Toxic In-Game Voice Chat Moderation using Multimodal LLMs

RADHIKA SINGH

May 24, 2025

A capstone project report

submitted to the Faculty of the Graduate School

The University at Buffalo, State University of New York

In partial fulfillment of the requirements for the degree of

Master of Science

To my family and friends

Contents

List of Tables

List of Figures

Acknowledgments

I extend my sincere gratitude to the Computer Science Department of the University at Buffalo, SUNY, and Dr. Hongxin Hu for providing me with an opportunity to pursue my research interests in the area of Responsible and Safe AI. Special appreciation goes to my mentor and friend Keyan Guo for his timely help and input on the project. I dedicate this work to my family and friends who have supported my pursuits.

Chapter 1

Abstract

Voice chat moderation in online multiplayer games presents a unique challenge due to the dynamic, unstructured nature of spoken communication. Unlike text, voice is ephemeral, harder to log, and significantly more resource-intensive to process. As platforms like Roblox adopt voice chat to enhance interactivity, ensuring a safe environment for younger players has become critical. Existing moderation systems either rely heavily on keyword filtering or simplistic models, often failing to detect nuanced toxic behavior, leading to both false positives and negatives.

This project addresses these limitations by leveraging Multimodal Large Language Models (MLLMs), specifically the GPT-4o audio model, to develop a robust, real-time voice moderation pipeline. Using benchmark datasets like DeToxy [5] and MuTox [3], we evaluated the performance of traditional, commercial, and open-source models across precision, recall, and F1-score. Results showed that existing systems, including Roblox's in-game moderation and AWS Transcribe, underperformed in balancing accuracy with recall.

We introduce a novel architecture combining Few-Shot and Chain-of-Thought (CoT) prompting techniques, allowing the model to reason through toxicity with high interpretability. The architecture includes category-specific definitions, examples, and tone analysis to improve multilabel classification. Preliminary experiments show significant improvements, achieving an F1-

score of 0.75 with GPT-4o + CoT prompting, outperforming all baselines. This approach demonstrates promise in building safer, AI-powered moderation tools for real-time voice communication.

Chapter 2

Introduction

Multiplayer cooperative and competitive games have emerged as a dominant genre in the gaming industry, attracting millions of players worldwide. Due to their high degree of interactivity, where players can engage with one another in real-time, these games are particularly popular among younger audiences. Platforms like Roblox, which provide immersive and customizable gaming experiences, have taken player communication to the next level by incorporating voice chat features.

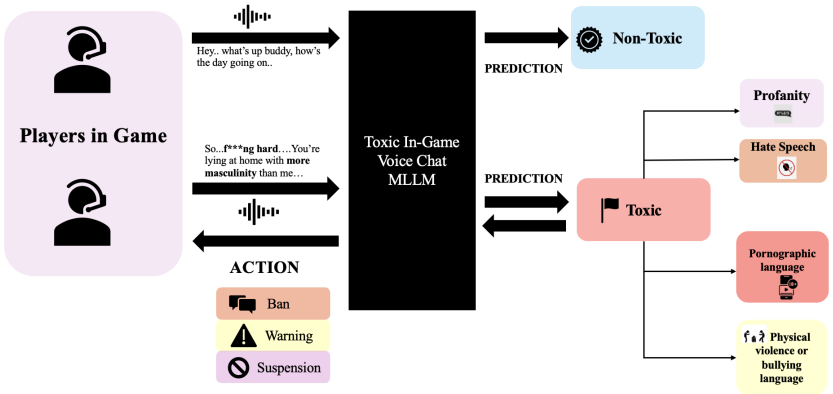

Figure 2.1: Roblox In-Game Voice Chat Moderation System

However, the introduction of voice chat brings significant safety concerns.

Unlike text, which can be quickly monitored and filtered based on specific words, voice chat is far less structured and difficult to control. Voice moderation requires advanced real-time analysis tools and poses new complexities for platform administrators and content supervisors.

This project aims to address these challenges by proposing innovative technological solutions that leverage multimodal large language models (MLLMs) to help platforms effectively moderate voice chat, ensuring safe and inclusive environments for all users, especially vulnerable younger audiences. [14]

Chapter 3

Motivation, Goal, and Challenges

Motivation

Voice chat presents unique challenges for moderation compared to text chat, primarily due to its ephemeral and dynamic nature. While conducting our research, we came across some major challenges that children and parents face on gaming platforms. These challenges motivated us to further investigate and develop an approach to address the problem. [17, 11]

Goal

The goal of this project is to rapidly monitor toxic voice chat on online gaming platforms in real time, thereby increasing the safety of children and curbing the spread of toxic behavior on these platforms. [7]

Challenges

Challenge 1. Voice chat moderation, unlike text moderation, cannot be quickly filtered based on words alone. While text can be easily logged and reviewed, voice chat is typically not recorded, making it harder to locate and remove problematic content. Additionally, processing audio data requires more computational resources than text, making real-time moderation more complex and expensive.

Challenge 2. Existing systems struggle to accurately distinguish toxic from non-toxic speech, resulting in inaccurate predictions. This occurs due to several factors, including bias in training data, the subjective nature of toxicity, and the difficulty in capturing nuanced context and intent. These models are also vulnerable to evasion tactics, such as the use of slang or speech variations to bypass filters.

Chapter 4

Benchmark Datasets

Detoxy Dataset

The DeToxy is the first publicly available toxicity-annotated dataset for the English language. It is released by TEGAnalytics, Cisco, IIT Delhi, and IIT Madras. DeToxy is sourced from various openly available speech databases and consists of over 2 million utterances. It is claimed by the authors of the dataset that this would act as a benchmark for the relatively new and unexplored Spoken Language Processing task of detecting toxicity from spoken utterances and boost further research in this space.

Ground Truth

Finally, strong unimodal baselines are provided for this dataset and compared with traditional two-step and E2E approaches. Text-based approaches are largely dependent on gold human-annotated transcripts for their performance and also suffer from keyword bias. However, the presence of speech files in the DeToxy dataset helps make the annotation unbiased and produces cleaner data.

MuTox Dataset

The MuTox dataset was released by FAIR and Meta. It is the first highly multilingual audio-based dataset with toxicity labels that covers 14 different linguistic families. The dataset comprises 20,000 audio utterances for English and Spanish and 4,000 for the other 28 languages.

Ground Truth

To demonstrate the quality of this dataset, the MuTox audio-based toxicity classifier is trained, allowing zero-shot toxicity detection in a wide range of languages. This classifier performs on par with existing text-based trainable classifiers, while expanding the language coverage more than tenfold. Compared to a wordlist-based classifier that covers a similar number of languages, MuTox improves the F1 score on average by 100%. This significant improvement underscores the potential of MuTox in advancing the field of audio-based toxicity detection.

Dataset Distribution

Since English is a widely spoken language on Roblox, we collected the audio from both datasets that were in English for the preliminary experiment and are for inference testing.

Dataset	**Class**	**Count**	**Class**	**Count**
MuTox	**Toxic**	257	**Non-Toxic**	1207
Detoxy	**Toxic**	500	**Non-Toxic**	500
Total	**Toxic**	757	**Non-Toxic**	1707

Table 4.1: Toxic and Non-Toxic counts across datasets

Chapter 5

Evaluation

Evaluation Metrics

Precision, Recall, and the F1 score are all metrics used to evaluate the performance of classification models. They measure different aspects of a model's ability to correctly classify data.

Precision: Precision measures the accuracy of positive predictions. It gives information on how many of the instances the model labeled as positive were positive.

$$Precision = \frac{TP}{TP + FP} \tag{5.1}$$

Recall: Recall measures the model's ability to find all the positive instances. It gives information on how many of the actual positive instances the model correctly identified.

$$Recall = \frac{TP}{TP + FN} \tag{5.2}$$

F1 Score: The F1 score is the harmonic mean of precision and recall. It provides a balanced measure of the model's performance, taking into account both false positives and false negatives.

$$F1\ Score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{5.3}$$

Chapter 6

Preliminary Results

Failure of Existing Detectors

We did a black-box testing of the collected audios on the Roblox In-Game Voice Chat Moderation System by creating an aggregated system at our end. Through the aggregated system, the audios were played live in a Roblox game, and the metrics were calculated accordingly. For the Open Source Toxicity Classifiers, we chose Roblox Voice-Safety Classifier to which we provided the audios and calculated the metrics. We tested toxicity on Text-based LLMs like Perspective API. We made a combination of OpenAI API (Whisper+Moderation), where the Whisper API converted audio into text, and the Moderation API classified the audio based on text. This process was similar to the one conducted by the authors of Detoxy in their paper while annotating the data. Next, we tested the audio dataset on AWS Transcribe, which is a commercial system. Finally, we conducted tests on commercial and open-source Multimodal LLMs like GPT-4o-Audio-Preview, GAMA, and Qwen2-Audio. [7]

Models	**Precision**	**Recall**	**F1-Score**
Roblox In-Game Voice Moderation	0.74	0.05	0.10
Roblox Voice-Safety Classifier	0.31	0.47	0.37
OpenAI (Whisper+Moderation)	0.31	0.41	0.35
AWS Transcribe	0.26	0.34	0.29
Perspective API	0.31	0.77	0.45
GAMA	0.00	0.00	0.00
Qwen2-Audio	0.89	0.42	0.45
GPT-4o-Audio-Preview	0.46	0.61	0.51

Table 6.1: Preliminary Results: Toxic/Non-Toxic Audio Classification

Chapter 7

Our Approach

On obtaining the preliminary results, we decided to construct our architecture using a Multimodal LLM with Prompt Engineering, as it requires less computation. In our case, we chose the GPT-4o model as its Precision, Recall, and F1-score were better than the other Multimodal LLMs. To conduct this experiment, we came up with an initial architecture for the prompt to continuously learn and evaluate. [1, 12, **?**]

Preliminary Approach

Few-Shot Learning

Few-shot prompting refers to the process of providing an AI model with a few examples of a task to guide its performance. This method is useful in scenarios where either extensive training data is unavailable or there aren't enough resources for computation. In techniques like zero-shot prompting, which requires no examples, or one-shot prompting, which relies on a single example, few-shot prompting uses multiple examples to improve performance. Few-shot learning is essential in situations for Generative AI where gathering large amounts of correctly labelled data is challenging. This technique also supports getting the predefined output format by guiding the model in reasoning through specific examples, ensuring consistency in the desired

structure. [9]

Chain of Thought + Few-Shot Learning

Chain of Thought (CoT) prompting is a technique for improving the reasoning capabilities of Multimodal LLMs by generating intermediate reasoning steps. This helps the MLLM generate more accurate answers. It can be combined with few-shot prompting to get better results on more complex tasks that require reasoning before responding, as it's a challenge with a zero-shot chain of thought. CoT has a lot of advantages. First of all, it's low-effort while being very effective, and works well with off-the-shelf MLLMs as a result, there is no need for finetuning. Interpretability with CoT prompting can also be achieved, as the Multimodal LLM's responses can be improved due to the reasoning steps, thereby increasing robustness and performance of the model in prediction. [4, 10, 2, 13, 8, 6, 16, 15]

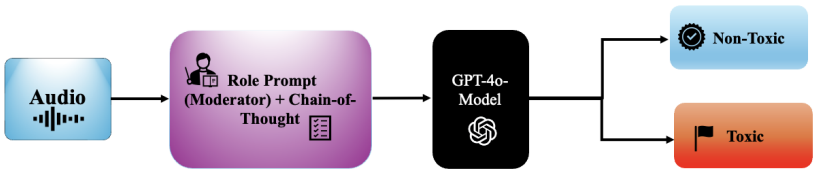

Figure 7.1: Chain-of-Thought + Few-Shot + Gpt-4o Approach

Architecture Overview

We have considered five steps while constructing the architecture.

Figure 7.2: Chain-of-Thought + Few-Shot + GPT-4o Architecture

Provide Definition and Examples

The definition of each type of toxic and non-toxic classes is provided in the prompt, along with the audio attached for each definition. Here, the text of the audio is written; however, while testing, we have attached two audio files for each sub-category.

Toxic Categories

Profanities: Slurs and language that is regarded as obscene, repulsive, or excessively vulgar, including scatological expressions.
Examples: shit, *asshole*, *fucking*

Hate Speech: Language used to demean, disparage, belittle, or insult groups of people.
Examples: women are sluts, *men are trash*, *wetbacks*, *towelheads*

Pornographic Language: Language that refers to sexual acts or describes body parts in a vulgar or explicit way.
Examples: suck my dick, *cumshot*

Violent or Bullying Language: Language used to bully, threaten, or silence individuals.
Examples: son of a bitch, *shut the fuck up*, *retard*

Appearance of Obscenity: Language that resembles profanities or pornographic content but is not directed at individuals.
Example: school sucks!

Non-Toxic Categories

Common Innocuous Slang: Informal or colloquial language that is not considered offensive.
Example: cops (referring to police officers)

Appearance of Hate: Language that expresses hate but is not directed at any person or group.
Example: I hate this movie!

Identification and Tone Measurement

Using the definition and the examples provided for each subcategory of toxic and non-toxic classes, the phrases were identified at first. The context of the phrase was extracted by the Multimodal LLM (GPT-4o-audio). It is then verified with the context of each of the subcategory class, if it matches any of the category, the audio's tone is measured. The tone measurement is done to validate the speaker's intent behind the conversation, as it results in high false positives and false negatives.

Conclusion

After Identification and Tone Measurement, the audios are labelled as toxic and non-toxic, along with subclasses of each category, with reasoning, making our architecture **Multilabel Toxic-Detection Multimodal LLM**.

Models	**Precision**	**Recall**	**F1-Score**
GPT-4o + COT + Few-Shot Prompt	**0.66**	**0.84**	**0.75**

Table 7.1: Performance of GPT-4o with Chain-of-Thought+Few-Shot Prompting

Future Work

After seeing a slight improvement in the performance, we are further going to work on improving the prompts and retest on the GPT-4o audio model. Since we got a better combination of Precision, Recall, and F1-Score for Qwen2 Multimodal as well, we will do the same experiments with it in parallel.

Chapter 8

Conclusions

From the observations, we conclude that our approach has significantly improved the evaluation metrics. However, false positives, reflected in the precision score, persist, indicating the need for a more comprehensive analysis of factors specific to toxic audio.

Bibliography

[1] Lee Boonstra. Prompt engineering, 2024.

[2] Mateo Clement. Optimizing llms for complex queries: The power of prompt engineering in few-shot learning.

[3] Marta R. Costa-jussà, Mariano Coria Meglioli, Pierre Andrews, David Dale, Prangthip Hansanti, Elahe Kalbassi, Alex Mourachko, Christophe Ropers, and Carleigh Wood. Mutox: Universal multilingual audio-based toxicity dataset and zero-shot detector. *arXiv preprint arXiv:2401.05060*, 2024. arXiv:2401.05060v2 [cs.SD], last revised 27 Jun 2024.

[4] Xian Fu. Enhancing multimodal large language models on demonstrative multi-image instructions. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, page 11429–11434, New York, NY, USA, 2024. Association for Computing Machinery.

[5] Sreyan Ghosh, Samden Lepcha, S Sakshi, Rajiv Ratn Shah, and S. Umesh. Detoxy: A large-scale multimodal dataset for toxicity classification in spoken utterances. *arXiv preprint arXiv:2110.07592*, 2022. Submitted to Interspeech 2022.

[6] Eunseo Jeong, Gyunyeop Kim, and Sangwoo Kang. Multimodal prompt learning in emotion recognition using context and audio information. *Mathematics*, 11(13):2908, 2023.

[7] Mahesh Kumar Nandwana, Yifan He, Joseph Liu, Xiao Yu, Charles Shang, Eloi Du Bois, Morgan McGuire, and Kiran Bhat. Voice toxicity detection using multi-task learning. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 331–335, 2024.

[8] Zhiqiu Lin, Samuel Yu, Zhiyi Kuang, Deepak Pathak, and Deva Ramanan. Multimodality helps unimodality: Cross-modal few-shot learning with multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19325–19337, 2023.

[9] Huan Ma, Changqing Zhang, Yatao Bian, Lemao Liu, Zhirui Zhang, Peilin Zhao, Shu Zhang, Huazhu Fu, Qinghua Hu, and Bingzhe Wu. Fairness-guided few-shot prompting for large language models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 43136–43155. Curran Associates, Inc., 2023.

[10] Anwesha Mohanty, Venkatesh Balavadhani Parthasarathy, and Arsalan Shahid. The future of mllm prompting is adaptive: A comprehensive experimental evaluation of prompt engineering methods for robust multimodal performance, 2025.

[11] Marcus Märtens, Siqi Shen, Alexandru Iosup, and Fernando Kuipers. Toxicity detection in multiplayer online games. In *2015 International Workshop on Network and Systems Support for Games (NetGames)*, pages 1–6, 2015.

[12] Ahlam Husni Abu Nada, Siddique Latif, and Junaid Qadir. Lightweight toxicity detection in spoken language: A transformer-based approach for edge devices, 2023.

[13] Arlo Octavia and Meade Cleti. Enhancing large language model performance through prompt engineering techniques.

[14] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022.

[15] Minjun Son and Sungjin Lee. Advancing multimodal large language models: Optimizing prompt engineering strategies for enhanced performance. *Applied Sciences*, 15(7):3992, 2025.

[16] Yaoxun Xu, Yixuan Zhou, Yunrui Cai, Jingran Xie, Runchuan Ye, and Zhiyong Wu. Multimodal emotion captioning using large language model with prompt engineering. In *Proceedings of the 2nd International Workshop on Multimodal and Responsible Affective Computing*, pages 104–109, 2024.

[17] Zinan Zhang, Sam Moradzadeh, Xinning Gui, and Yubo Kou. Harmful design in user-generated games and its ethical and governance challenges: An investigation of design co-ideation of game creators on roblox. *Proceedings of the ACM on Human-Computer Interaction*, 8(CHI PLAY):1–31, 2024.