Rohrbach et al. 2018; Ji et al. 2022) on content hallucination, we call this phenomenon as *inappropriate content hallucination*. In this paper, we study the interaction between YouTube Kids video content and ASR (automatic speech recognition) systems to assess this phenomenon. ASR systems have prominent use in multimedia systems to generate transcripts for video content at scale and have found use in a broad range of applications that include

[*] Krithika Ramesh was equally advised by Ashiqur R. KhudaBukhsh and Sumeet Kumar. She worked on this project while working as a research intern at the Indian School of Business.

[†] Ashiqur R. KhudaBukhsh and Sumeet Kumar are joint corresponding authors.

[1] Source: https://web.archive.org/

[2] See the project page: https://github.com/sumeetkr/UnsafeTranscriptionofKidsContent, for more examples.

| Ground truth | Transcription |
|---|---|
| …its fluffy and so so soft toy… | …its fucking and so so soft toy… |
| …if you like this craft keep on watching until the end so you can see related videos… | …if you like this crap keep on watching until the end so you can see related videos… |
| …stretchy and sticky and now we have a crab and its green… | …stretchy and sticky and now we have a crap and its green… |
| …duck pulls away here comes arthur mr conductor tries as hard as duck gets a little bit of boost but its not enough… | …duck pulls away here comes arthur mr conductor tries as hard as fuck gets a little bit of boost but its not enough… |
| …in order to be strong and brave like heracles… | …in order to be strong and rape like heracles… |

Table 1: Examples where ASR systems hallucinated taboo-words originally not present in the source.

call routing (Riccardi et al. 1997), transcribing meetings and lectures (Ranchal et al. 2013), IoT appliances (Mehrabani, Bangalore, and Stern 2015), and medical scribing (Finley et al. 2018). Our study uncovers a disturbing pattern of (inadvertent) introduction of inappropriate words by ASR systems while transcribing kids' videos described through the following illustrative example.

## An Illustrative Example

Consider the real-world example from a highly popular YouTube Kids channel presented in Figure 1. All words present in the utterance are correctly transcribed except for the word `crab`, which is incorrectly transcribed as `crap`. Among the broad taxonomy of offensive/inappropriate words such as profanity, slur, insults, and slang, following developmental psychology literature on children (Jay 1992), we broadly categorize such inappropriate words as *taboo-words*.

We do not know what percentage of kids use YouTube Kids application (as opposed to YouTube) and what fraction of kids watching videos turn their subtitles on while watching the videos. However, documented evidence exists indicating that (1) *same language subtitling* (Vanderplank 2016b) and captioned media in foreign language (Vanderplank 2016a) both improve learning in children; and (2) for children with disabilities, captions provide a critical learning resource (Vanderplank 2016a). In the example presented in Figure 1 where every other word in the audio matches with the transcript but the taboo-word, it is not hard to envision that there is a potential risk for kids to incorporate this taboo-word into their vocabulary without even knowing that it is inappropriate.

As indicated in Table 1, such hallucinations of taboo-words are far from a one-off incident and some of the videos containing highly inappropriate hallucinated taboo-words could have been exposed to millions of viewers. Of course, ASR for kids adds additional challenges and new methods have investigated possible mitigation strategies (Wu et al. 2019; Yeung and Alwan 2018; Plantinga and Fosler-Lussier 2019). However, to our knowledge, no such study considered this novel risk of introducing unsafe content through transcription errors. Our study thus adds value to the ongoing conversation around the potential risks and harms to the society that can be caused by over-reliance on AI systems and applying them to a demographic potentially underrepresented in the training data (in our case, kids).

Further, one key point to remember is that none of these taboo-words were present in the actual content – they are (accidentally) introduced by a downstream AI application. As already mentioned, with complex and connected AI applications where outputs of one AI application may form inputs to others[3], these inadvertently generated taboo-words can pose unseen challenges for downstream applications.

While our first goal in this paper is to attract the attention of the research community to this novel threat of accidental introduction of inappropriate content through downstream AI applications, in order to facilitate robust testing of ASR systems, one of our key contributions is a novel data set of challenging audio inputs in which major ASR systems hallucinated taboo-words[4]. In addition, we demonstrate that some of these challenging instances can be corrected using high-performance language models.

**Contributions:** Our contributions are the following.

1. *Social:* Via a comprehensive study of 7,103 videos on 24 YouTube Kids (YTK) channels, our study indicates that prominent ASR systems often hallucinate taboo-words in children's content.
2. *Resource:* We release a first-of-its-kind data set of 652 challenging audio inputs in which prominent ASR systems have hallucinated taboo-words along with the ground truth transcriptions. We release a lexicon of 1,301 taboo-words for kids that draws from developmental psychology literature, a curated corpus MPAA rated movie subtitles data set and a well-known hate lexicon.
3. *Method:* We present a novel application of masked language models to fix some of these errors.

## Related Work

Although transcripts generated using ASR systems lack in terms of quality as compared to manual transcription performed by human annotators, ASR systems have found use in a broad range of applications that include call routing (Riccardi et al. 1997), transcribing meetings and lectures (Ranchal et al. 2013), video subtitling (Sawaf 2012), IoT appliances (Mehrabani, Bangalore, and Stern 2015), and medical scribing (Finley et al. 2018). In addition to challenges such as background noise (Rajnoha and Pollák 2011)

[3] In fact, existing multimodal approaches (Alghowinem 2018) have consulted ASR outputs to detect inappropriate content in videos.

[4] Data set, lexicon, and additional details are available at our project page: https://github.com/sumeetkr/UnsafeTranscriptionofKidsContent.

and speech variability (Benzeghiba et al. 2007), and specifically while transcribing kids' speeches (Wu et al. 2019; Yeung and Alwan 2018; Plantinga and Fosler-Lussier 2019), ASR systems face issues with comprehending speech where stuttering, erratic pauses, dysarthric speech, etc. are present (Mengistu and Rudzicz 2011).

Given the disparity in the latency between manual transcription and ASR transcripts, Gaur et al. (2016) propose allowing the transcriptionists to utilize the ASR outputs as a starting point for their transcription work, as manual transcription would otherwise take up to almost five times the length of the audio. It was noted that this did indeed reduce the normalized average latency, provided that the WER (Word Error Rate) of the ASR system was less than 30%.

Other approaches to improve the quality of the audio transcripts involve grammatical correction. Several lines of work exist that have used language models to improve ASR systems (see, e.g., (Arisoy et al. 2015; Chen et al. 2017; Shin, Lee, and Jung 2019; Namazifar et al. 2021)). Namazifar et al. (2021) introduced WLM-SC, a generalized version of masked language models that uses warping, and trained the model on data sets containing grammatical errors to become robust to these word-level errors. WLM-SC, when used for sentence correction, demonstrates that WLM-SC not only improves the WER of automatic transcriptions but that of human transcriptions as well.

Kim et al. (2019) draw a comparison between multiple ASR systems, and manual transcription ones as well, which outstrip the automatic ones in terms of their WER. In addition, they also attempt to find a correlation between nonverbal behavior cues and unintelligible speech, showing that the variability of the speech intensity is lower when the speech is not clear. Our work contrasts with existing research on ASR systems in two key ways: (1) unlike traditional performance metrics like word error rate (WER), we focus on a potentially harmful content hallucination of ASR systems, a phenomenon never studied before to our knowledge; and (2) we release a novel benchmark data set of challenging hallucinated examples with ground truth.

At a philosophical level, our work is closely related to (Bender et al. 2021; Gehman et al. 2020) that explore the potential risks associated with large opaque models, including the lack of diversity in the data they are trained on, and the biases they exhibit. In particular, the work discusses how biases can reflect in the derogatory language that could potentially be produced by the model, in the form of racial slurs and derogatory terms that target marginalized communities, and the difficulty of filtering out such terms from our data.

## Data and Design Considerations

### Data Set

We create a new data set by collecting videos from top YouTube Kids (YTK) channels. In order to construct a data set highly relevant for kids, among the vast number of channels on YTK, we focus on the top-ranked channels based on their popularity (i.e., number of views). We use two rankings from Wall Street Journal [5] and Statista [6] to get a few very popular channels.

For these channels, we next retrieve all English language videos in those channels using a widely used library[7]. Our data set comprises a total of 7,013 videos extracted across 24 channels that fall under the YouTube Kids category. We consider both the transcriptions provided by the YouTube API and Amazon Transcribe in our experiments. Further details regarding the data set are listed in Table 2.

Some of the videos involved music and rhymes, some involved interaction between two or more characters, and some involved no form of speech at all. Despite there being no verbal interactions in some of these videos, we found that in many of these cases, the services nevertheless produced transcriptions, which in some cases contained toxic language and hate speech, in direct violation of YouTube Kids' guidelines. [8]

### Speech to Text Methods

We next give a brief description of the two speech-to-text transcript services that we consider.

- Amazon (AWS) Transcription: As per the description on Amazon website[9], AWS Transcribe uses a deep learning method to convert speech to text. In addition to subtitling, Amazon Transcribe also generates metadata including the number of speakers and which transcription was a result of the speech by which speaker. For generating AWS transcriptions, we first obtained the audio of the YouTube videos. Then we used AWS transcribe service to transcribe the audio files to text.
- YouTube (Google) Transcription: In addition to AWS, we also consider YouTube transcriptions. YouTube transcriptions are created when a video is uploaded. As per YouTube[10], automatic captions are generated by machine learning algorithms, so the quality of captions may vary across videos. Though there are a number of languages for which captions could be available, for the analysis in this paper, we only consider videos for which English language transcripts (also called captions) are available via YouTube API[11].

### Transcription Quality

It may well be the case that the overall transcription quality of our data set is low and the presence of taboo-words are mere artifacts of noisy outputs. We first thus validate the quality of transcripts.

**Mutual Agreement** Let $\mathcal{T}_i^{\mathcal{A}_j}$ denote the transcript obtained from video $v_i$ using transcription algorithm $\mathcal{A}_j$. Let

[5] https://www.wsj.com/articles/kids-love-these-youtube-channels-who-creates-them-is-a-mystery-11554975000

[6] https://www.statista.com/statistics/785626/most-popular-youtube-children-channels-ranked-by-subscribers/

[7] https://github.com/ytdl-org/youtube-dl

[8] https://support.google.com/youtube/answer/9528076?hl=en

[9] https://aws.amazon.com/transcribe/

[10] https://support.google.com/youtube/answer/6373554

[11] https://developers.google.com/youtube/v3/docs/captions

| Channel Name | Channel View Count | Channel Subscriber Count | # of Videos | # of Google Transcripts with taboo-words | # of AWS Transcripts with taboo-words |
|---|---|---|---|---|---|
| Sesame Street | 20 Billion | 23 Million | 2405 | 432 (2) | 763 (122) |
| Ryan's World | 50 Billion | 32 Million | 1437 | 892 (9) | 1,138 (383) |
| 3KidsTV | 2 Million | 10 Thousand | 39 | 8 (0) | 12 (0) |
| Sesame Studios | 343 Million | 615 Thousand | 320 | 62(0) | 113(10) |
| Barbie | 3 Billion | 11 Million | 98 | 26 (1) | 32 (6) |
| Moonbug Kids - Cartoons & Toys | 110 Million | 3 Million | 632 | 497(31) | 499(63) |
| Elizabeth & Eva TV | 42 Million | 172 Thousand | 157 | 57 (1) | 87 (30) |
| Rob The Robot - Learning Videos For Children | 215 Million | 380 Thousand | 113 | 90 (3) | 103 (24) |
| SimpleCrafts - 5 Minute Crafts For All | 211 Million | 718 Thousand | 736 | 257 (1) | 333 (16) |
| Kids Toys Play | 524 Million | 461 Thousand | 112 | 76 (1) | 91 (22) |
| Mister Max | 14 Billion | 22 Million | 72 | 8 (0) | 31 (12) |
| Blippi - Educational Videos for Kids | 12 Billion | 15 Million | 78 | 57 (0) | 64 (15) |
| Like Nastya | 69 Billion | 86 Million | 224 | 44 (1) | 76 (31) |
| New Sky Kids | 2 Billion | 2 Million | 40 | 37 (0) | 39 (14) |
| Kiddopedia | 339 Million | 717 Thousand | 116 | 53 (0) | 65 (2) |
| Baby Einstein | 575 Million | 768 Thousand | 124 | 23 (0) | 37 (9) |
| Fun Kids Planet | 828 Million | 1 Million | 121 | 86 (3) | 107 (26) |
| ChuChuTV Surprise Eggs Learning Videos | 4 Billion | 7 Million | 64 | 16 (0) | 23 (5) |
| ChuChu TV Nursery Rhymes & Kids Songs | 36 Billion | 54 Million | 59 | 21 (1) | 30 (13) |
| Funny Kids Playtime with Jade & James - ToysReview | 151 Thousand | 1 Thousand | 29 | 24 (3) | 27 (9) |
| Dipo Dipo | 8 Million | 19 Thousand | 25 | 0 (0) | 1 (1) |
| Two Kids TV | 154 Million | 308 Thousand | 11 | 1 (0) | 1 (0) |
| Cupcake Squad | 788 Million | 2 Million | 1 | 1 (0) | 0 (0) |

Table 2: List of YouTube channels considered. Channel statistics reflect data as on 20 Jan 2022. Numbers in braces () indicate highly inappropriate taboo-words' count.

$\mathcal{N}(w_k, \mathcal{T}_i)$ denote the total number of occurrences of the word $w_k$ in transcript $\mathcal{T}_i$. Further, let $\mathcal{V}(\mathcal{T}_i)$ denote the vocabulary of a transcript $\mathcal{T}_i$. The mutual agreement between two transcripts of the same video is the fraction of words that have appeared identical number of times over both transcripts:

$MA\ (\mathcal{T}_i^{\mathcal{A}_j}, \mathcal{T}_i^{\mathcal{A}_{j'}}) = \frac{\Sigma_{w_k \in \mathcal{V}} I(\mathcal{N}(w_k, \mathcal{T}_i^{\mathcal{A}_j}) = \mathcal{N}(w_k, \mathcal{T}_i^{\mathcal{A}_{j'}}))}{|\mathcal{V}|}$ where $\mathcal{V} = \mathcal{V}(\mathcal{T}_i^{\mathcal{A}_j}) \cup \mathcal{V}(\mathcal{T}_i^{\mathcal{A}_{j'}})$, i.e., the union of the vocabularies of the transcripts generated by algorithms $\mathcal{A}_j$ and $\mathcal{A}_{j'}$ when applied to video $v_i$.

Our intuition is if both transcription algorithms perform a reliable job on a given video, the mutual agreement value will be high. We note that this is a quite stringent criterion as every single error while transcribing a video may contribute to a reduced *MA*. Figure 2 indicates that more than 42% of the videos have mutual agreement higher than 50%. In addition, we conduct a human inspection of the videos and confirm that several videos in our data set have overall high-quality transcripts.

## Developing a Set of Taboo-words for Kids

Deciding on the set of taboo-words is one of the major design considerations in this project. Several factors such as subjectivity, cultural contexts, and audience can determine if a word is perceived as inappropriate or not. Consequently, there exists no broad consensus on hate lexicons. As a starting point, we select a well-known, publicly available hate lexicon[12] containing more than 1,300 words (denoted as $\mathcal{H}_1$). While describing this lexicon as a reasonable starting point to block or filter offensive content, curators of this lexicon acknowledge the subjectivity of this lexicon stating that the list contains words that many people may **not** find offensive.

Our second choice of the lexicon is strongly grounded in prior literature in developmental psychology (Sutton-Smith and Abrams 1978; Jay and Jay 2013). We obtain a list of 76 words presented in Jay (1992) as taboo-words for children.

[12] Available at https://www.cs.cmu.edu/~biglou/resources/bad-words.txt

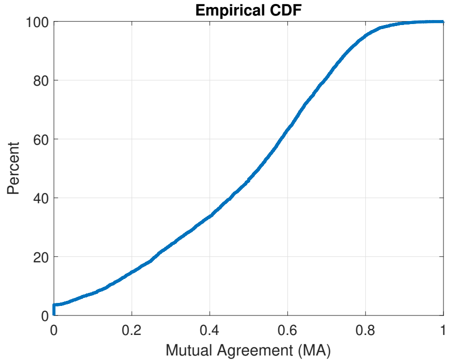


Figure 2: A CDF plot of mutual agreement in our data set between Amazon Transcribe and Google Speech-to-Text.

These words are collected from actual usage of these words by children within the age range of 1–10 in a field study (Jay 1992) (denoted as $\mathcal{H}_2$).

We note that both $\mathcal{H}_1$ and $\mathcal{H}_2$ complement each other and combining them may have certain merits. Since $\mathcal{H}_2$ consists of words actually used by kids in a field study, it precludes certain inappropriate words with strong sexual connotations as these words require adult-level understanding. For instance, unlike $\mathcal{H}_1$, words like `cocksucker` or `rape` are absent in $\mathcal{H}_2$. In our work, we are focusing on content hallucinations from ASR systems. Thus casting a net wider than $\mathcal{H}_2$ has understandable benefits (in fact, as shown in Table 1, one of the content hallucination indeed produces the taboo-word `rape`).

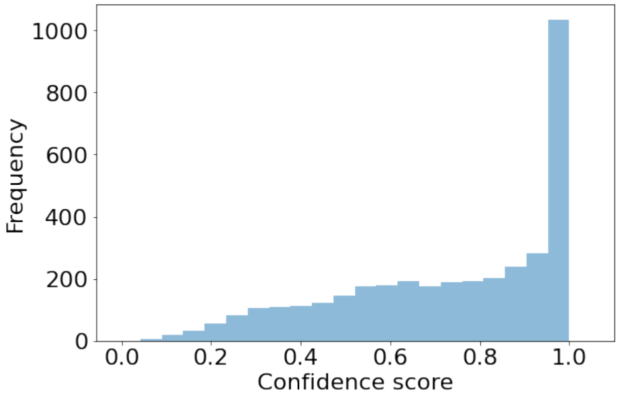


Figure 3: Plot of the confidence scores of all taboo-words and their frequency from Amazon Transcribe.

While words belonging to $\mathcal{H}_2$ serve an important purpose to give us a broad understanding of what could be construed as inappropriate for kids, we note that some of these words could be heavily context-dependent. For instance, the word `dog` may be used in a completely non-taboo scenario. Similarly, we find that $\mathcal{H}_1$ also contains certain words that can be used in non-taboo scenarios. For example, the word `killer` can be used in a completely harmless context of `killer whale`.

We thus combine both $\mathcal{H}_1$ and $\mathcal{H}_2$ and analyze to what extent these words are present in children's movies. We construct a data set, $\mathcal{D}_{Disney\text{-}Pixar}$, consisting of English subtitles[13] of all movies released in or after 2000 with an MPAA movie rating of (G) implying that these movies are certified as safe for the general audience and nothing would offend parents for viewing by children. Overall, we obtained 57 movies. Our choice of these well-regarded movie franchises with MPAA certifications of (G) serves two key purposes. First, a wide variety of entertainment content targeted for kids ensures that the corpus consists of a rich set of kids' entertainment contexts. Second, the MPAA rating indicates that these contexts are certified as appropriate for children. Our intuition is if a word $w \in \mathcal{H}_1 \cup \mathcal{H}_2$ appears on multiple occasions in $\mathcal{D}_{Disney\text{-}Pixar}$, it possibly indicates that that $w$ can be used in non-taboo scenarios for kids. Table 3 shows top 10 words from $\mathcal{H}_1$ and $\mathcal{H}_2$ that have appeared at least five or more times in $\mathcal{D}_{Disney\text{-}Pixar}$. We further note that in $\mathcal{D}_{Disney\text{-}Pixar}$, we observe zero mentions of words with strong sexual connotation such as `rape` and `fuck` and scatological references such as `shit`.

| $\mathcal{H}_1$ | $\mathcal{H}_2$ |
|---|---|
| kid, fairy, fairies, girls, dead, fight, god, bigger, stupid, shoot | dog, god, stupid, hell, pig, fat, chicken, nuts, silly, butt |

Table 3: Top ten words (ranked by frequency) from $\mathcal{H}_1$ and $\mathcal{H}_2$ present in $\mathcal{D}_{Disney\text{-}Pixar}$. $\mathcal{D}_{Disney\text{-}Pixar}$ consists of English subtitles of all Disney and Pixar movies with MPAA rating (G) released in or after 2000.

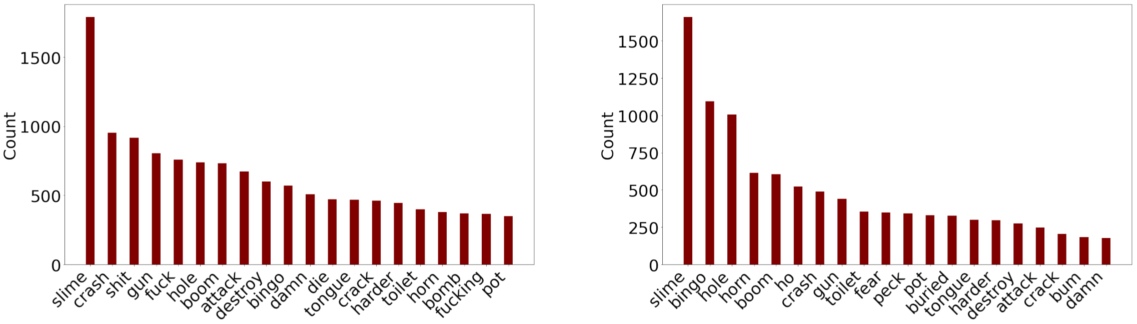
We remove all words from $\mathcal{H}_1 \cup \mathcal{H}_2$ that have appeared for five or more times in $\mathcal{D}_{Disney\text{-}Pixar}$. In addition, we manually remove words indicating nationality of a person (e.g., `Italian`, `American`) or religious words (e.g., `Muslim`). After this step, our combined lexicon of taboo-words, $\mathcal{H}$, consists of 1,301 words.

## Extent of Presence of Taboo-words

Overall, we find that 330 taboo-words were present in YouTube transcripts and 386 taboo-words were present in AWS transcripts in our data set. Figure 4 presents the distribution of the top twenty words belonging to $\mathcal{H}$ present in our transcript data set. We observe the considerable presence of inappropriate taboo-words such as `shit` in transcripts generated by Amazon Transcribe and Google Speech-to-Text.

While Figure 4 indicates a worrisome finding of considerable presence of taboo-words in video transcripts, some of these words may not be hallucinated and could be present in contexts safe for kids. For example, `ho`, which also implies the disparaging and offensive slang `whore`, could be simply present in a Santa Claus video. We thus manually inspected $\mathcal{H}$ and shortlisted a set of 16 words that are unambiguously inappropriate and analyzed their presence in the video transcripts. As shown in Figure 5, we find that the video tran-

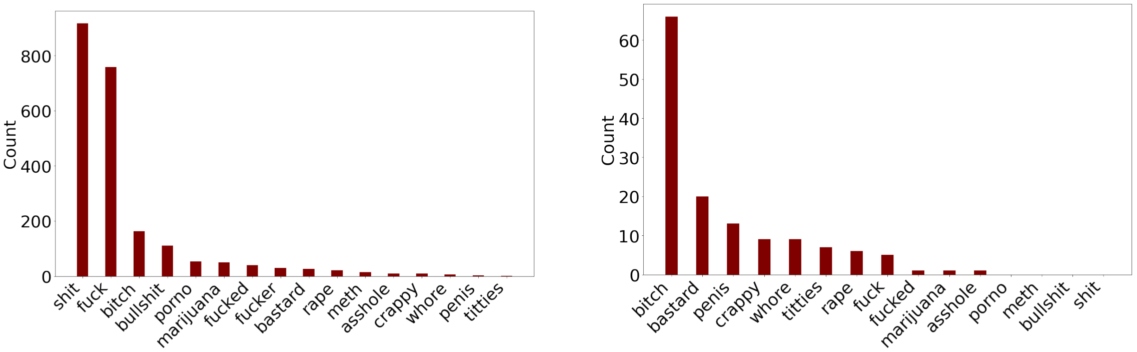
[13] Subtitles are obtained from Subscene.org



(a) Amazon Transcribe

(b) Google Speech-to-Text

Figure 4: Top twenty taboo-words (potentially) hallucinated by Amazon Transcribe and Google Speech-to-Text in our data set.



(a) Amazon Transcribe

(b) Google Speech-to-Text

Figure 5: Top few highly inappropriate taboo-words (potentially) hallucinated by Amazon Transcribe and Google Speech-to-Text in our data set.

scripts also exhibit substantial presence of these highly inappropriate words. In fact, Table 2 indicates that nearly one in ten videos contains at least one or more of these highly inappropriate taboo-words in transcription generated either by AWS or Google Speech-to-Text. In addition, we show some less frequent but highly inappropriate words in Figure 6.

***RQ 1:*** *Are these taboo-words indeed hallucinated, or are they indeed present in the actual audio content?*

We first randomly sample 100 contexts containing five highly inappropriate taboo-words ({`shit`, `fuck`, `crap`, `rape`, `ass`}) and manually inspect if these words are hallucinated or if they are indeed present in the source. Two annotators independently listened to the audio clips and confirmed that none of the taboo-words occurred in the actual audio. Upon manual inspection, we identify the following high-level potential factors to these content hallucinations: (1) background music; (2) baby talk; (3) kids' speech; (4) ESL (English as Second Language speakers) speakers; and (5) songs and rhymes. Note that, we do not intend to be formal or exhaustive, but rather to be illustrative of the broad range of potential reasons that can cause such hallucination.

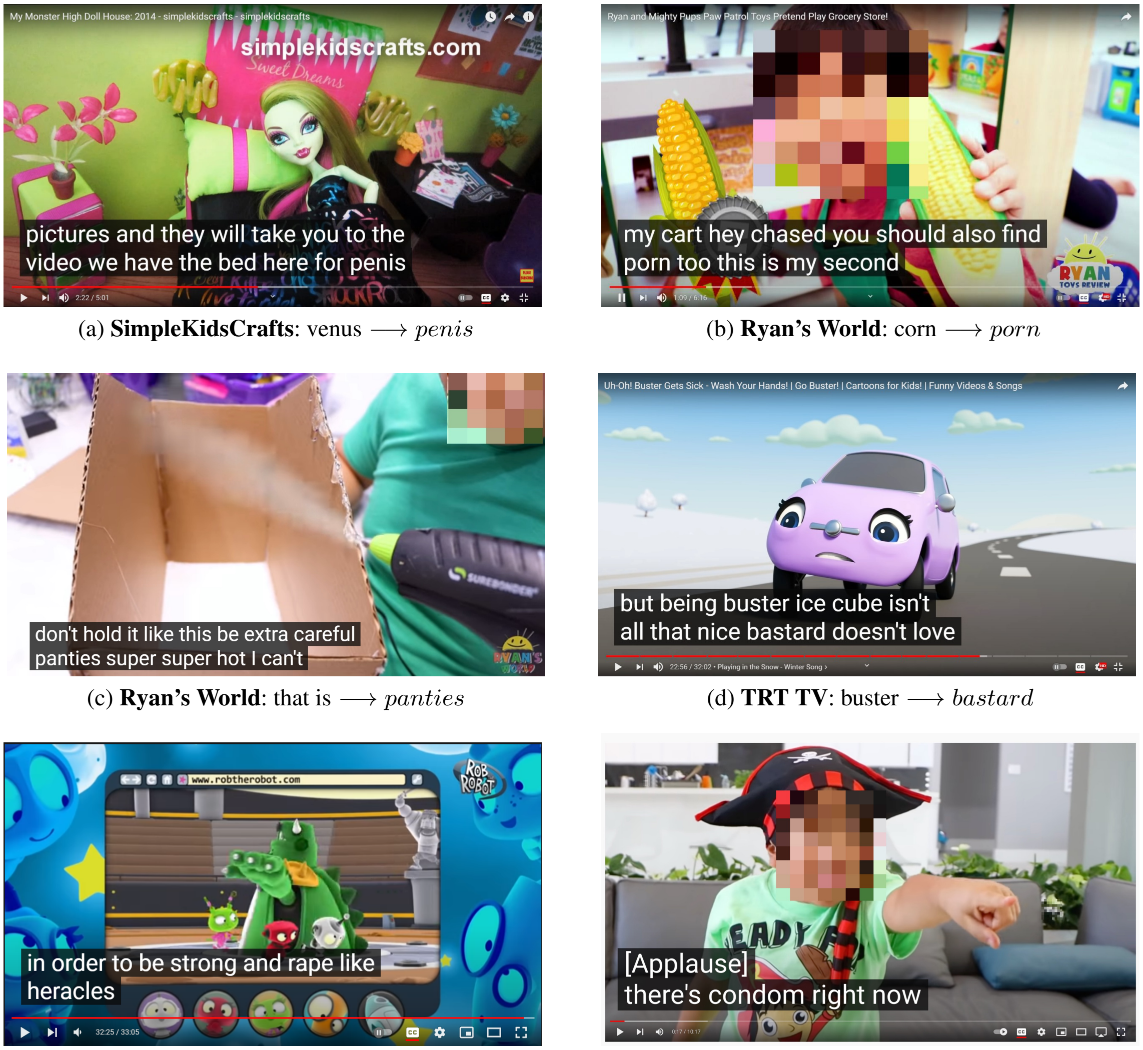
***RQ 2:*** *How confident is the transcription method while hallucinating a taboo-word?* Amazon Transcribe presents a confidence estimate of individual words. Figure 3 indicates that a vast number of taboo-words were high-confidence predictions of this system.

Once we establish that these taboo-words are largely absent in the audio inputs and the transcription methods often produce them while reliably transcribing a large part of the adjacent audio inputs, we turn our focus into creating a data set consisting of challenging audio inputs where high-performance commercial ASR systems hallucinate taboo-words.

We construct a data set, $\mathcal{D}_{taboo}$, consisting of 284 YouTube transcriptions and 368 Amazon Transcribe transcriptions that satisfy the following conditions: (1) the snippet contains a taboo-word; and (2) both transcription algorithms exhibit considerable agreement within transcribing the snippet (algorithm sketch is presented in the project page). All samples contain consensus labels from two annotators.

## Corrective Method

Our method to correct taboo-words from video transcripts relies on cloze tasks performed by recent high-performance language models such as `BERT` (Devlin et al. 2019), XLM (Chi et al. 2021), DistilBERT (Sanh et al. 2019), XLNet (Yang et al. 2019), and Megatron (Shoeybi et al. 2019) . When presented with a sentence (or a sentence stem) with a missing word, a cloze task is essentially a fill-in-the-


My Monster High Doll House: 2014 - simplekidscrafts - simplekidscrafts
simplekidscrafts.com
pictures and they will take you to the
video we have the bed here for penis
Ryan and Mighty Pups Paw Patrol Toys Pretend Play Grocery Store!
my cart hey chased you should also find
porn too this is my second
don't hold it like this be extra careful
panties super super hot I can't
Uh-Oh! Buster Gets Sick - Wash Your Hands! | Go Buster! | Cartoons for Kids! | Funny Videos & Songs
but being buster ice cube isn't
all that nice bastard doesn't love
www.robtherobot.com
in order to be strong and rape like
heracles
[Applause]
there's condom right now

(a) **SimpleKidsCrafts**: venus ⟶ *penis*

(b) **Ryan's World**: corn ⟶ *porn*

(c) **Ryan's World**: that is ⟶ *panties*

(d) **TRT TV**: buster ⟶ *bastard*

(e) **Rob The Robot - Learning Videos For Children**: brave ⟶ *rape*

(f) **Ryan's World**: combo ⟶ *condom*

Figure 6: Examples of hallucinated taboo-words from YouTube along with corresponding ground truths.

blank task. For instance, in the following cloze task: *During the* [MASK], *it rains a lot*, monsoon is a likely completion for the missing word. BERT's masked word prediction has a direct parallel to cloze task introduced in the psycholinguistics literature (Taylor 1953). BERT's cloze task has been previously used in the (1) extracting relational knowledge (Petroni et al. 2019); (2) mining political insights (Palakodety, KhudaBukhsh, and Carbonell 2020); (3) assessing the quality of translation (Zhang* et al. 2020); and (4) estimating linguistic quality (Sarkar, Mahinder, and KhudaBukhsh 2020).

Our method's intuitions are the following. It is highly likely that the ground truth is phonetically (and lexically) similar to the taboo-word (e.g., $\langle \texttt{crap}, \texttt{crab} \rangle$; $\langle \texttt{seat}, \texttt{shit} \rangle$; $\langle \texttt{rape}, \texttt{rake} \rangle$). Our method thus first constructs a set of candidate words based on some notion of proximity (lexical or phonetic). Next, it conducts a constrained cloze task using a language model that only considers the candidate set as potential completions. Assuming that the transcriptions for the context minus the taboo-word are reliable, we expect that the context would be able to guide the language models towards the correct completion.

Let $\texttt{LM}_{cloze}(w, \mathcal{S})$ denote the completion probability of the word $w$ when a language model, LM, has a masked cloze task $\mathcal{S}$ as input. When we have a text snippet with a taboo-word, we construct $S$ by masking the taboo-word. For example, if our snippet is the following: *I love to eat* ***crap*** *and lobster for dinner.*, our cloze task will be *I love to eat [MASK] and lobster for dinner.* Let our candidate set, $\mathcal{C} = \{\texttt{crab}, \texttt{crap}, \texttt{craft}\}$, be consisting of words that are similar to the taboo-word based on some similarity measure.

Our corrective method will output $c^*$ obtained by:

$$c^* = \operatorname*{argmax}_{c \in \mathcal{C}} \mathrm{LM}_{cloze}(c, \mathcal{S})$$

For language models, we consider several well-known high-performance language models: `BERT`, `XLM`, `XLNet`, `DistilBERT`, and `Megatron`. For a given taboo-word, we generate the candidate set using two different approaches. In our first approach, we consider words with a low Levenshtein distance from the taboo-word. Our second approach considers highly phonetically similar words to the taboo-word (details are present on the project page). In order to conduct a fair assessment of the language models, we restrict our experiments to samples where (1) single-word substitutions would suffice to fix the error, and (2) the ground truth is included in the language model's vocabulary.

Table 4 summarizes our correction results. **P@1** performance indicates the fraction of instances where the top predicted completion in the cloze test is indeed the ground truth. Following, standard practice (Petroni et al. 2019), we also report the **P@5** and **P@10** performance where **P@K** performance indicates that the top $K$ completions for the cloze test contains the ground truth. The computing infrastructure used for running experiments is described in the project page along with information on the hyper-parameters used for each model.

| Corrective method | ASR algorithm | P@1 | P@5 | P@10 |
|---|---|---|---|---|
| ⟨`BERT`, *Levenshtein*⟩ | Amazon | 19.1% | 40.7% | 45.9% |
| ⟨`BERT`, *Phonetic*⟩ | Amazon | 7.7% | 25.8% | 35.4% |
| ⟨`XLM`, *Levenshtein*⟩ | Amazon | 6.2% | 24.4% | 40.7% |
| ⟨`XLM`, *Phonetic*⟩ | Amazon | 6.2% | 22.0% | 35.9% |
| ⟨`XLNet`, *Levenshtein*⟩ | Amazon | 5.3% | 23.0% | 39.7% |
| ⟨`XLNet`, *Phonetic*⟩ | Amazon | 4.8% | 15.8% | 34.0% |
| ⟨`DistilBERT`, *Levenshtein*⟩ | Amazon | 16.8% | 38.8% | 45.0% |
| ⟨`DistilBERT`, *Phonetic*⟩ | Amazon | 8.1% | 25.4% | 35.4% |
| ⟨`Megatron`, *Levenshtein*⟩ | Amazon | 25.4% | 40.7% | 45.0% |
| ⟨`Megatron`, *Phonetic*⟩ | Amazon | 10.5% | 27.8% | 35.0% |
| ⟨`BERT`, *Levenshtein*⟩ | YouTube | 22.3% | 37.6% | 44.1% |
| ⟨`BERT`, *Phonetic*⟩ | YouTube | 5.3% | 21.8% | 31.8% |
| ⟨`XLM`, *Levenshtein*⟩ | YouTube | 18.2% | 42.9% | 47.7% |
| ⟨`XLM`, *Phonetic*⟩ | YouTube | 7.06% | 21.2% | 34.7% |
| ⟨`XLNet`, *Levenshtein*⟩ | YouTube | 10.0% | 28.8% | 39.4% |
| ⟨`XLNet`, *Phonetic*⟩ | YouTube | 0.6% | 17.7% | 35.9% |
| ⟨`DistilBERT`, *Levenshtein*⟩ | YouTube | 18.8% | 36.5% | 44.1% |
| ⟨`DistilBERT`, *Phonetic*⟩ | YouTube | 10.0% | 22.9% | 31.8% |
| ⟨`Megatron`, *Levenshtein*⟩ | YouTube | 28.2% | 39.4% | 46.5% |
| ⟨`Megatron`, *Phonetic*⟩ | YouTube | 10.6% | 28.2% | 31.8% |

Table 4: Performance on our benchmark data set $\mathcal{D}_{taboo}$.

As observed in Table 4, we find that among all the combinations, ⟨`Megatron`, *Levenshtein*⟩ obtained the best performance on the data, fixing more than 25% of the errors produced by Amazon Transcribe and over 28% of the errors hallucinated by Google Speech-to-Text. We were not surprised by our modest success at fixing these hallucinated taboo-words. While the inner workings of Amazon Transcribe and Google Speech-to-Text are opaque, there exists substantial literature on ASR systems leveraging language models. Furthermore, here, we are attempting to fix non-trivial, extremely challenging transcription errors hallucinated by industry-scale, commercial solutions. We are rather encouraged by a high **P@5** performance indicating that a human-in-the-loop setting can be aided by our method for corrective purposes.

| | |
|---|---|
| rape \| brave | monsters in order to be strong and **rape** like heracles we even had a chariot race in order to be fast like heracles but when orbit rescued emma we realized theres more to a hero than just being |
| bitch \| beach | glasses so you can see that and then here we have his sandals his sandals are completely made out of plastic they are orange and they have the same flames at the top and then we have a little **bitch** towel that came with him and what is really cool about this towel is a motif which |
| crap \| craft | if you have any requests or **crap** ideas that you would like us to explore kindly send us an email |

Table 5: Examples of correctly fixed snippets. In the left column, the hallucinated taboo-word is followed by the ground truth. In the right column, the example is presented with the taboo word marked in bold.

We next turn our focus to some of the correctly fixed examples. As shown in Table 5, we notice that when sentences are well-formed and present with enough context, language models are often successful at fixing the error. However, given the nature of kids' channels, many of the transcripts contain incoherent, ill-formed sentences, thus making it extremely challenging for LMs to predict the masked word correctly.

| | |
|---|---|
| cocktail \| copter \| social | also exceptionally strong for their size they can lift 10 to 50 times their own weight thats like being a little hero that can lift their own **cocktail** over their head other cool features |
| penis \| venus \| pets | you need is in the description but as were passing through the pictures you can click on the pictures and they will take you to the video we have the bed here for **penis** and the side drawers as well and here we have the arts and crafts studio which is basically everything from |
| bastard \| buster \| stars | indeed if you are in trouble then who will help you out here at super **bastard** quest without a doubt |

Table 6: Examples of incorrectly fixed snippets. In the left column, the hallucinated taboo-word is followed by the ground truth, and the top prediction using cloze test. In the right column, the example is presented with the taboo word marked in bold.

## Conclusions and Discussions

In this paper, we have found a disturbing result that commercial ASR systems may hallucinate taboo-words in video content for children. On a data set consisting of highly popular videos with worldwide consumption, we show that such hallucinations are far from occasional errors. We release a one-of-its-kind challenging data set of audio inputs where high-performance ASR systems have hallucinated taboo-words. We also show that some of these hallucinations can be corrected using language models.

Our work raises several important points to ponder.

***1. Which words are inappropriate for kids?*** Deciding on the set of inappropriate words for kids was one of the major design issues we ran into in this project. We considered several existing literature, published lexicons, and also drew from popular children's entertainment content. However, we felt that much needs to be done in reconciling the notion of inappropriateness and changing times. For example, we found that both $\mathcal{H}_1$ and $\mathcal{H}_2$ contain terms such as `gay` and `queer`. The field study that yielded $\mathcal{H}_2$ was conducted in early 1990s. Additionally, these words may or may not appear as abusive content based on the context it is present in. Since then, the continual struggle for LGBTQ+ rights and equality has made massive strides. Although queer studies is a developing field, Campo-Arias (2010) demonstrates that a child's age when they become aware of their sexual orientation varies, and it is possible that it could occur during childhood. In addition to this, children's attitudes toward queer people are also positively influenced by media exposure (Zhang, Feng, and Shen 2019), but they can also vary due to cultural differences (Bos, Picavet, and Sandfort 2012). We thus strongly feel that these lexicons need revisiting from experts to set better ethical guidelines for kids' content reflective of modern times.

***2. Risks of black-box AI systems.*** Several recent lines of work have reported instances where state-of-the-art content-filtering systems got blindsided by unseen (Sarkar and KhudaBukhsh 2021) or adversarial content (Gröndahl et al. 2018). Recent studies have revealed that biases in large language models often influence toxic content generation in neural text generation models (Gehman et al. 2020). In many such cases, we are dealing with opaque systems where it is impossible to know on what data these large systems are trained on, a risk aptly discussed in (Bender et al. 2021). At a philosophical level, we see our work making a small contribution in this growing discussion of responsible, inclusive, and trustworthy AI in the following key ways. First, we show that downstream AI applications can introduce highly inappropriate taboo-words in kids' content originally not present in them. The benefits of these ASR systems are undeniable. That said, we cannot disregard the fact that such systems, when applied to content with high visibility to a vulnerable and impressionable community, need rigorous checks and balances. Our findings, backed up with a challenging benchmark data set, is a small step towards that. Second, we do not know the distribution of kids' speech examples in the training data these opaque systems are trained on, nor do we know how well these data sets represent ESL (English as Second Language) speakers. However, our analysis of hallucinations reveal that many of the errors were caused in the presence of ESL speakers and kids. Our results thus potentially point to ways these data sets can be more inclusive. Finally, our work draws the attention of the community to form a deeper understanding of intermediate risks in a chain of black-box systems, where one system's outputs are inputs to another.

***3. Mitigation strategies.*** In our cloze test experiments to fix some of these hallucinated taboo-words, we notice that language models often have a propensity towards predicting the taboo-word as the most likely completion. For instance, according to `BERT`, a constrained cloze test *I love [MASK]* with candidate sets {`porn`, `corn`} yields `porn` as the likelier completion. In fact, 16.90% (Google) and 14.95% (Amazon) of top predictions by the Megatron model were taboo-words. This indicates that although language models can bring in improvement, they alone cannot fix the problem as these models also possibly suffer from a similar issue of being trained on content largely meant for an adult audience as opposed to kids.

While we observe limited success in fixing some of the hallucinated taboo-words, our experiments revealed potential avenues for improvement. We observe that many hallucinated audio inputs had visual signals that can be leveraged. For instance, in examples where `crab` is confused with `crap`, object recognition information can complement textual information to correct such mistakes. A multimodal method to robustify ASR systems could be a worthy future research challenge. Our experiments with language models produced a modest **p@1** improvement. However, a better **p@10** performance indicates that a human-in-the-loop setting, coupled with suggestions from language models, especially given these contents are consumed by kids worldwide, can offer more safety to kids.

***4. Integration challenges between YouTube Kids and general YouTube.*** YouTube Kids allows keyword-based search if parents (or guardians) enable it in the application. Of the five highly inappropriate taboo-words, {`shit`, `fuck`, `crap`, `rape`, `ass`}, we find that `rape`, `fuck`, and `shit` are not searchable through the kids app (understandably). We also find that most English language subtitles (including subtitles with many hallucinated taboo-words) are disabled on the kids app. However, as shown in Figure 6, the same videos have subtitles enabled on general YouTube. It is unclear how often kids are only confined to the YouTube Kids app while watching videos and how frequently parents (or guardians) simply let them watch kids' content from general YouTube. Our findings indicate a need for tighter integration between YouTube general and YouTube Kids to be more vigilant about kids' safety.

## Acknowledgements

We acknowledge the support received from Srini Raju Centre for IT and the Networked Economy (SRITNE) at the Indian School of Business (ISB). We thank ISB Research Assistant, Mallikarjuna Tupakula's help in the data collection and acknowledge Philip G. Bittenbender for his inputs in annotation.

# References

Alghowinem, S. 2018. A Safer YouTube Kids: An Extra Layer of Content Filtering Using Automated Multimodal Analysis. In *Proceedings of SAI Intelligent Systems Conference*, 294–308. Springer.

Arisoy, E.; Sethy, A.; Ramabhadran, B.; and Chen, S. 2015. Bidirectional recurrent neural network language models for automatic speech recognition. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5421–5425. IEEE.

Bender, E. M.; Gebru, T.; McMillan-Major, A.; and Shmitchell, S. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, 610–623.

Benzeghiba, M.; De Mori, R.; Deroo, O.; Dupont, S.; Erbes, T.; Jouvet, D.; Fissore, L.; Laface, P.; Mertins, A.; Ris, C.; et al. 2007. Automatic speech recognition and speech variability: A review. *Speech communication*, 49(10-11): 763–786.

Bos, H. M. W.; Picavet, C.; and Sandfort, T. G. M. 2012. Ethnicity, Gender Socialization, and Children's Attitudes Toward Gay Men and Lesbian Women. *Journal of Cross-Cultural Psychology*, 43(7): 1082–1094.

Campo-Arias, A. 2010. Essential aspects and practical implications of sexual identity. *Colombia MÃ*, 41: 179 – 185.

Chen, X.; Ragni, A.; Liu, X.; and Gales, M. J. 2017. Investigating bidirectional recurrent neural network language models for speech recognition. In *Proceedings of Interspeech 2017*, 269–273. International Speech Communication Association (ISCA).

Chi, Z.; Huang, S.; Dong, L.; Ma, S.; Singhal, S.; Bajaj, P.; Song, X.; and Wei, F. 2021. XLM-E: Cross-lingual Language Model Pre-training via ELECTRA. arXiv:2106.16138.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 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)*, 4171–4186.

Finley, G.; Edwards, E.; Robinson, A.; Brenndoerfer, M.; Sadoughi, N.; Fone, J.; Axtmann, N.; Miller, M.; and Suendermann-Oeft, D. 2018. An automated medical scribe for documenting clinical encounters. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, 11–15.

Gaur, Y.; Lasecki, W. S.; Metze, F.; and Bigham, J. P. 2016. The Effects of Automatic Speech Recognition Quality on Human Transcription Latency. In *Proceedings of the 13th International Web for All Conference*, W4A '16.

Gehman, S.; Gururangan, S.; Sap, M.; Choi, Y.; and Smith, N. A. 2020. RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. In Cohn, T.; He, Y.; and Liu, Y., eds., *Findings of the Association for Computational Linguistics: EMNLP 2020*, volume EMNLP 2020 of *Findings of ACL*, 3356–3369.

Gröndahl, T.; Pajola, L.; Juuti, M.; Conti, M.; and Asokan, N. 2018. All you need is" love" evading hate speech detection. In *Proceedings of the 11th ACM workshop on artificial intelligence and security*, 2–12.

Han, W.; and Ansingkar, M. 2020. Discovery of Elsagate: Detection of Sparse Inappropriate Content from Kids Videos. In *2020 Zooming Innovation in Consumer Technologies Conference (ZINC)*, 46–47.

Jay, K. L.; and Jay, T. B. 2013. A child's garden of curses: A gender, historical, and age-related evaluation of the taboo lexicon. *The American Journal of Psychology*, 126(4): 459–475.

Jay, T. 1992. *Cursing in America*, volume 10. Philadelphia: John Benjamins.

Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y.; Madotto, A.; and Fung, P. 2022. Survey of Hallucination in Natural Language Generation. *arXiv preprint arXiv:2202.03629*.

Kayhan, O. S.; Vredebregt, B.; and van Gemert, J. C. 2021. Hallucination In Object Detection—A Study In Visual Part VERIFICATION. In *2021 IEEE International Conference on Image Processing (ICIP)*, 2234–2238. IEEE.

Kim, J. Y.; Liu, C.; Calvo, R. A.; McCabe, K.; Taylor, S. C.; Schuller, B. W.; and Wu, K. 2019. A comparison of online automatic speech recognition systems and the nonverbal responses to unintelligible speech. *arXiv preprint arXiv:1904.12403*.

Mehrabani, M.; Bangalore, S.; and Stern, B. 2015. Personalized speech recognition for Internet of Things. In *2015 IEEE 2nd World Forum on Internet of Things (WF-IoT)*, 369–374. IEEE.

Mengistu, K. T.; and Rudzicz, F. 2011. Comparing Humans and Automatic Speech Recognition Systems in Recognizing Dysarthric Speech. In *Proceedings of the 24th Canadian Conference on Advances in Artificial Intelligence*, Canadian AI'11, 291–300.

Namazifar, M.; Malik, J.; Li, L. E.; Tür, G.; and Hakkani-Tür, D. 2021. Correcting Automated and Manual Speech Transcription Errors using Warped Language Models. *CoRR*, abs/2103.14580.

Palakodety, S.; KhudaBukhsh, A. R.; and Carbonell, J. G. 2020. Mining Insights from Large-Scale Corpora Using Fine-Tuned Language Models. In *ECAI 2020 - 24th European Conference on Artificial Intelligence*, volume 325 of *Frontiers in Artificial Intelligence and Applications*, 1890–1897.

Papadamou, K.; Papasavva, A.; Zannettou, S.; Blackburn, J.; Kourtellis, N.; Leontiadis, I.; Stringhini, G.; and Sirivianos, M. 2020. Disturbed YouTube for kids: Characterizing and detecting inappropriate videos targeting young children. In *Proceedings of the international AAAI conference on web and social media*, volume 14, 522–533.

Petroni, F.; Rocktäschel, T.; Riedel, S.; Lewis, P.; Bakhtin, A.; Wu, Y.; and Miller, A. 2019. Language Models as

Knowledge Bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2463–2473.

Plantinga, P.; and Fosler-Lussier, E. 2019. Towards Real-Time Mispronunciation Detection in Kids' Speech. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 690–696. IEEE.

Rajnoha, J.; and Pollák, P. 2011. ASR systems in noisy environment: Analysis and solutions for increasing noise robustness. *Radioengineering*, 20(1): 74–84.

Ranchal, R.; Taber-Doughty, T.; Guo, Y.; Bain, K.; Martin, H.; Robinson, J. P.; and Duerstock, B. S. 2013. Using speech recognition for real-time captioning and lecture transcription in the classroom. *IEEE Transactions on Learning Technologies*, 6(4): 299–311.

Riccardi, G.; Gorin, A. L.; Ljolje, A.; and Riley, M. 1997. A spoken language system for automated call routing. In *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, 1143–1146. IEEE.

Rohrbach, A.; Hendricks, L. A.; Burns, K.; Darrell, T.; and Saenko, K. 2018. Object Hallucination in Image Captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 4035–4045.

Sanh, V.; Debut, L.; Chaumond, J.; and Wolf, T. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

Sarkar, R.; and KhudaBukhsh, A. R. 2021. Are Chess Discussions Racist? An Adversarial Hate Speech Data Set (Student Abstract). In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021*, 15881–15882.

Sarkar, R.; Mahinder, S.; and KhudaBukhsh, A. 2020. The Non-native Speaker Aspect: Indian English in Social Media. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, 61–70.

Sawaf, H. 2012. Automatic speech recognition and hybrid machine translation for high-quality closed-captioning and subtitling for video broadcast. *Proceedings of Association for Machine Translation in the Americas–AMTA*, 14.

Shin, J.; Lee, Y.; and Jung, K. 2019. Effective sentence scoring method using BERT for speech recognition. In *Asian Conference on Machine Learning*, 1081–1093. PMLR.

Shoeybi, M.; Patwary, M.; Puri, R.; LeGresley, P.; Casper, J.; and Catanzaro, B. 2019. Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism. *CoRR*, abs/1909.08053.

Sutton-Smith, B.; and Abrams, D. M. 1978. Psychosexual material in the stories told by children: The Fucker. *Archives of Sexual Behavior*, 7(6): 521–543.

Taylor, W. L. 1953. "Cloze procedure": A new tool for measuring readability. *Journalism quarterly*, 30(4): 415–433.

Vanderplank, R. 2016a. *Captioned media in foreign language learning and teaching: Subtitles for the deaf and hard-of-hearing as tools for language learning*. Springer.

Vanderplank, R. 2016b. 'Effects of'and 'effects with'captions: how exactly does watching a TV programme with same-language subtitles make a difference to language learners? *Language Teaching*, 49(2): 235–250.

Wu, F.; García-Perera, L. P.; Povey, D.; and Khudanpur, S. 2019. Advances in Automatic Speech Recognition for Child Speech Using Factored Time Delay Neural Network. In *Interspeech*, 1–5.

Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J. G.; Salakhutdinov, R.; and Le, Q. V. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In Wallach, H. M.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E. B.; and Garnett, R., eds., *Advances in Neural Information Processing Systems (NeurIPS)*, 5754–5764.

Yeung, G.; and Alwan, A. 2018. On the difficulties of automatic speech recognition for kindergarten-aged children. *Interspeech 2018*.

Zhang, S.; Feng, S.; and Shen, Z. 2019. How Do Background Factors Influence Children's Attitudes toward Gays and Lesbians? *Psychology*, 10: 1572–1594.

Zhang*, T.; Kishore*, V.; Wu*, F.; Weinberger, K. Q.; and Artzi, Y. 2020. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations (ICLR)*.