Comparing the Effectiveness of Fine-tuning vs. One-Shot Learning on the Kidz Bopification Task by Semantic Savants

Jeonghoon Kim
kimx5154@umn.edu

Matthew Olson
olso8286@umn.edu

Marco Berriodi
berri079@umn.edu

Abstract

Kidz Bopification refers to the process of removing explicit content from song lyrics and substituting it with sanitized alternatives. Our project's objective is to develop a language model capable of performing Kidz Bopification automatically. We utilized an existing lyrics API to extract data, preprocessed it, and trained two models employing different machine learning algorithms: a T5-based encoder-decoder model and a ChatGPT-based (GPT-3.5) decoder-only model. By comparing the performance of these models, our aim is to identify the most effective approach for Kidz Bopification. This research carries broader ethical and social implications, as technologies that censor or modify language can influence cultural expression and free speech. Nevertheless, our project offers a valuable resource for parents, educators, and others seeking to provide children with age-appropriate versions of popular songs. Ultimately, our project contributes to the broader discourse on content moderation in the digital era, promoting a more enjoyable experience for audiences worldwide.

1 Introduction

1.1 Motivation

Many popular songs contain explicit or inappropriate contents that may not be suitable for children. In recent years, the amount of explicit language in popular music has grown rapidly [Bannister, 2021], which can have negative implications for parents of young children who enjoy listening to music. Although many songs have clean versions available, these often only temporarily mute profanity, which disrupts the song's original message and rhythm and ignores less obvious forms of explicit lyrics like metaphors and sexual imagery.

For this reason, we developed a tool that can automatically clean song lyrics by removing explicit content and replacing it with age-appropriate substitutes. This tool would enable parents, educators, and others to provide children with versions of popular songs that are safe and appropriate for their age. In addition, we want to explore the broader implications of creating technology that can modify or censor language and consider the potential ethical and social impacts of such tools. By exploring the ethical and social implications of creating technology that can modify or censor language, we aim to provide parents and educators with a tool that can provide children with safe and appropriate versions of popular songs while promoting positive language and meaning.

There is one organization currently addressing this problem in a more thorough way: Kidz Bop, an American children's music group that releases compilation albums in which children cover versions of popular songs with explicit material substituted for cleaner lyrics. Therefore, we dub the process of identifying explicit lyrics and generating appropriate substitute lyrics that, if possible, maintain the original meaning and flow of the song, "Kidz Bopification".

Manual Kidz Bopification is time-consuming and contextually challenging. In order to reduce the time and cost, the use of NLP tools to develop a model that can automatically Kidz Bopify any set of lyrics is a promising direction.

1.2 Future impact and Benefits

Our work on Kidz Bopification may appeal to a diverse range of stakeholders, including parents, educators, music industry professionals, and researchers in natural language processing and content moderation. Each of these groups provides a unique opportunity to make an impact.

First, our work can provide a means to produce more family-friendly content within the music industry. This can lead to increased revenue for artists and record labels, as well as more opportunities for the production of music that is suitable for all ages. Our tool could also benefit companies such as Kidz

Bop, which require editors to manually revise all lyrics, as our tool could significantly aid in this process.

Secondly, our work addresses concerns among parents and educators about children's exposure to inappropriate content in popular music. By offering a tool that automatically sanitizes song lyrics, we empower parents and educators to create safer, more suitable listening experiences for children, potentially fostering positive values and attitudes in young listeners. This avenue is especially promising when combined with AI-music generation techniques, whereby a song could automatically be made clean by first running the lyrics through a Kidz Bopification model and generating a new song using those lyrics.

Finally, our work can contribute to the broader conversation around content moderation in the digital age. By exploring the ethical and social implications of developing technology that modifies or censors language, we can contribute to a more informed and nuanced understanding of these issues. Our tool can also serve as a starting point for the development of similar tools for other types of media, such as movies or television shows. Overall, our work has the potential to make a significant impact on the way that music is produced and consumed, and on the way that technology is used to regulate content in the digital age.

2 Literature Survey

Researchers have attempted various approaches to sanitizing song lyrics for children, with the most common method involving manual editing by human editors, e.g. Kidz Bop. However, this approach is time-consuming, labor-intensive, and often subjective, leading to inconsistencies in the level of sanitization across different songs. Automated approaches to sanitizing lyrics have also been attempted but have not been widely applied in the music industry, and they have been limited in their effectiveness due to the complexity and nuances of language. Most of these approaches rely on keyword matching or rule-based methods, which may not accurately capture the meaning of the original lyrics or produce natural-sounding substitutes [Bannister, 2021].

One of the main challenges of the current practice is achieving a balance between sanitization and preservation of the original meaning and structure of the lyrics. Additionally, the use of automated tools raises ethical and social questions about the potential impact on cultural expression and free speech, as well as concerns around the accuracy and appropriateness of the substitutes generated by the tool [Zhu et al., 2021].

With children's increasing exposure to potentially explicit music through streaming services and social media, concerns have been raised about the impact of explicit content in song lyrics on children's development. Studies have shown a positive correlation between children's exposure to profanity in media and aggression [Coyne et al., 2011]. Researchers have explored the use of machine learning and natural language processing techniques for detecting explicit content in lyrics. Several attempts have been made to classify music as explicit or clean, but little research has been done on using NLP models to automatically alter lyrics to make them less explicit [Zeng et al., 2021, Chin et al., 2018, Rospocher and Eksir, 2023].

Our proposed approach aims to address some of these limitations by developing a language model that can accurately and efficiently perform Kidz Bopification on real-world data while preserving the original meaning and structure of the lyrics [Zhu et al., 2021].

3 Problem Definition

Our ultimate goal is to create a model that can Kidz Bopify with a high degree of accuracy, quality, and scalability in a reasonable time frame. However, this leads to a problem: there are multiple strategies for building such a model, so which is the best option? In the current paradigm, there exist two major strategies for completing an NLP task: fine-tuning an existing model for the downstream task and using in-prompt learning with a large language model (LLM) like GPT-3.5 to complete the task. Thus, this report is devoted to comparing the effectiveness of fine-tuning and one-shot learning on the Kidz Bopification task.

3.1 Task Definition

The fundamental task for our models is to take in the lyrics of a song or snippet of the song and output a version of those lyrics where the explicit material, including non-obvious material such as innuendo, is replaced with inoffensive substitute lyrics. Furthermore, the general semantic meaning (assuming it is not offensive), as well as the rhythm and flow of the lyrics should be maintained as much as possible.

4 Approach

Our team pursued two methods to perform the Kidz bopification task: utilizing ChatGPT (specifically the GPT-3.5-turbo chat completions from the OpenAI API) and fine-tuning the T5 (Text-To-Text Transfer Transformer) model. Prior to building the models, we extracted original and clean Kidz Bop versions of lyrics from over 1100 popular songs using the Genius API. Following data preprocessing and T5 model training, we evaluated the performance of both ChatGPT and T5 models by running them on a test set to assess their ability to generate kid-friendly lyrics.

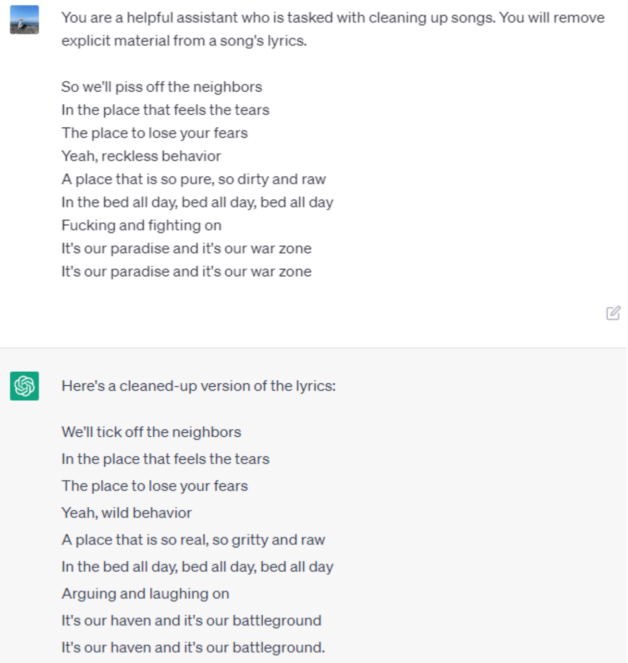
Our approach is novel in its use of advanced NLP models to generate kid-friendly versions of popular song lyrics. Previous attempts to "clean up" song lyrics for children have often relied on manual editing or simple rules-based approaches. In contrast, our approach leverages the latest advances in deep learning and NLP to automatically generate new versions of the lyrics based on large amounts of existing data. With our method, we can avoid the time-consuming job of manual editing. Additionally, by utilizing technologies like "Uber-Duck.ai" to create songs that mimic the voices of real singers, we can broaden the scope of our approach to the audio itself rather than just lyrics. This could also lead to application in other fields, including video production, while also potentially applying a real-time profanity filter to TV, audio, or video content.

4.1 Preprocessing

For data preprocessing, we divided the 1100 songs into matching verses and choruses to increase the sample size and decrease the sample length. Next, we removed unnecessary and repeating parts such as '[verse]' and '[chorus]' because such data could potentially affect the T5 model's performance. Given the large volume of data, manually checking all 2864 extracted verses/choruses from the 1100 songs was not feasible. However, we manually reviewed a representative sample of 25 percent of the data to ensure its quality.

4.2 Method 1: ChatGPT

For the ChatGPT version, we first split the preprocessed data into training (2291 samples) and test (573 samples) sets. We then utilized the OpenAI API to generate Kidz Bopified versions of lyrics with a one-shot prompt. The example given to ChatGPT was chosen by using TFID to determine the most similar verse in the training set. We experimented with various task prompts to identify ones that effectively sanitized the lyrics without altering the original content too much. We anticipated that ChatGPT would perform well due to its extensive training on a vast corpus of diverse text data, including social media, web pages, and books. As a result, it has an excellent understanding of natural language and can produce coherent and grammatically correct sentences. Furthermore, ChatGPT enables flexible and customizable text generation, which is advantageous for our task since we need to replace certain words or phrases with kid-friendly alternatives while preserving the overall meaning and structure of the original lyrics.



You are a helpful assistant who is tasked with cleaning up songs. You will remove explicit material from a song's lyrics.

So we'll piss off the neighbors
In the place that feels the tears
The place to lose your fears
Yeah, reckless behavior
A place that is so pure, so dirty and raw
In the bed all day, bed all day, bed all day
Fucking and fighting on
It's our paradise and it's our war zone
It's our paradise and it's our war zone

Here's a cleaned-up version of the lyrics:

We'll tick off the neighbors
In the place that feels the tears
The place to lose your fears
Yeah, wild behavior
A place that is so real, so gritty and raw
In the bed all day, bed all day, bed all day
Arguing and laughing on
It's our haven and it's our battleground
It's our haven and it's our battleground.

Figure 1: Example of the Result of the ChatGPT based on the prompt; "You are a helpful assistant who is tasked with cleaning up songs. You will remove explicit material from a song's lyrics: [Verse from Pillow Talk by Zayn]"[OpenAI, 2023]

4.3 Challenges of ChatGPT

While Figure 1 shows that ChatGPT performs well, there were still instances where it was not able to capture certain nuances in the original lyrics. Moreover, since the models rely on existing data, there is a risk of perpetuating biases and stereotypes that may have been present in the source lyrics. This is because ChatGPT is a language

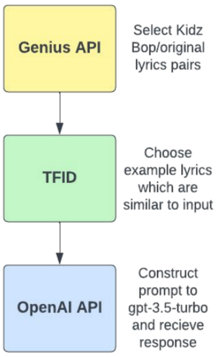

Genius API
Select Kidz Bop/original lyrics pairs
TFID
Choose example lyrics which are similar to input
OpenAI API
Construct prompt to gpt-3.5-turbo and recieve response

Figure 2: ChatGPT Method Workflow

model that has been trained on a massive amount of diverse text data. Additionally, the accuracy of the models depends on the quality and diversity of the prompts, which may be limited or biased in certain ways. In addition to that, the task of "kidz bop" songs is subjective, and opinions on what constitutes appropriate language for children may vary. Based on our analysis, we found that the ChatGPT model occasionally overcorrected lyrics that were not explicitly inappropriate, such as substituting the word "Son" with "kid". Despite these limitations, it was relatively straightforward to implement the function utilizing the ChatGPT API since we only needed to experiment with different prompts to generate the most optimal results.

4.4 Method 2: T5

For the T5 version, we split the preprocessed data into training (2291 samples) and test (573 samples) sets and fine-tuned the T5 model with the training data. We expected the T5 model would also show decent performance because T5 (Text-to-Text Transfer Transformer) is a powerful language model that has been pre-trained on a large amount of text data, making it well-suited for a variety of natural language processing tasks. Its ability to perform text-to-text transfer allows it to be fine-tuned for specific tasks, such as the Kidz bopification task, by providing input-output pairs during training. T5 also has the ability to handle sequence-to-sequence tasks, which is important for the task of transforming lyrics into a kid-friendly version while preserving the meaning and structure of the original lyrics. Overall, T5's flexibility and pre-training make it a strong candidate for the Kidz bopification task.

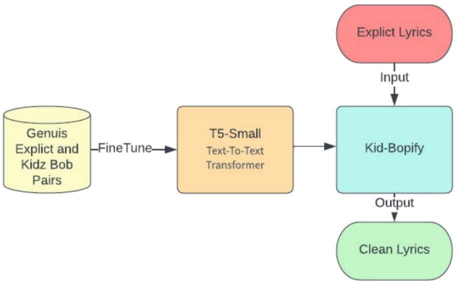


Figure 3: Fine-tuning T5 Model Method Workflow

4.5 Challenges of the T5 Model

Although T5 is a powerful language model, we encountered some limitations during the training process. One of the main challenges was determining appropriate hyper-parameters for our task, such as the learning rate, batch size, and maximum sequence length. This required extensive experimentation and parameter-tuning, which can be time-consuming and computationally expensive.

Additionally, due to the size and complexity of the T5 model, it required a significant amount of computational resources to train effectively. Training on over 2000 data for 2 epochs took more than 6 hours to complete. If we had better resources or utilized cloud-based services like Google Cloud Platform, which can be expensive, we would have been able to conduct further research on the model and enhance the performance.

Finally, we also encountered some limitations in the quality of the generated Kidz bop lyrics, particularly in cases where the original lyrics contained complex language or wordplay. If a song contained too much explicit content, sometimes the Kidz Bop version lyrics had to be changed to an entirely new set of lyrics, deviating from the original song while preserving the rhythm/beat. Therefore, there may be limits to the ability of these models to fully capture the nuances of human language and creativity, which also contains an issue with the quality of data since it does not partially change the lyrics.

4.6 Evaluation Metrics

In evaluating our models with GPT-3.5 (ChatGPT) and T5 models for sanitizing kid-friendly versions of popular song lyrics, we focused primarily on human evaluation. Although the process of human evaluation is subjective, we believed that it would be more accurate in terms of evaluating the meaning and potential explicitness of the lyrics. Also, we thought automatic evaluation is necessary too because it is efficient and objective, and can provide a quick way to compare the performance of different models or techniques. Our goal was to ensure that the generated lyrics were not only appropriate for children, but also maintained the essence and meaning of the original lyrics. Therefore, we thought of using both human and automatic evaluation methods to provide a more comprehensive and nuanced assessment of the effectiveness of our Kidz Bopification model. By combining the strengths of both evaluations, we can gain a more comprehensive understanding of the performance of our model and identify ways to further improve it.

For human evaluation, 60 samples were evaluated (20 by each of us) with 4 important aspects of lyrics: 1) explicitness, 2) retaining original meaning, 3) retaining flow/rhythm, and 4) retaining original length. We evaluated on a linear scale from 0-5.

For automatic evaluation, 100 samples were evaluated between the test (kidz bop lyrics) and output from ChatGPT and T5 model. We chose to utilize both the BLEU score and the BERT score to evaluate our model's performance. Although the BLEU score has its limitations as it only considers lexical similarity and disregards meaning and context, it remains a valuable tool in assessing the quality of generated text. Furthermore, the BERT score is another popular evaluation metric used to measure the similarity between two pieces of text. It takes into account both the content overlap as well as the fluency of the generated text. BERT score is particularly useful when evaluating text generated by deep learning models, as it can help identify cases where the generated text is grammatically correct but semantically incorrect. Using the BERT score alongside the BLEU score can provide a more comprehensive evaluation of the generated text, as the BLEU score is based on n-gram overlap and does not capture the semantic meaning of the text. By incorporating the BERT score into your evaluation process, you can ensure that your generated text is not only grammatically correct but also semantically accurate.

5 Results

5.1 Human Evaluations

As mentioned earlier, we employed four distinct metrics for the human evaluation phase of our project to assess the effectiveness of our model. These criteria encompassed:

1. **Explicitness:** gauges the degree to which potentially offensive content has been filtered out.

2. **Retaining the original meaning:** measures how well the model preserves the core message and intent of the lyrics.

3. **Retaining original rhythm:** evaluating the model's ability to maintain the musical flow and structure of the lyrics.

4. **Retaining original length:** an assessment of how closely the model-generated lyrics match the original lyrics in terms of overall length.

By considering these four dimensions, we aimed to provide a comprehensive and nuanced evaluation of our model's performance in handling diverse aspects of lyrical content. We used these metrics on 60 samples of one verse or hook of a song from our dataset. To ensure comprehensive coverage of songs, we distributed the evaluation task among team members, with each member assigned to evaluate 20 samples based on the four criteria.

As illustrated in Table 1, the explicitness scores for both models were notably low, indicating effective explicit content filtering. During the human evaluation process, we noticed that the Kidz Bop version lyrics in our test set predominantly included low to moderate levels of explicitness. While both models may not be perfect, the generated output from both models effectively filtered out explicit content to a reasonable extent. Based on the results of our human evaluation, we found that the ChatGPT prompting methods outperformed the Fine-tuned T5 model. The ChatGPT approach not only effectively replaced explicit content but also preserved the original meaning, rhythm, flow, and length of the lyrics.

5.2 Automatic Evaluations

Similarly, for automatic evaluations, we utilized different types of evaluation metrics, specifically the BLEU and BERT scores, to gain a more comprehensive understanding of the models' performance. we observed a similar trend in the automatic evaluations, where the performance of ChatGPT surpassed that of the fine-tuned T5 model. This is because BLEU's output is always a number between 0 and 1 [Lin and Och, 2004]. This value

Methods	Explicit	Retains Meaning	Retains Rhythm	Retains Length
ChatGPT Prompting	0.049	3.98	4.31	4.72
Fine-tuned T5 Model	0.54	2.49	2.26	1.98

Table 1: Results of human evaluation of 60 samples. Each human evaluation score is on a 0-5 scale: Explicitness (0 = None, 5 = Heavy), Retains original meaning (0 = Not at all, 5 = completely), Retains flow/rhythm (0 = Not at all, 5 = completely), Retains original length (0 = Not at all, 5 = completely).

Methods	Avg BLEU	Median BLEU	Avg BERT	Median BERT
ChatGPT Prompting	0.38	0.35	0.82	0.85
Fine-tuned T5 Model	0.32	0.17	0.75	0.76

Table 2: Average results of automatics evaluation of 100 samples. BLEU and BERT scores were calculated automatically based on the original Kidz Bop lyrics.

indicates how similar the candidate text is to the reference texts, with values closer to 1 representing more similar texts. Therefore, if BLEU is closer to 1, it means that the length, and meaning would be similar to the original text. Plus, BERTScore is an automatic evaluation metric that calculates a similarity score between each token in the generated sentence and each token in the reference sentence. By utilizing BERTScore, we can assess the generated text's similarity to the reference text at a deeper level, taking into account nuances in meaning and context. This is particularly valuable in tasks where maintaining the meaning and coherence of the text is crucial, such as generating kid-friendly lyrics. BERTScore's ability to capture semantic information enhances its effectiveness in evaluating the quality of the generated lyrics and provides a more accurate measure of performance compared to simpler lexical-based metrics. It leverages the pre-trained contextual embeddings from BERT models and determines the similarity between words in the generated and reference sentences using cosine similarity. Considering the higher BERTScore values for ChatGPT compared to the fine-tuned T5 model, it indicates a higher level of quality and semantic similarity between the generated and reference text. This suggests that ChatGPT outperforms the fine-tuned T5 model in terms of text generation. Therefore, the higher BERTScore values achieved by the ChatGPT prompting method serve as evidence that ChatGPT performs better than the fine-tuned T5 model.

5.3 Assessing Outcomes

After extensive evaluation through both automatic and human assessments, we arrived at the conclusion that ChatGPT outperformed our attempts to develop a fine-tuned T5 model. We found some test samples that failed to preserve the original rhythm, meaning, and length of the songs. This is important since preserving the original rhythm meaning and length could potentially change the original song somehow. For example, we observed that the model encountered difficulties in generating consistent and accurate outputs. In some cases, it would repetitively output the same lyrics or simply reproduce the input itself, which was acceptable when the input was not explicit. However, the model performed poorly when confronted with repeating lyrics, such as:

- **Original:**

  you prolly think that you are better now, better now you only say that 'cause i'm not around, not around you know i never meant to let you down, let you down woulda gave you anything, woulda gave you everything

- **Fine-tuned T5 Model:**

  cause i'm not around, not around you know I never meant to let you down, let you down woulda gave you anything, woulda gave you anything, woulda gave you anything, woulda gave you anything, woulda gave you anything, woulda gave you anything,

- **ChatGPT Prompting:**

  You probably think that you are awesome now, awesome now You only say that 'cause I'm not in town, not in town You know I never meant to bring you down, bring you down Would have given anything, would have given everything

The generated output from fine-tuned T5 model, like the example above, is not sufficient. We believe that there are three main reasons for the incorrect output from the T5 model. Firstly, the model might not have been adequately trained due to limited computational power and memory, preventing the use of larger T5 models. Secondly, our dataset could have been improved as some verses did not align perfectly with the Kidz Bop versions, resulting in significant changes in song lyrics. Lastly, the model would benefit from more data as there are numerous nuanced steps involved that the model could potentially learn from [Jacob, 2023]. The solutions to these problems are straightforward. If we have a larger dataset and enough resources on GPU and RAM, we could have potentially improved the performance.

On the other hand, the generated output from ChatGPT performed relatively well, although it occasionally made unnecessary changes to certain words. To address this minor issue, we could experiment with different prompts to limit unnecessary changes. However, it is important to note that apart from this minor concern, ChatGPT demonstrated outstanding performance, surpassing our expectations and outperforming the T5 model.

6 Analysis

During the human evaluation process, we observed that our dataset lacked an ideal balance between explicit and non-explicit songs. Instead of featuring a diverse mix of highly explicit, moderately explicit, and non-explicit songs, our dataset predominantly consisted of moderately explicit or non-explicit tracks. This limitation can be attributed to the fact that our testset only includes songs with Kidz Bop versions. Consequently, as illustrated in Table 1, the explicitness scores for both models were notably low, indicating effective explicit content filtering.

For the rest of the metrics, we found ChatGPT one-shot Prompting had clear better results and was much more reliable and consistent results. ChatGPT was very good at retaining the meaning, rhythm, and length of the songs with scores averaging 4 or more. The fine-tuned model really struggled at this as it either removed the explicitness from the song and did not retain any structure of the song or did not change anything about the song due to no explicit material. The fine-tuned model also sometimes would get in loops that caused it to continually repeat itself.

6.1 Replicability

Others can easily replicate our findings with access to the Genius API, ChatGPT, and a basic understanding of the Hugging Face library and NLP. To reproduce our results, one can follow the steps outlined in Figures 2 and 3. All our code can be found at https://github.umn.edu/semantic-savants/kidz-bopify.

By following the approaches in Figure 2 and 3 and leveraging these resources, others should be able to achieve comparable results to our study. Since the existing dataset from Genius API has over 1100 songs that have original and Kidz bop versions in pairs, it would be easy to replicate. However, the preprocessing plan should be made after analyzing the data.

6.2 Discussion

While conducting our research and encountering unexpected limitations during the training of the T5 model, we came up with the idea of incorporating heuristics to identify explicit content, such as implementing a "*profanity filter*" and replacing such content using Word2Vec. This is because Word2vec represents each word by a vector. The vectors learned by word2vec preserve the similarity of words, that is vectors of dog and canine will be more similar than vectors of cat and canine. Word2vec learns these representations from a large corpus of text, basically, the words that co-occur together end up having similar vector representations[Birajdar, 2021]. Therefore, although it may not fully comprehend the entire context, we anticipate that Word2Vec could still excel in replacing explicit words, making it worth exploring as a potential solution.

Another thing to discuss is the ethics of this Kidz Bopify on the world and the effect it can have on the world. The first thing is free speech as this tool in the future can easily just filter out all the bad words of what someone is saying or singing and change what they are saying. This would prohibit people from using free speech as they would be constantly censored. The next ethics is Bias to certain types of speech or just one speaking style. Lastly is the use of this data as we did not create any of these songs and are using other people's songs and ideas without asking them.

To combat these ethical problems we would monitor the use of this model to make sure its fair use and no one or entity is misusing this. We would also try to use all different types of songs to try to have this not have any bias. Lastly, we would have a system for this tool for song artists to opt out of the use of their songs as data and maybe compensate the artist for the use of their songs as data in this project.

7 Conclusion

Our project aimed to solve the problem of how best to develop an automated language model capable of Kidz Bopification. For this task, we compared ChatGPT prompting techniques and fine-tuning on a T5 model, evaluating their performances through human and automated evaluations. The human evaluation focused on four main aspects: clarity, original meaning, retention of flow/rhythm, and retention of the original length. Our results showed that the ChatGPT query method outperformed the adjusted T5 model by effectively replacing explicit content while preserving the song's original meaning, rhythm, flow, and duration. We also performed automated evaluations using 100 samples, which incorporated BLEU and BERT scores to measure the performance of our models. Consistent with the findings of our human evaluation, the ChatGPT model demonstrated superior performance to the fine-tuned T5 model.

Although our models have made considerable progress in the Kidz Bopification process, there is still a lot of room for improvement. We found instances where the model could not preserve the songs' original rhythm, feel, and length, which could alter the original song entirely. However, our work provides a valuable starting point for creating standard versions of popular songs. It has the potential to benefit a wide range of stakeholders, including parents, educators, music industry professionals, and researchers in natural language processing and content moderation. As we continue to refine our models and explore new techniques, our goal is to improve the performance and applicability of our Kidz Bopification tool, making it a valuable resource for parents, educators, and the music industry to be able to filter out explicit lyrics from songs. Additionally, we hope to contribute to the broader discourse on content moderation in the digital age, promoting a more enjoyable and age-appropriate experience for audiences worldwide.

References

Mark Bannister. dand-p4-billboard, Mar 2021. URL https://github.com/mspbannister/dand-p4-billboard/blob/master/Billboard_analysis__100417_.md#the-billboard-hot-100-exploring-six-decades-of-number-one-singles.

Nikhil Birajdar. Word2vec research paper explained, 2021. URL https://towardsdatascience.com/word2vec-research-paper-explained-205cb7eecc30.

Hyojin Chin, Jayong Kim, Yoonjong Kim, Jinseop Shin, and Mun. Y. Yi. Explicit content detection in music lyrics using machine learning. In *2018 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 517–521, 2018. doi: 10.1109/BigComp.2018.00085.

S. M. Coyne, L. A. Stockdale, D. A. Nelson, and A. Fraser. Profanity in media associated with attitudes and behavior regarding profanity use and aggression. *PEDIATRICS*, 128(5):867–872, Oct 2011. doi: https://doi.org/10.1542/peds.2011-1062.

John Jacob. What is flan-t5? is flan-t5 a better alternative to gpt-3?, Feb 2023. URL https://exemplary.ai/blog/flan-t5.

Chin-Yew Lin and Franz Josef Och. ORANGE: a method for evaluating automatic evaluation metrics for machine translation. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 501–507, Geneva, Switzerland, aug 23–aug 27 2004. COLING. URL https://www.aclweb.org/anthology/C04-1072.

OpenAI. Chatgpt, 2023. URL https://chat.openai.com/.

Marco Rospocher and Samaneh Eksir. Assessing fine-grained explicitness of song lyrics. *Information*, 14 (3), 2023. ISSN 2078-2489. doi: 10.3390/info14030159. URL https://www.mdpi.com/2078-2489/14/3/159.

Mingliang Zeng, Xu Tan, Rui Wang, Zeqian Ju, Tao Qin, and Tie-Yan Liu. Musicbert: Symbolic music understanding with large-scale pre-training. *CoRR*, abs/2106.05630, 2021. URL https://arxiv.org/abs/2106.05630.

Hongyuan Zhu, Ye Niu, Di Fu, and Hao Wang. Musicbert: A self-supervised learning of music representation. *ACM*, 1:3955–3963, 2021. URL https://dl.acm.org/doi/abs/10.1145/3474085.3475576?casa_token=6rP6g_jZvDEAAAAA:vXmzykF-Hk9WkALH6vp2BnsvTsdwTiFXbqaff5Z0DDCnUD_XdjxB_pg8yfLuYl0Uqi3Oce7GlRQ.