Utrecht University

Department of Information and Computing Science

Game and Media Technology Master Thesis

Toxicity in AI-Generated Video Streams and Chats: How do Viewers React to Harmful Messages and Stereotypes?

First examiner:
Dr. Julian Frommel

Second examiner:
Dr. Almila Akdag

Candidate:
Joris Lambooij

Student Number:
5931126

February 17, 2025

Utrecht University

1 Abstract

Biases and harmful messages in AI-generated content are no new things. With the advent of generative AI models, these ingrained biases could become even more of a problem, as they do not rely on biased human input anymore. Identifying and measuring how toxic content and reactions to it appear and spread is thus vital for a healthy future of AI-generated content. Biases in AI-generated content have been studied extensively, but with the constant and rapid evolution of AI models, new forms of content are always emerging, with their own formats, styles, problems, and risks. This project aims to investigate one such form in particular: AI-generated video stream. In this project, a select number of AI-generated video streams and their communities have been analyzed, with the goal to answer questions surrounding toxicity: When and where does toxic content appear? How do viewer reactions manifest? Are there observable feedback loops between the generative AI and users? This project has analyzed transcripts of AI-generated video streams, alongside the chat logs that accompanied them. Both sets of data were then analyzed using the Perspective API to determine toxicity levels. Finally, a qualitative step of an inductive thematic analysis was performed to answer the underlying research questions of why this toxicity manifests, what it manifests in response to, and what form it takes when it does manifest.

Quantitative results were inconclusive, unable to provide strong evidence for a connection between content and comments. The thematic analysis did find interesting results, however, finding that characters belonging to marginalized communities had higher toxicity scores associated with them than their non-marginalized peers. Not all humor was explicit and offensive, however, as there was also plenty of toxicity purely born out of a frustration over technical difficulties. Furthermore, the thematic analysis found a stark difference between different AI stream communities and suggested that the source material for the AI streams has a big impact on the AI stream community as well, shaping the way that toxicity forms within said communities.

2 Introduction

2.1 Generative AI and toxicity

Artificial Intelligence (AI) has undeniably become a big aspect of contemporary culture. One of its most intriguing applications is content generation, where generative AI models can produce text, images, and since recently even videos that can mimic human creations. However, as AI-generated content has now entered the mainstream consciousness, concerns about its flaws and biases have become more prevalent as well. In particular, AI models have come under scrutiny for racist and sexist biases in their algorithms [1], [2]. Even when developers are aware of these problems and try to correct their model, that is only a band-aid fix, possibly resulting in comically inappropriate overcorrections [3].

Toxic messages, encompassing harmful stereotypes and even outright hate speech, pose significant risks to the health and well-being of online communities [4]. With AI's capacity to generate these kinds of messages in vast quantities and at inhuman speeds, it is imperative that we as a society stay vigilant about new developments and how those developments impact communities. In order to combat toxicity, it first has to be identified and located. For the purposes of this project, "toxicity" refers to harmful messages and stereotypes, direct personal attacks, and aggressive or offensive language. This includes sexism, racism, queerphobia, and ableism. This project aimed to do just that: To study AI-generated video content, particularly instances where it creates harmful messaging, and how that messaging affects the consumers of that content. The answers to these questions will hopefully guide future research and development of moderation techniques and tools.

2.2 Selected streams

This research project has examined AI-generated video streams in particular. These streams, while set up and managed by a human creator, rely entirely on automated, AI-powered processes to create content. In particular, language models like ChatGPT are used to create a script, which is then processed by scripted render engines like Unity. So, while there is still human involvement in the initial setup and moderation of the stream, all of the content ultimately is determined by the AI's input. Therefore, this project refers to them as "AI-generated video streams" as a shorthand.

The AI-generated video streams selected for this project all follow a similar format, in which they attempt to create a never-ending series of "episodes" from a popular TV series. The examples that this project will look at include these "infinite episodes" of *Family Guy* [5], *Spongebob Squarepants* [6].Initially, AI versions of *Breaking Bad* [7] and *Seinfeld* [8] were also considered, but ultimately not included for different reasons. The AI-generated *Breaking Bad* simply did not have the viewership of chat activity to draw any kind of substantial conclusions. The *Seinfeld* stream did have the viewership, but could only be found on the platform Twitch, while every other stream was found on YouTube. The difference in viewership and community culture on both platforms made a direct comparison more difficult, and because of time constraints, this stream was ultimately excluded as well. This left the project with two streams, and two communities to investigate. *Family Guy* and *Spongebob*. While both of these streams deal with different subject matters, different formats, and styles of the TV series that they attempt to parody, their internal workings are mostly identical.

2.3 How it works

Users can submit prompts via some form of delivery system, most often the Discord servers that the corresponding communities have set up. These prompts are then given to a text generator such as ChatGPT, which is asked to generate a short episode in the style of whatever show the stream is attempting to parody. That output is then processed through a series of AI voice models that mimic the voices of the characters in the show, some simple visuals are animated using various modeling tools such as Blender or Unity, and the complete product is then streamed to the web for viewers to watch. This entire process is mostly automated with minimal oversight, as admitted by the creators themselves in a short how-it-is-made video posted on a similar YouTube channel. [9].

This process means that these streams present a semi-interactive nature, where viewers can submit their own prompts which the AI then turns into a short episode. As a result, harmful or toxic messages can appear as either the result of ingrained AI biases or because of extrinsic viewer-submitted prompts. For this research project, both types have not been differentiated, though it is questionable if such a thing would even be possible. Models like ChatGPT operate as a sort of black box, where we humans have little to no insight into the inner workings. Whether or not toxic attitudes have been inserted into the original prompt is impossible to ascertain just by looking at the output. Furthermore, as the focus was on viewers' perceptions of and reactions to toxic messages regardless of their origin, this was not considered an important distinction to make.

2.4 Methods

To analyze these AI-generated streams, both automated and manual methods have been applied. First, a quantitative analysis using the Perspective API [10] identified toxicity within the stream itself and the corresponding comment sections. Using this data, the project plotted the toxicity levels over time and was able to identify sudden spikes and upticks in toxicity. Those peaks have then been further analyzed using both quantitative and qualitative methods.

The qualitative analysis took a closer look at the types of toxicity and their intent. The project attempted to identify the triggering incident, and then manually code each comment that was submitted within a minute of the trigger. The goal was to identify what caused toxicity in these AI-generated live streams, what kinds of content provoked it, how and why toxicity manifested, and to determine if there were observable feedback loops of toxic content provoking toxic comments, which in turn provoked more toxicity in the content again.

3 Related Work

3.1 Biases in AI

The prevalence of biases in AI algorithms is well-documented. From racist and sexist stereotypes being deeply ingrained in many generative models [1], [2], to outright discriminatory practices when relying on automated analyses instead of human judgments [11], AI models are still far from neutral.

One example is in image generation, where models such as Google's Gemini AI produce images based on user queries. These however often end up producing overwhelmingly white, male people when asked to generate any sort of person, regardless of the race- and gender-neutrality of the prompt. Google's Gemini AI was then overcorrected to promote diversity even in contexts where it was explicitly inappropriate [3]. This is just one example of the inherent racial and gender bias in large-scale AI models. Other image generators and other types of generative AI such as text generators all display similar biases [2].

The already ingrained biases in the original works that these streams take inspiration from mean that harmful messages could not only arise from the AI's own biases, but also from any existing biases and stereotypes that the original relied on, employed, or ridiculed. The latter option would not even have to have been used maliciously, as simply mentioning these biases and stereotypes could prompt the AI to utilize them, even in a manner that the original might never have intended.

3.2 Biases Against AI

Research has also been conducted into the human perception of AI-generated content. There is enough evidence to suggest that we humans perceive and treat AI-generated content differently than human-generated content. Whether or not content was created by a comparatively simple algorithm, deep learning, or generative models, humans consistently display a different attitude towards anything machine-made.

Dietvorst et al. [12] examine a phenomenon the authors call "AI Aversion". Human viewers, knowing that something was algorithm-created, consistently display lower trust towards the content, and perceive its quality to be worse than equivalent content they know was produced by humans. This mostly occurs after seeing the algorithm err, giving a false prediction or outcome. Similar studies into this phenomenon have come to similar results. Humans tend to favor things made by other humans, rating those works as significantly better (whatever "better" means in their respective context) as those made by algorithmic or AI means [13].

A follow-up study suggested that this phenomenon may perhaps be a little more complicated than initially observed [14]. That study theorized that the phenomenon may not necessarily have been a bias against algorithms and AI, but rather a preference for things created by our fellow humans, i.e. "Human Favoritism". If participants were given even a modicum of control over the results, the previously observed aversion diminished significantly.

In the context of AI-generated video streams, these two phenomena, AI Aversion and Human Favoritism, could be interesting to observe as well. Do viewers react to the content differently because it is "only a computer"? Or does the semi-interactive nature of these streams, allowing some human input, serve as enough of a sense of control that the content is taken as seriously as human-generated content would?

3.3 Toxicity in Online Communities

Research has been conducted on the impact of toxicity within online communities. Small amounts of toxicity can be fine, sometimes even beneficial [4], as these can be classified as friendly banter more so than hateful conduct. However, anything exceeding that is more likely to be detrimental to the long-term health of communities, often contributing significantly to their quickened deterioration. It has also been shown that a failure to address toxicity can lead to a positive feedback loop, where these messages get increasingly normalized [15].

Toxic content, or more specifically controversial content, is also a popular way to increase one's reach on social media. It is not a new phenomenon, as sparking controversy and moral outrage has always provided a way to garner attention. In the digital age, this has become an even more prevalent tactic to increase one's reach [16]. Toxicity not only increases the reach of a message to reach more people but can also help to keep those people engaged for longer. People are much more likely to react to and interact with controversial or toxic content [17].

Live stream chats are not that different, often acting as community and community spaces in their own right. Hamilton et al. [18] even go so far as to describe live streams can act as virtual third places, where people can socialize and bond. And so the identification and moderation of toxicity there is just as important. The size and format of a live stream and its community also have a big impact on the prevalence of toxicity, as do the gender of the creator and gender distribution of the audience [19]. There has been little research into toxicity in the communities around AI-generated live streams, and how these kinds of communities and live streams compare to "traditional" live streams. This project will thus try to provide a starting point for that comparison, to investigate how toxicity manifests in these streams.

3.4 Toxicity Triggers

Almerekhi et al. [20] postulate that in online discussions toxicity is often a response to certain triggers. These are often subjects of ongoing political debates, as words like *woman*, *Muslim*, or *Israel* are examples of words that can trigger a toxic response. As discussed before, TV series are no stranger to using stereotypes about marginalized people. Whether or not those stereotypes themselves are harmful may not even matter, as the mere mention can spark toxic reactions.

When AIGVC generates biased content, whether prompted or not, it is likely to invite toxic reactions similarly. Searching for these triggers whenever toxicity flares up can provide a more detailed explanation of the causes of said toxicity.

3.5 Stereotypes in Media

Harmful stereotypes in media are no new phenomenon, especially in comedy series. Series like *Family Guy* and *South Park* have been critiqued for their reliance on negative stereotypes for jokes, particularly their portrayals of Jewish people [21]. These series are also no stranger to sexist stereotypes about women [22].

Family Guy and *South Park* are by no means the first or only offenders, of course. Nor will they be the last. However, both series are widely popular and have a reputation for being "provocative", never shying away from making crude or controversial jokes. Fundamentally though, both shows do feature pro-social messages, even if veiled in satire and irony. Arguments can even be made that this format is actually beneficial, reaching audiences that would otherwise not engage with such topics [23].

But, the question remains: What will a generative AI model do with these source materials? Research into this question is sparse, as this is a new development. A study by Gross [24] investigating the use of generative AI in the creation of satirical news has pointed out major flaws and dangers, that while the generated news articles did look like satire, they distinctly lack the human intent to critique a problem in society. That is arguably the most important aspect of satire though, even the entire purpose, and without it, it is not proper satire.

Being Large Language Models, ChatGPT and its equivalents are unable to engage with text and subtext critically. Instead, they merely attempt to imitate a style of content, predicting what it would look like if someone (a human) had written it. Hamid [25] likens it to the Chinese Room thought experiment by Searle [26], [27], arguing that intelligent-sounding speech is not enough to consider the model intelligent.

It is important to note that Searle's thought experiment of the Chinese Room has been disputed and countered, with opponents saying it is too simplistic and not applicable to questions of machine consciousness [28], some even going so far as to say that the very premise of the thought experiment is unhelpful and counterproductive [29]. Although this project does not attempt to answer that particular question, it remains an important point of consideration when discussing nuanced topics such as satire, and the necessary *intent* required to execute them successfully.

And so, with all of this in mind, this project attempts to answer several questions: Are the text generator models able to capture the nuance required to create proper, successful satire? Or will the generated parodies contain offensive jokes as a poor imitation of the original? How will the audience perceive these questions, and how will they react? Will the reactions be dismissive, attributing any and all wrongdoing to "it is just a flawed AI"? Or will they be disapproving, demanding that the content be better moderated?

3.6 Summary

In summary, AI-generated live streams can form communities, just like other live streams often do. In all online communities, toxicity is inevitable, especially once they grow to a certain size. The questions thus are: What form does this toxicity take in the communities of AI-generated content? How does the source material affect the toxicity in either the streams themselves or the communities around them?

In AI-generated live streams specifically, toxicity can not only arise from within the community but also from the content of the stream itself. This is not necessarily exclusive to AI-generated live streams, of course. Human streamers can be toxic too, after all. However, research around AI-generated content and consumers' reactions to it shows that humans treat the two differently and that there is an important distinction to be made there. And so this project aims to bridge that gap: To analyze toxicity within the communities around AI-generated live streams, and to identify the ways in which toxicity manifests itself within them as a reaction to the content.

4 Research Goals

The primary goal of this research project is to analyze AI-generated video streams, evaluating the frequency and intensity of any kind of toxicity that appears either on screen or in the live chat that is continuously reacting to the content on screen. This project is interested in quantifying the relationship and identifying correlations between the two.

The secondary goal of this project is to identify how this toxicity manifests. Toxicity can come in many forms, and not all are created equal. This then poses the question, of how toxicity manifests, what types of toxicity manifest, and how often and why.

As such, the research questions this project aims to answer are as follows:

1. How does toxicity develop over the course of AI-generated live streams?
   (a) Does biased/harmful content appearing on screen invite toxicity in the live chat?
   (b) Does toxicity in the live chat lead to more biased/harmful prompts being submitted to the AI?
   (c) Do these two phenomena lead to a feedback loop of toxicity and harmful content?
2. What form does toxicity take in AI-generated live streams?
   (a) Which topics discussed within the live chat trigger more toxic reactions?
   (b) Which types of toxicity appear more often than others?
3. How does the source material affect the toxicity in the AI versions?

5 Methodology

This project aimed to systematically analyze toxicity levels. To gather the data, the transcripts of the YouTube video streams were needed. This was done by scraping the video stream's subtitles, which were automatically generated by YouTube's closed captions system. Concurrently, the chat logs also had to be scraped to capture the live reactions to the content on screen. Both comments and the stream's content were time-stamped when recorded this way, which would later allow for easy synchronization of the two data sets.

Once a sufficient amount of data had been collected, both automated and manual methods of analysis were used, as outlined in the following sections.

5.1 Data Collection

The data collection for this project was done over the course of May 2024. The availability of YouTube videos and all associated aspects can fluctuate over time. The AI-generated imitations of existing TV shows suffer from this especially, as their legality is still disputed. This project vouches for the accuracy of all claims related to data availability for only May of 2024, with any preceding or following changes not taken into account. As of writing this report in February 2025, it also appears that the *AI Sponge Rehydrated* YouTube channel has already been taken down over claims of copyright infringement. *ai_peter* remains operational, at least at the time of writing.

The data for this project was gathered directly from the streams' respective YouTube channels. Each channel keeps a library of the streams available for later playback. These Videos-On-Demand, also referred to as *VODs*, each contain the video and audio of the stream itself, as well as a record of the live chat reactions including timestamps. YouTube videos also usually offer automatically generated subtitles if none were provided by the author, though they are not always available. For this project, it was the archived live chat and the subtitles that were important, so any stream with sufficient chat activity and available subtitles was chosen for each of the chosen YouTube channels.

Third Party Tools YouTube's UI does not provide an official way to extract either the subtitles or the comments of any given VOD. So, third-party tools were used to download this data. To acquire the subtitles, the website *downsub.com* provided a way to download a YouTube video's subtitles in SRT format. SRT is a file format for subtitles which includes timestamps for each line of subtitles, exactly what this project called for.

The corpus for the comments was acquired using a browser extension, *YCS - YouTube Comment Search* [30]. This extension handled the downloading of all comments for each stream's VOD, alongside the timestamp and the name of the commenter.

Selection Criteria The streams from which this project aimed to capture the data were the AI parodies of the popular TV series *Family Guy* and *Spongebob Squarepants*. The streams could all be found on YouTube, on their respective channels of *ai_peter*[5] and *AI Sponge Rehydrated*[6]. These two streams in particular had been selected because they represented a contrast of styles and audiences in their original form. *Family Guy* is a show aimed at teenagers and young adults, whereas *Spongebob Squarepants* is a show aimed at children but enjoyable for all ages. The two shows are therefore quite different in style, which in turn could mean that both AI-generated streams would also be different.

In order to qualify for this project, a stream should be of sufficient length, provide auto-generated subtitles, and have decently high chat activity. For the length, it was decided that a stream needed to run for at least one hour so that there could be observable developments in the live chat throughout the stream. The threshold for sufficiently high chat activity was set at 500 comments per hour. If a stream met this criterium, it meant that on average there would be at least one comment every 7.2 seconds. This kind of threshold was necessary, as streams with low chat engagement would yield data too sparse to reliably analyze. A higher threshold would have been desirable, but with the availability of streams that have enough data and auto-generated captions, this was the necessary compromise.

Synchronization Since both datasets were already timestamped, synchronizing the two was relatively straightforward. This allowed a detailed analysis of toxicity levels in both sets, as well as any correlation that may or may not have occurred between or across the two. Within the selected datasets, sections where toxicity appears - or where levels of existing toxicity suddenly spike up - were then selected for more detailed analysis, as detailed in section 5.3.2.

5.2 Collected Streams

Family Guy By far the biggest stream, *Family Guy* had the largest availability of streams and data. For this channel, the six longest streams with available subtitles were chosen, totaling a combined runtime of ca. 32 hours. These streams all originally went up in 2023.

Spongebob The AI-generated version of *Spongebob* had a much smaller community. As a result, finding VODs that met the selection criteria set above was more challenging. For this part of the project, only three streams qualified, totaling a combined runtime of ca. 22 hours. These streams all took place in 2024.

Collected streams Streams from the following dates were selected and collected as data:

Date	Length (h)	Comments
Jun 15, 2023	11	18,190
Jun 17, 2023	12	28,573
Aug 8, 2023	5	9,469
Aug 21, 2023	1	1,161
Sep 26, 2023	2	1,066

Table 1: Family Guy streams

Date	Length (h)	Comments
Mar 26, 2024	11	33,838
Apr 6, 2024	1	3,180
May 5, 2024	10	14,916

Table 2: Spongebob streams

5.3 Analysis

This project aimed for a hybrid approach, combining automated and manual data analysis methods to ensure both an objective, quantitative measurement, as well as a qualitative, human judgment.

For the purposes of this project, *content* or *stream* refers to the spoken words that appear in the AI-generated live stream, the lines that each character of the parodied show is saying. *Lines* refers to the individual spoken lines, the full or partial sentences that make up a character's speech. On the other end, *chat* refers to the live chat reactions to said live stream, consisting of *comments* posted by *commenters* or *chatters*.

5.3.1 Quantitative

Different frameworks exist to quantify toxicity in an automated manner. This project incorporated the Perspective API [10] to automate the analysis and objectively identify and measure the toxicity of the data in its entirety.

Perspective is a free API, created to combat toxicity online. It is intended to help with the moderation of online spaces, providing an automated tool to more easily identify toxicity. The dataset was rated using this API, to create a quantitative analysis of toxicity in the live chat, to plot that toxicity over time, and to identify key sections to inspect in further detail in the qualitative analysis.

Comments: The dataset containing the live chat comments was sorted chronologically, in the order in which the chat comments appeared in the live chat. Each comment was then evaluated using the Perspective API, which graded it on a scale from 0 to 1, assessing the overall level of toxicity. The API also offers more fine-grained scales of different types of toxicity, but for this project, only the overall generalized assessment was utilized.

Those toxicity scores were plotted over time. It is important to note here that chat comments were not uniformly distributed over time, as sometimes there happened to be a drought of comments over an extended period, while other times a flood of comments reacted to something unexpected. This natural variation in density was not necessarily detrimental though, as an increase in the overall number of comments would both signal that something significant has happened and increase the overall reliability of that section of the data by providing a bigger sample set.

Content: The same has been done to a transcript of the stream's content. Using YouTube's automatic subtitle generation, the content of the stream was captured into plain text files, which were then analyzed by the same Perspective API in much the same manner.

This dataset was more consistent, with a predictable amount of words per minute, so the temporal fluctuations of the chat were not a consideration here. Furthermore, chat comments have the possibility of masking their true intended toxicity using creative spelling, euphemisms, or other evasive methods. The AI-generated video content on the other hand does not, as the technical setup that these AI-generated streams use was not designed or instructed to use such techniques.

However, this dataset still presented a different challenge: YouTube's automated captions system captures only the words spoken, with little regard for punctuation, context, or coherency. To still preserve accuracy and context, and to hopefully give the Perspective API better data to work with, these individual lines needed to be preprocessed.

To solve this issue, a sliding window approach was used, where each line was bundled with both its predecessor and its successor into one single data point for the Perspective API to evaluate. Each of these combined lines was then fed to the API for grading, using the same parameters as the comments. The resulting toxicity scores were then once again paired with the corresponding timestamps to plot the toxicity levels throughout the stream.

Interpolation Once all lines and comments had been graded by the Perspective API, the next step was to interpolate the scores to provide a continuous function approximating the overall trends. At the same time, outliers had to be weeded out to improve the reliability of the next step, which was the question of peak detection. To achieve this, the data for chat was divided into non-overlapping sections of 10 seconds each. In each section, the mean of the toxicity scores of all lines and comments that occurred within that period was calculated.

The same was also done for the stream. While this dataset did have an inherently more consistent frequency of data points, it still was not quite uniform. To maintain a consistent approach, the toxicity scores of the individual lines of the stream were also interpolated into the same 10-second windows, utilizing the timestamps of the first line of each line triplet.

After the interpolation step came the identification of peaks in the toxicity scores. This was done using a moving window over the last 60 seconds. If the current section of 10 seconds was considerably higher in toxicity than the preceding 50 seconds, a peak was registered. A difference was counted as a peak if it was at least 2.5 times larger than the standard deviation observed in the preceding 50 seconds of toxicity. Once again, this step was performed for both the stream and chat data, using the same method and parameters for both.

Once individual peaks had been identified, the preceding 10 and the following 60 seconds were then selected as that peak's surrounding section. The timestamps of these sections were then used to compare toxicity levels, both within and between streams and comments.

Within-Comparison With the peaks identified, both datasets could now be split again into two parts: Peaks and non-peaks. Now those two sets were compared against each other to determine if there was a significant difference between them. While this may seem redundant, given how these peaks were specifically selected when they were deemed to be significantly higher than the preceding average, this step nevertheless served two purposes: It would confirm whether or not the detected peaks were complete outliers or indicators of heightened toxicity levels even after the initial peak. Additionally, it would also provide a point of reference for the next step, the between-comparison. Kruskal-Wallis tests were used to determine if the two sets were significantly different in their toxicity levels. This test attempted to answer the question of whether the periods following a spike in toxicity would also demonstrate heightened levels of toxicity, or whether these spikes themselves were statistical outliers, not influencing the course of the toxicity levels in any major way.

Between-Comparison Importantly, the test described previously was also done by cross-referencing data across the two datasets of stream and chat. Concretely, this meant taking the timestamps of peaks in the stream but referencing the toxicity levels within the chat. By cross-referencing the datasets in this manner, the question was investigated of whether a rise in toxicity in the stream would cause a rise in toxicity in the comments.

These cross-references used the same Kruskal-Wallis tests, though the squared means of toxicity levels within and across these peaks were also calculated for a more detailed analysis of the nature of these potential influences.

Evaluation: With toxicity scores assigned, processed, and compared, the first and third research questions could now be investigated: The Kruskal-Wallis tests comparing peak and non-peak sections across the two datasets would provide insight into the influence that the stream had over the chat, and in what capacity toxicity in the former would invite toxicity in the latter.

The mean toxicity of each stream and accompanying chat also provided some insights into this potential relationship. It furthermore allowed a comparison between the two communities of AI-generated *Family Guy* and AI-generated *Spongebob Squarepants*.

5.3.2 Qualitative

To answer the second research question and its subquestions, a qualitative approach was also necessary. Sections of the chat were selected for closer inspection, to see what commenters were talking about, and how they were talking about it.

Sifting through all the comments manually would have taken a long time, but with the work done in the previous step, the dataset could be reduced to only the most interesting sections. This limitation allowed the next step of looking at these sections in more detail, using qualitative, human judgment.

Additionally, automated processes run the risk of missing toxic comments that intentionally attempt to circumvent moderation algorithms: Substituting letters with numbers, symbols, or emotes, or the use of emotes in the first place, all are used to evade most traditional keyword-based searches. Another aspect that the automated process could have missed was censored or veiled language, that either the commenter, the stream's moderation team, or YouTube's automated systems had censored. For these reasons, the qualitative analysis looked at select parts of the dataset with a human eye.

Coding The analysis made use of a combination of inductive coding. Every comment within relevant sections identified in the previous step (10 seconds before to 60 seconds after) was evaluated and inductively coded to synthesize the expressed intent in a formal way, along with the subject of discussion.

The results of this coding process were then examined, to distill the overall attitude that commenters displayed in reaction to these topics. Furthermore, this also allowed for a more accurate analysis of certain key topics.

The guidelines for applying codes to each comment were as follows. Most codes were not mutually exclusive, and every comment received as many codes as were applicable.

- First of all, if a character from the show was mentioned by name, or if it could be inferred through context clues that the comment was talking about them, that character's name was assigned as a code to that comment.
- Comments talking about real-world religions, religious figures, or iconography, were assigned one of the following labels, whichever would apply: "Christianity", "Judaism" or "Islam".
- Many comments used sexually explicit language. This could range from simple innuendo to explicit descriptions of sexual acts. These comments were given the label "Sexual".
- Similarly, many comments invoked language about human waste. When this was used in a joking manner, the label "Toilet Humor" applied.
- Death was something that was frequently mentioned, most often about a character from the show, either commenting on their on-screen "death", wishing they would die, or suggesting a prompt that involved their death. These were assigned the label "Death".

- Many comments discussed AI, either about the specific stream they were watching or in more general terms referring to the broader field. This also includes concerns about AI ethics and safety. These were assigned the label "AI".
- Certain topics were categorized and summarized under one umbrella term, which was then assigned as a label to all comments that touched on the subject. Examples include "Technical Problems" for anything to do with the stream not working properly.
- Similarly, anytime that minority groups were being mentioned or discussed, an appropriate label was used. Examples include "POC" for any mention of people of color. There was an effort made to distinguish what was merely mentioning and what was discriminatory, though the line between those two categories was quite blurry most of the time.
- Lastly, there are too many subjects and labels that emerged during the analysis, which would be too numerous to list them all here. Some stream-specific labels will be explained in later sections, but most other topics should be self-explanatory enough.

Combination with the quantitative analysis Once all relevant comments had been processed, it was then also possible to combine both the qualitative and quantitative steps to calculate the aggregate toxicity scores of these labels. By comparing the toxicity levels of different labels, certain inferences and conclusions could then be drawn about the attitudes of the chat.

Evaluation After the comments had been coded, an evaluation took place to answer the research questions: Through the coding process, discussion topics and the general emotional atmosphere of the comments and the content had been distilled. This could then be used to identify categories of toxicity, and which ones appear more often than others. By examining the development over time in the comments, the relationship between content and comments could be evaluated. The combination of these quantitative analysis and inductive coding then highlighted which topics were triggering toxicity more than others, and which topics would invite disproportionate levels of toxicity.

5.3.3 Topic Modelling

Lastly, to combine the results from both the quantitative and qualitative analyses, topic modeling was used to get a clear picture of what people were talking about and how. The peak sections identified during the quantitative analysis were grouped and fed into a topic modeling algorithm. For this project, the Python package Gensim [31] was used. Gensim provides functions to apply the Word2Vec algorithm to create a Machine Learning model that groups keywords into different, depending on how often these words are used close together. A range of options for the number of topics were tested, and for each resulting model, a coherence score was calculated to assess the model's accuracy. Then, the model with the highest coherence score was selected to present in this paper.

For this project, the topics discussed in the comments of the streams were of interest. This was done with up to 1000 comments at a time, or up until the start or end of a peak section, whichever condition would be fulfilled sooner. These aggregates of up to 1000 comments served as the "documents" that Gensim's Word2vec function uses to extract the topics from. The 1000-comment limitation was introduced mostly for compatibility with the *Spongebob* streams, as these would have otherwise had too few documents to work with. Not splitting the comments this way resulted in overfitted models with coherence scores close to 1, which did not provide much additional insights beyond what had already been observed manually.

The next step was the preprocessing and cleaning up of these strings. Following Gensim's design guidelines all punctuation was removed, as well as a list of stopwords. Stopwords are words that do not carry any real meaning in and of themselves but rather work to enhance, modify, or relate the meanings that the nouns, verbs, and adjectives around them describe. This list of stopwords was taken from the Natural Language Toolkit (NLTK) Python package [32], but further expanded upon with a selection of stream/chat-specific stopwords that were observed during the development of this project. Additions include "oh", "uh", "lol" and other filler words, as well as a selection of stream-specific phrases that were spammed so much that including them would be pointless.

After preprocessing, Gensim's Word2vec function could be applied. All streams from one community were taken at the same time, creating one overarching topic model. The parameters for this function were set at a vector size of 500. The minimum word frequency was 5, and the skip-gram version for the architecture of the NN was used. Otherwise, default parameters were kept, including a window size of 5 and a learning rate of 0.025 Once the algorithm was finished, each word now had a ranking in each of the resulting topics. For each topic, the 10 highest-ranking words for that topic were selected for presentation.

6 Results - *Family Guy*

6.1 Quantitative Analysis

Date	15-06	17-06	20-08	21-08	26-09
Length (h)	11	12	5	1	2
Comments	18,190	28,573	9,469	1,161	1,066
Average Stream Toxicity %	4.2	2.8	14.6	12.3	11.4
Average Chat Toxicity %	3.8	2.8	5.0	5.2	3.3
Median Stream Toxicity %	1.3	0.0	16.7	10.7	6.8
Median Chat Toxicity %	0.8	0.0	2.4	0.8	0.1

Table 3: Breakdown per individual stream. Toxicity scores have been mapped to a percentage value for better readability.

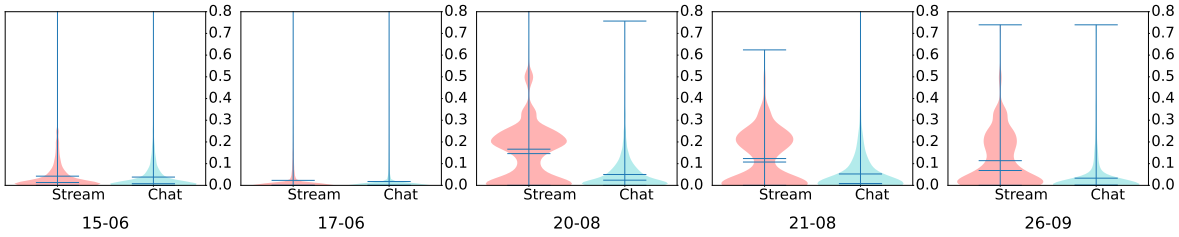


Figure 1: Toxicity score distributions for each stream.

Toxicity overall As seen in table 3, in all cases the stream exhibited higher toxicity levels than the comments, at least when evaluated in aggregate. This is likely due to the many spam comments and comments that do not really say anything, simple reactions that do not convey anything other than surprise. The median toxicity levels support this observation, as the difference seems even more pronounced there. The stream from June 17th is especially notable in this case, as it apparently featured so many stream lines and chat comments with a toxicity of 0 that it resulted in a median of 0. This observation is further supported by Fig. 1, which shows the distribution of toxicity scores for each stream. Densities for all streams are skewed towards the lower end, but the shorter streams have a visibly more spread-out distribution than the longer streams.

Cross-Examination Table 4 provides a more detailed breakdown of the average toxicity levels in the peak sections. More specifically, it analyzes the toxicity levels during the time periods defined by peaks in toxicity, from 10 seconds before to 50 seconds after. Peaks in toxicity levels were recorded both within and across datasets, meaning each peak in either stream or chat defined two separate sections, one within its own dataset and one in the other. Both of these resulting sections thus covered the same period as defined by the timestamp of the peak, one in the stream and one in the corresponding chat.

Date	15-06	17-06	20-08	21-08	26-09
Δ Stream to Stream Tox %	+1.1	+1.2	+2.8	+2.4	+2.8
Δ Chat to Chat Tox %	+0.8	+1.1	+1.1	+1.4	+1.1
Δ Stream to Chat Tox %	+0.1	+0.0	-0.3	+0.8	+0.1
Δ Chat to Stream Tox %	-0.1	-0.1	+0.1	-2.0	-0.4

Table 4: Differences in toxicity levels between peak and non-peak sections. The first two rows compare peaks and non-peaks of each dataset individually. The last two rows cross-reference the peaks from one dataset, taking the time periods of peaks in that set but examining toxicity scores found in the other set at the same time. See Fig. 2 for reference. Toxicity scores have been mapped to a percentage value for better readability.

Figure 2: Visual illustrations of table 4. The figures represent toxicity levels in the first 10 minutes of the stream/chat from September 26th. The red graphs illustrate stream toxicity, the blue graphs illustrate chat toxicity levels. Sections marked in red denote peaks identified in the stream, while blue sections are peaks identified in the chat. So, the combination stream-to-chat examines the toxicity levels in the chat, but using the sections dictated by peaks in the stream. Chat-to-stream does the reverse.

Figure 3: Stream to Stream

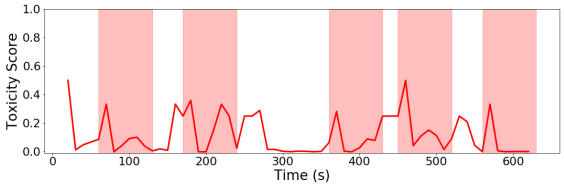


Figure 4: Chat to Chat

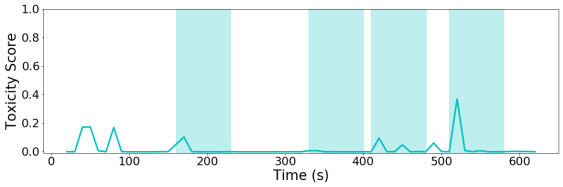


Figure 5: Stream to Chat

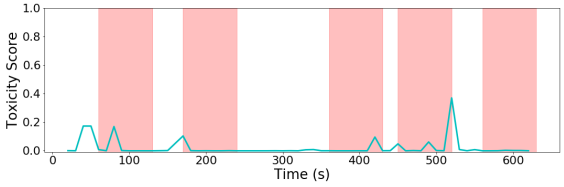
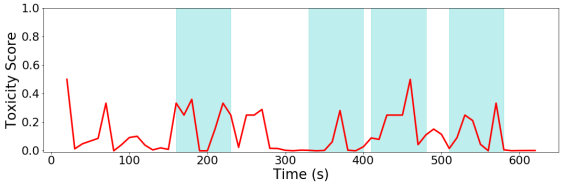


Figure 6: Chat to Stream



Stream-to-stream therefore refers to all the sections defined by timestamps in the stream, and analyzing toxicity levels in the stream as well. *Stream-to-chat* on the other hand takes the timestamps from the stream peaks and analyzes data from the chat, thereby describing the reactions in the chat that responded to whatever was happening in the stream at that time. All Δ values are normalized to the average toxicity within the target dataset, meaning they describe how much higher or lower the toxicity levels are than the average toxicity in the rest of that stream or chat.

Once a single spike in toxicity occurred, the overall toxicity levels quickly returned to their regular values. Table 4 illustrates this with the stream-to-stream and chat-to-chat rows, showing relatively low differences in toxicity between peak and non-peaks within each dataset. These rows show the difference in toxicity levels between the peaks and non-peaks of the respective datasets, broken down per stream. A positive value means that, on average, these peak sections contained more toxic lines or comments than the non-peak sections. Kruskal-Wallis tests were performed to determine if these differences were significant, but all of these tests yielded p-values of 1.0, failing to find significant differences.

The stream-to-chat cross-evaluation took the timestamps of the peaks within the stream and applied those timestamps to the chat to take a snapshot of the reactions to that stream peak. When comparing these peak-reaction sections to the rest of the chat, toxicity levels were heightened.

The *within* toxicity Δs of stream-to-stream and chat-to-chat, predictably feature only positive values. This indicates that toxicity levels remain somewhat elevated after a peak, at least for a little while. However, all values are rather low, suggesting that either the methodology is flawed and unable to properly describe the developments of toxicity levels, or the developments themselves are simply small and not as dramatic as expected.

Similarly, there do not seem to be many insights to be drawn from the *across* Δs. The stream-to-chat values are considerably lower than the stream-to-stream values, indicating that there is no strong connection from the stream to the chat in terms of how toxic each development gets at any given time. The chat-to-stream values are even going into the negatives. Considering that the average toxicity levels in the chats seem to be generally lower than the toxicity levels in the streams as seen in 3, this seems to line up with that observation even when specifically looking at the peak sections.

Date	15-06	17-06	20-08	21-08	26-09
Sample size N	3965	4319	1522	413	732
Kruskal-Wallis H statistic	1.02	0.11	0.29	3.95	0.01
p-value	0.69	0.26	0.41	0.95	0.09

Table 5: Results of the Kruskal-Wallis tests determining if peak and non-peaks are different.

Kruskal-Wallis test To confirm whether or not the peak and non-peak sections were significantly different, statistical tests were also used. Since the peak and non-peak sections resulted in differently sized datasets, where normality could not be assumed, Kruskal-Wallis tests were chosen. These tests focused specifically on the stream-to-chat data, to investigate the influence that toxicity in the stream had over toxicity in the chat. For each stream, the chat was split into peak sections and non-peak sections as defined by the timestamps of peaks within the corresponding stream. As each test was a direct comparison between only two sets of data, the degrees of freedom were $df = 1$ for each test.

Table 5 details the results of these tests for each stream. Kruskal-Wallis tests could not find a significant difference in toxicity scores between the peak and non-peak sections of the chat when it was divided along the timestamps dictated by the stream peaks. The sample size for each stream is the number of 10-second intervals that were interpolated, as outlined in 5.3.1. The p-values for all five streams were greater than a significance level $\alpha = 0.05$, so no insights could be extracted from this test.

6.2 Thematic Analysis

The thematic analysis yielded several key insights into the nature of the toxicity found in these live streams. In the *Family Guy* stream, the following patterns were observable:

Technical Difficulties: Whenever the stream encountered technical problems, toxic reactions were quick to follow. Technical problems could appear in the form of the entire stream shutting down unexpectedly, characters' voices suddenly peaking in volume, or character models phasing through scene geometry and disappearing from view.

Generally speaking, toxicity that was a reaction to these kinds of issues was the mildest and least offensive form of toxicity, expressed more in general frustration than targeted, personalized attacks on any one person or group of persons, whether real or fictional.

Women Characters: The same cannot be said for the toxicity that was aimed at women characters specifically. *Family Guy* and its AI-generated livestream counterpart both feature a number of female characters, most notably the characters of Meg, Lois, and Bonnie. In both the original and the AI-generated version, Meg especially is often the butt of a joke, being ridiculed, harassed, and mistreated for the viewers' amusement. This targeted mistreatment is also mirrored in the live chat, both in reaction to the stream and to other live comments. Examples of this include topic suggestions involving Meg's suffering, also disproportionally involving her dying in some way. No other character received this kind of treatment. This phenomenon is a mirror version of the original show and the way Meg is treated there.

Meg is not the only female character, however. As mentioned earlier, the character of Bonnie is also featured. She does not appear to be hated as much as Meg is, instead, the toxicity surrounding her takes the form of sexualizing and objectifying her. Yet again, this mirrors the way in which *Family Guy* depicts and treats the character, as one of Bonnie's defining traits in the *Family Guy* universe is her attractiveness.

Offensive Humor Alongside the sexism that is on display whenever female characters are being discussed, other minority groups frequently also find themselves the target of toxicity from the live chat. Homophobia and transphobia are not uncommon, and neither are negative attitudes towards ethnic and religious minorities. Often, these sentiments were expressed in the form of jokes. This is a common strategy for users online to mask their true intentions, hiding in the ambiguous nature and plausible deniability of humor. The majority of the comments that were identified as people joking fell into one of two categories: They were either sexual comments, most often directed at the female characters in the stream. Alternatively, these comments in some way referenced Christianity, more specifically invoking the name of Jesus Christ in their comment. Most other Joke-type comments did reference minority groups in one way or another. No single category of these seemed more dominant than others, indicating that the commenters posting these comments were indiscriminate with the targets of their jokes.

In many cases, the comments posted seemed more interested in humor for the sake of humor, with no deeper intent or thought put behind the words. No matter the subject, nothing and nobody was safe from being the target of the joke.

Community In-Jokes The final overarching pattern that emerged was community in-jokes. This category was still a noticeable category of toxicity because these in-jokes were among other reasons perpetuated because of the annoyed reactions of other chatters who felt the joke had run its course long ago. Each time such a joke began to take over the conversation, the chat was flooded by two sides of the same coin: People spamming the joke over and over, and people reacting with annoyance and frustration to this spam, which only encouraged the aforementioned spammers to continue. These jokes turned into ironic versions of themselves, quickly going through the usual cycle that internet memes go through. While this category of toxicity was not the worst in terms of problematic language, that did not stop individuals from hurling insults and other toxic language at each other. Generally speaking though, nothing in this category exceeded the normal expectations one would have for any internet community.

6.3 Toxic Subjects

Subject	Count	Average Toxicity
AIDS	7	0.54
Alt-Right	10	0.47
Toilet Humor	19	0.42
Conspiracy	17	0.42
Islam	6	0.42
Transgender	11	0.39
Death	84	0.38
Politics	11	0.38
Sexual	168	0.37
Homosexuality	47	0.37

Character	Count	Average Toxicity
Meg	110	0.29
Mort	13	0.25
Chris	73	0.21
Bonnie	46	0.18
Brian	88	0.15
Peter	217	0.14
Quagmire	60	0.13
Joe	43	0.13
Herbert	44	0.13
Lois	11	0.12

Table 6: Notable Toxicity Scores of Key Topics. The left table contains the 10 most toxic topics that were observed at least 5 times, ordered from most to least toxic. The right table does the same for all the characters.

The most toxic subjects Table 6 breaks down a selection of key subjects and the average toxicity score associated with those subjects. As laid out previously, discussions involving minority groups or their struggles tend to exhibit higher toxicity scores. Despite the relatively small sample sizes, a consistent pattern is visible, mirroring trends in other communities and the wider landscape of political discourse. More marginalized communities attract more polarized discussions, thereby inviting more toxic behavior.

Subjects with high toxicity were, for the most part, as one might expect: Anything political, religious, or culturally divisive. "AIDS" took the top spot in terms of toxicity, but appeared only 7 times in total, which is comparatively low. Far more consistently toxic was the subject of homosexuality. This came most often in the form of using homosexuality as an insult, calling things or people "gay" as a means to express one's dislike for it, though more explicit cases of homophobia were also observable, contributing to a higher average toxicity score. Islam and Judaism also scored fairly high in terms of toxicity, though not nearly as often as the subject of sexuality. Predictably, both of these terms also scored significantly higher in toxicity than their hegemonic counterpart, Christianity.

One particularly concerning entry here is "Alt-Right". This refers to instances of commenters explicitly posting far-right dog whistles. While 10 seems relatively low, this counts only the observed instances in the peak sections, and only those explicit enough for this project to not give them the benefit of the doubt.

Other high-toxicity subjects also include the terms "Toilet Humor" and "Sexual" with scores of The Perspective API generally seemed to score these kinds of comments higher in toxicity due to the uncouth language used, though on a closer look during the qualitative analysis, perhaps not all of these comments deserved such a high toxicity score. Many did boil down to little more than juvenile humor, with little to no malicious intent behind them.

On the character side, a similar pattern emerges. In order from top to bottom, the table features: A woman, a Jew, an overweight person, and another woman before it finally gets to the heterosexual white male characters. And by that point, toxicity scores have already plummeted to about half what they were at the top.

The character of Meg was also notable. She is a character that was often discussed, and with significantly higher associated toxicity than other characters like Peter, Brian, Quagmire, or Stewie. As will be discussed in more detail later, she also scored close to a term that seemed to be strongly related to her mention: Death. The fact that comments involving death, dying, or killing scored high in toxicity with the Perspective API was no surprise, though it was notable how often the subjects of Meg and Death appeared together, hence their relatively close positioning within this plot.

The only other character who received similar levels of toxicity is the character of Mort. Mort, perhaps not coincidentally, is also a member of a minority group as he is "the Jewish character" of the show [21].

Fig. 7 provides an overview of subjects and their associated average toxicity. The figure took all the comments that have been labeled a subject, then from that calculated the average toxicity of the comments that each label has been associated with.

Each labeled subject is then placed in accordance with the average toxicity (y-axis) and the frequency with which it occurred (x-axis). Note that the plot has been adjusted to a logarithmic scale (using the natural log of the number of occurrences), as otherwise certain areas of the figure would be too densely packed to analyze. Only subjects that appeared at least twice have been included, and subjects with toxicity scores of less than 0.1 were also excluded.

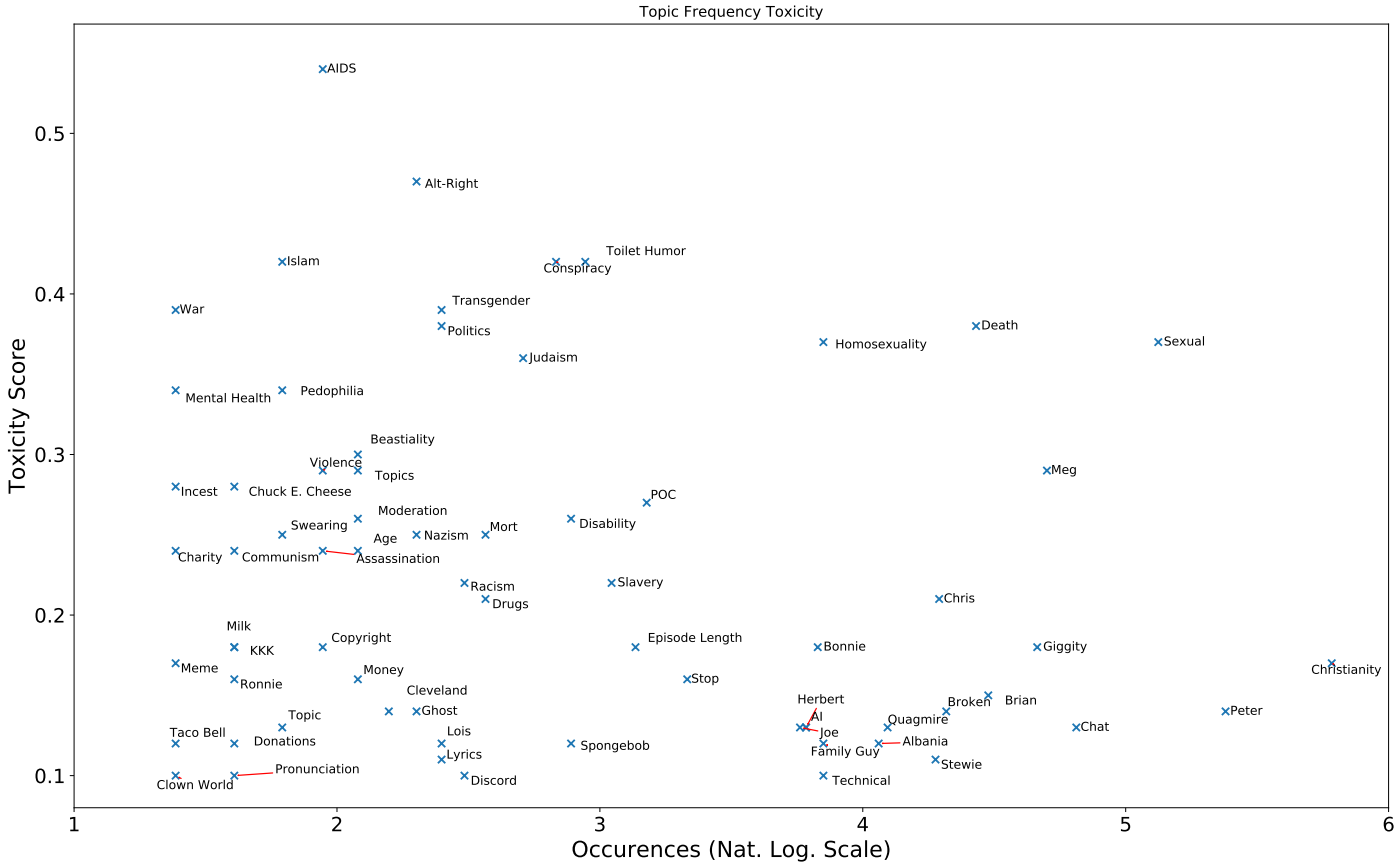


Figure 7: Toxicity associated with specific subjects. Plot has been transformed to a logarithmic scale for better legibility.

Clarifications Most subjects are self-explanatory, though a few require at least some context.

- "Topics" refers to the topic suggestions that people can submit through the Discord server. A group of commenters expressed their dislike of the topics people were suggesting, venting their frustrations about the quality of the suggestions and the preferential treatment that these paid suggestions were getting over regular ones.
- "Giggity" is Quagmire's catchphrase, used to respond to sexual innuendos or poorly worded statements that can be interpreted in a sexual way.
- "Technical Problems" refers to commenters expressing frustration over technical problems in the stream, like voices being either too loud or too quiet, 3D models not displaying properly, or the whole system freezing up or crashing.
- "Ghost" here is related to the previous point. 3D models clipping into walls were being compared to ghosts.
- "Politics" refers to any discussion about contemporary real-world politics and politicians.
- "Conspiracy" refers to the invoking of conspiracy theories, ranging from relatively harmless examples like the moon landing being fake to dangerous extremist ones like a secret Jewish cabal controlling governments from the shadows.
- "Alt-Right" refers to a specific collection of comments, all invoking alt-right rhetoric and humor. While it was difficult to distinguish between what was sincere and what was "just a joke", all of these comments were deliberately offensive in their language.
- "Episode Length" here refers to people complaining about episodes being either too short or too long.

Characters Quite a few entries in this graph refer to characters in the show. To delve into each character and their portrayal would be beyond the scope of this project, so the following points are the shortest, most condensed explanations of a few key characters [33]:

- *Peter*, whose full name is Peter Griffin, is the central character. He is a father of three children, none of whom he treats particularly well. He is also often portrayed as a careless, bumbling, childish idiot.
- *Lois* is Peter's wife and the mother of the aforementioned children. Though she plays a larger role in the original *Family Guy*, in the AI version her presence seems greatly diminished.
- *Meg* is the eldest of the three Griffin children. She is a self-conscious teenage girl, and often the target of ridicule.
- *Chris* is the middle child. Taking after his father, he too is often lazy and not particularly intelligent.
- *Stewie* is the youngest child. Although physically an infant, he is shown to be the most intelligent of the entire family, talking in a sophisticated upper-class accent and behaving with far more maturity than his siblings.
- *Brian* might be the family dog, but he too possesses the intelligence and self-awareness that many of the Griffin family lack. However, contrary to Stewie, he is more arrogant and self-righteous, often to his detriment.
- *Quagmire* is one of Peter's best friends. His defining character trait is his sex addiction. His catchphrase "giggity" which he uses to respond to innuendos has also become a popular reference on its own.
- *Joe* is Peter's disabled and wheelchair-bound neighbor and another one of his close friends. Joe works as a police officer and generally takes a calm and reserved attitude.
- *Bonnie* is Joe's soft-spoken wife. She generally takes a backseat to her husband and his friends' exploits, but she is friendly with Lois and has her moments.
- *Herbert* is a retired World War II veteran whose defining character trait is his pedophilic tendencies. A lot of jokes revolve around him making sexual advances towards Chris.
- *Mort* is another one of Peter's friends, though he does not appear as often as Quagmire or Joe. He also runs a pharmacy, and more importantly for this project, is of Jewish heritage.

6.4 Topic Modeling

Table 7 illustrates the topics that the topic extraction model identified. Different options for the number of topics k were tested, within the range of 6 to 12 topics. Fig. 8 shows the results of different models with different options for k. $k = 11$ was the option that produced the highest coherence score for the resulting model, so those 11 topics were extracted and will be presented here.

Topic 1: Pull the plug! This topic centers around technical issues. "Loudris" is a nickname given to the character of Chris, who seemed to suffer most often from audio issues making his voice way too loud in comparison to other characters. "Pull the plug" became a common phrase in response to technical issues, demanding a reboot of the system in hopes of fixing the issues.

Topic 2: Jesus and the car This topic seemed to represent a number of coincidental occurrences. "Jesus" is the central word here, which is strongly associated with words like "house" and "car". The former is a frequent cause of technical issues, causing people to invoke the name of Jesus Christ to express frustration or surprise. The car seems to be a total coincidence. The in-show character of Christ showed up at the same time that the car showed up. This was during the stream of August 21st. Jesus Christ gives a lecture about religion, specifically the differences between Christianity and Islam while standing in the driveway next to the car. This created a strong connection between him and the car, which was then reflected in the topic extraction model.

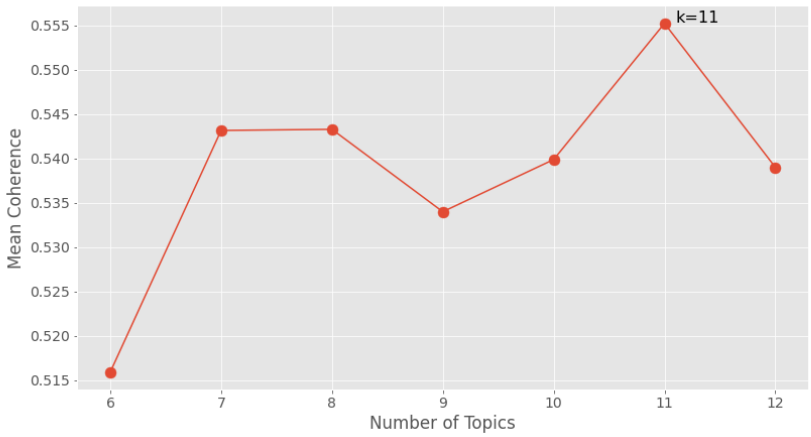


Figure 8: Coherence scores for different numbers of topics k.

Topic 3: Peter in the walls The stream was no stranger to technical problems. One such problem was character models clipping into scene geometry, sometimes completely disappearing into the walls. Hence, "Peter is in the walls" became a running joke. Whenever Peter's model would experience such clipping, chatters were quick to respond with this joke. "Giggity" here could also refer to his role as a "ghost", being invisible and therefore in a prime position to behave inappropriately.

Topic 4: I love this show This topic represents a community in-joke: "I love this show!". This joke emerged during the stream of August 20th, with some people expressing their love for the stream. Other people became annoyed, condemning the messages as spam. This response spurred on the first group of commenters, who began professing their love for the show even more aggressively to elicit more reactions.

Topic 5: "Crazy? I was crazy once. They put me in a room. [..] A rubber room with rats." This is an internet meme that was at its height of popularity in 2023, around the time of these streams. Most commonly, it was used in response to the word "crazy", hence this topic formed. Interestingly enough, the names that were associated most with this meme seemed to be Meg and Chris, again confirming them to be the target of ridicule.

Topic 6: Kill Meg, Giggity The idea to "kill" "Meg" was a surprisingly common topic suggestion. Unsurprisingly, both terms score fairly high in terms of toxicity, and it is no coincidence that they are ranked so closely together in the topic analysis either. The treatment of Meg is a point that will be discussed later on in more detail.

"Giggity" is Quagmire's catchphrase, both in the original show and the AI stream. "Quagmire", "Quaggy" and "Glen[n]" are naturally associated with the phrase, but it is interesting that Quagmire seems to be so closely connected to the killing of Meg.

Topic 7: Fish Milt Another community in-joke. During the stream of June 15th, "Fish Milt" became a commonly spammed meme, which seemed to have stuck around since then. It does not appear as if there is any deeper meaning here. More likely, the concept of fish milt, i.e. the seminal fluid of fish, was funny enough in and of itself to warrant spamming in the chat. Albania also makes an appearance here, another commonly spammed meme, see also Topic 11.

Topic 8: Brian's balls Brian was the target of a lot of sexual humor and comments, specifically his scrotum. This could be a reference to the fact that Brian is a dog, and a common practice for household pets is castration. Hence, Brian "knowing" and being angry at "Peter" for taking away his "balls" is a reflection of a wider internet meme, which imagines dogs to have the intelligence and self-awareness to realize what their owner has done to their body and express discontent over it.

Topic 9: Other characters This was a difficult topic to assess, as there did not seem to be a strong connection between the terms other than the fact that more secondary characters appeared here than in any other topic. "Mort", "Tricia", "Carter" and "Milton" together make up 4 of the terms in this topic, which already is more names in this topic than in any other. Peter also makes an appearance here, possibly the one thread to connect them all together. There also is a reference to the "news" here, likely the fictional news station of the show in which Tricia works. When her character was added to the stream, she was greeted with much applause.

Topic 10: Glenn Quagmire This topic seems to focus on the character of Quagmire in particular. It also refers to "summoning" or "bringing" him "back", from "hell" apparently. This is likely referring to a specific instance in the stream of June 15th, where a summoning was being performed on stream, causing the commenters in the chat to attempt their own summoning, of Quagmire.

Topic 11: Albania and the Honda Civic Albania and the Honda Civic were both running jokes in the chats across the different streams. There did not appear to be a deeper meaning behind them other than the fact that these connections were random and unexpected, and therefore funny.

Topic	1	2	3	4	5
Average Toxicity	0.13	0.14	0.16	0.16	0.27
Topic Words +	peter, 0.17	jesus, 0.16	giggity, 0.18	show, 0.05	rubber, 0.36
associated Toxicity	chris, 0.19	car, 0.07	peter, 0.17	love, 0.05	crazy, 0.37
	pull, 0.18	house, 0.07	walls, 0.19	joe, 0.13	rats, 0.36
	discord, 0.05	bonnie, 0.14	ai, 0.18	jesus, 0.16	room, 0.3
	plug, 0.17	joe, 0.13	ghost, 0.12	chris, 0.19	locked, 0.36
	loudris, 0.02	ai, 0.18	spammer, 0.32	madden, 0.1	meg, 0.3
	topic, 0.09	peter, 0.17	stewie, 0.15	peter, 0.17	chris, 0.19
	ai, 0.18	mort, 0.27	hamburger, 0.06	ai, 0.18	loud, 0.13
	stewie, 0.15	herbert, 0.13	theyre, 0.13	ia, 0.08	brian, 0.15
	based, 0.06	unity, 0.09	physics, 0.06	white, 0.49	invisible, 0.13

6	7	8	9	10	11
0.2	0.09	0.14	0.13	0.15	0.14
meg, 0.3	fish, 0.08	ball, 0.19	mort, 0.27	owner, 0.06	albania, 0.13
kill, 0.65	milt, 0.05	peter, 0.17	tricia, 0.11	quagmire, 0.16	albanian, 0.15
quagmire, 0.16	brian, 0.15	unity, 0.09	news, 0.13	summon, 0.04	peter, 0.17
particle, 0.02	bee, 0.02	brian, 0.15	stroke, 0.13	glenn, 0.03	civic, 0.02
giggity, 0.18	albanian, 0.15	quagmire, 0.16	carter, 0.02	bring, 0.11	honda, 0.0
quaggy, 0.1	whopper, 0.04	stroke, 0.13	car, 0.07	back, 0.14	chris, 0.19
stewie, 0.15	strokeie, 0.06	based, 0.06	milton, 0.03	hell, 0.46	money, 0.18
loud, 0.13	vs, 0.01	knows, 0.07	peter, 0.17	please, 0.13	ai, 0.18
glen, 0.02	stroke, 0.13	ai, 0.18	lois, 0.2	im, 0.2	doin, 0.26
cheese, 0.28	sure, 0.19	pull, 0.18	sorry, 0.17	donate, 0.11	

Table 7: Topic Modelling for peak sections. Each column represents the top 10 words of that topic, along with their associated toxicity scores. Average overall toxicity of all topic words combined is also noted.

7 Results - *Spongebob*

7.1 Quantitative

Date	26-03	06-04	05-05
Length (h)	11	10	1
Comments	30518	14271	3072
Average Stream Toxicity %	11.8	8.5	6.5
Average Chat Toxicity %	3.2	2.9	2.3
Median Stream Toxicity %	4.7	2.1	0.0
Median Chat Toxicity %	1.6	0.7	1.1

Table 8: Breakdown per individual stream. Toxicity scores have been mapped to a percentage value for better readability.

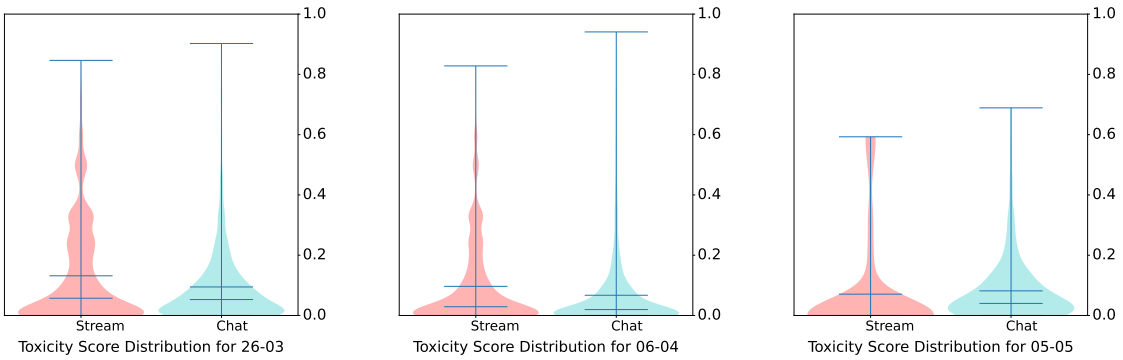


Figure 9: Toxicity score distributions for each stream.

Toxicity overall The availability and quality of data for the *Spongebob* streams was unfortunately much lower than that of the *Family Guy* streams. Nevertheless, with the same analysis performed, the average toxicity in the *Spongebob* streams and chats was similar to that of the *Family Guy* streams. Also similarly to the *Family Guy* streams, comments and lines with extremely low toxicity levels seem to dominate. This is further supported by the distribution densities illustrated in Fig. 9, which show very bottom-heavy distributions of the toxicity scores.

Date	26-03	06-04	05-05
Δ Stream to Stream Tox %	+3.1	+2.5	+2.3
Δ Chat to Chat Tox %	0.0	0.0	+0.4
Δ Stream to Chat Tox %	0.0	+0.0	+0.1
Δ Chat to Stream Tox %	+0.7	+0.7	+0.4

Table 9: Cross-examination per individual stream. Toxicity scores have been mapped to a percentage value for better readability.

Cross-Evaluation The cross-evaluation detailed in table 9 tells a similar story: While there were differences in toxicity levels between the peaks and non-peaks within the stream, when looking at the levels in the chat, a much more steady picture emerges. Neither the *stream-to-chat* nor *chat-to-chat* Δ are particularly high, indicating a remarkably stable level of toxicity. This also confirms a suspicion that developed during the execution of this project, where it became difficult for the algorithms used to detect any kind of peaks at all in the data of the *Spongebob* streams.

Date	26-03	06-04	05-05
Sample size N	4069	3474	260
Kruskal-Wallis H statistic	0.48	0.58	6.22
p-value	0.51	0.55	0.99

Table 10: Results of the Kruskal-Wallis tests determining if peak and non-peaks are different.

Kruskal-Wallis test The same Kruskal-Wallis tests were also performed on the AI Spongebob streams.

Table 10 details the results of these tests for each stream. As before, the tests tried to determine if there was a significant difference between the peaks and non-peaks in the chat, dictated by the peak timestamps from the stream. Once again, all p-values were greater than a significance level $\alpha = 0.05$, so no evidence was found that there was a significant difference.

7.2 Thematic Analysis

The thematic analysis of the *Spongebob* streams was difficult, as there was not a lot of data to draw conclusions from. Nevertheless, a few patterns were still observable.

Technical Problems As with the *Family Guy* streams, the *Spongebob* streams were no stranger to technical difficulties. Characters' voices were frequently too loud or too quiet, 3D models would not interact properly with each other, or the entire stream came crashing down on itself, requiring a full restart. These problems predictably caused frustration, which then in turn led to toxic comments in the chat. Nevertheless, as with the *Family Guy*, this toxicity was most often directed at the stream itself, at the technical processes that created the product, instead of any one person or group of persons.

Offensive Language Similarly to the *Family Guy*, and most other online communities, offensive language was often deployed for humor. Unlike the *Family Guy* streams though, the language used in the *Spongebob* was noticeably less discriminatory. On the less offensive end, drugs and toilet humor were frequently employed, just like in the *Family Guy* streams. The other end of the spectrum was far lower though. There still was swearing, ableism, and mild homophobia, but all of it was far more contained, far less frequent, and far less extreme than anything observed in the chats from the *Family Guy* streams.

Songs In the observed streams were repeated instances of characters "singing". While there was no actual music, melody, or for that matter, singing, the topic prompts that had been suggested pretended that the characters were indeed capable of musical performances. This seemed to have become a community favorite, and soon enough more "song prompts" were being submitted, watched, and laughed about.

7.3 Toxic Subjects

Subject	Count	Average Toxicity
Stink	9	0.65
Fat	13	0.39
Toilet Humor	19	0.38
Sexual	120	0.33
POC	9	0.31
Death	36	0.30
Fire	20	0.23
Drugs	39	0.22
Burger	16	0.20
Terrorism	23	0.19

Character	Count	Average Toxicity
Krabs	27	0.24
Larry	33	0.18
Bubble Bass	69	0.17
Sandy	20	0.15
Squidward	86	0.14
Plankton	140	0.12
Perch	28	0.10
Karen	26	0.10
Spongebob	125	0.10
Patrick	90	0.09

Table 11: Notable Toxicity Scores of Key Topics. The left table contains the 10 most toxic topics that were observed at least 5 times, ordered from most to least toxic. The right table does the same for all the characters.

The most toxic subjects Like table 6 did for the *Family Guy* streams, table 11 breaks down a selection of key subjects and their toxicity scores for the *Spongebob* streams. Although there are some repeat offenders and close parallels, this table as a whole paints a very different picture of the kinds of toxicity in both streams. Sexual humor and comments still scored high, predictably, as did the subjects of "Stink" and "Toilet Humor", both terms mirroring the *Family Guy* stream's own "Toilet Humor" in some way. "Sexual" also wins out by frequency, counting over 100 instances when other subjects barely breach the double digits. Most other entries in the table are either referencing very specific in-jokes or the events that happened on screen.

"Terrorism" refers to an episode prompt where Spongebob was planning a bomb attack on Plankton's restaurant. "Burger" references the Krabby Patty, the famous burger made by Mr. Krabs's fast food establishment. The only problematic entries in this table are "POC", "Drugs", "Death", and "Fat".

On the character side, the closest to an entry like Meg, who was a true target of toxicity in the *Family Guy* streams, would be Mr. Krabs or Plankton. As the primary villain of the original, Plankton would make sense. Plankton is not an explicit representation of any kind of demographic group, however, which may be a contributing factor in his comparatively low toxicity levels. In fact, the toxicity he received seems surprisingly low, given his role as a villain. He sits comfortably in the middle of the ranking, not receiving any more toxicity than the rest of the cast. Mr Krabs had the highest associated toxicity score, getting close to Meg's 0.29. In Mr. Krabs's case though, the frequency with which he was being discussed counts to only 27 observed instances. This is especially low when compared to Meg, who wins the most toxic responses not only by toxicity score but also lands 2nd place in terms of frequency, her 110 instances only being eclipsed by Peter's 217, the main character of *Family Guy*.

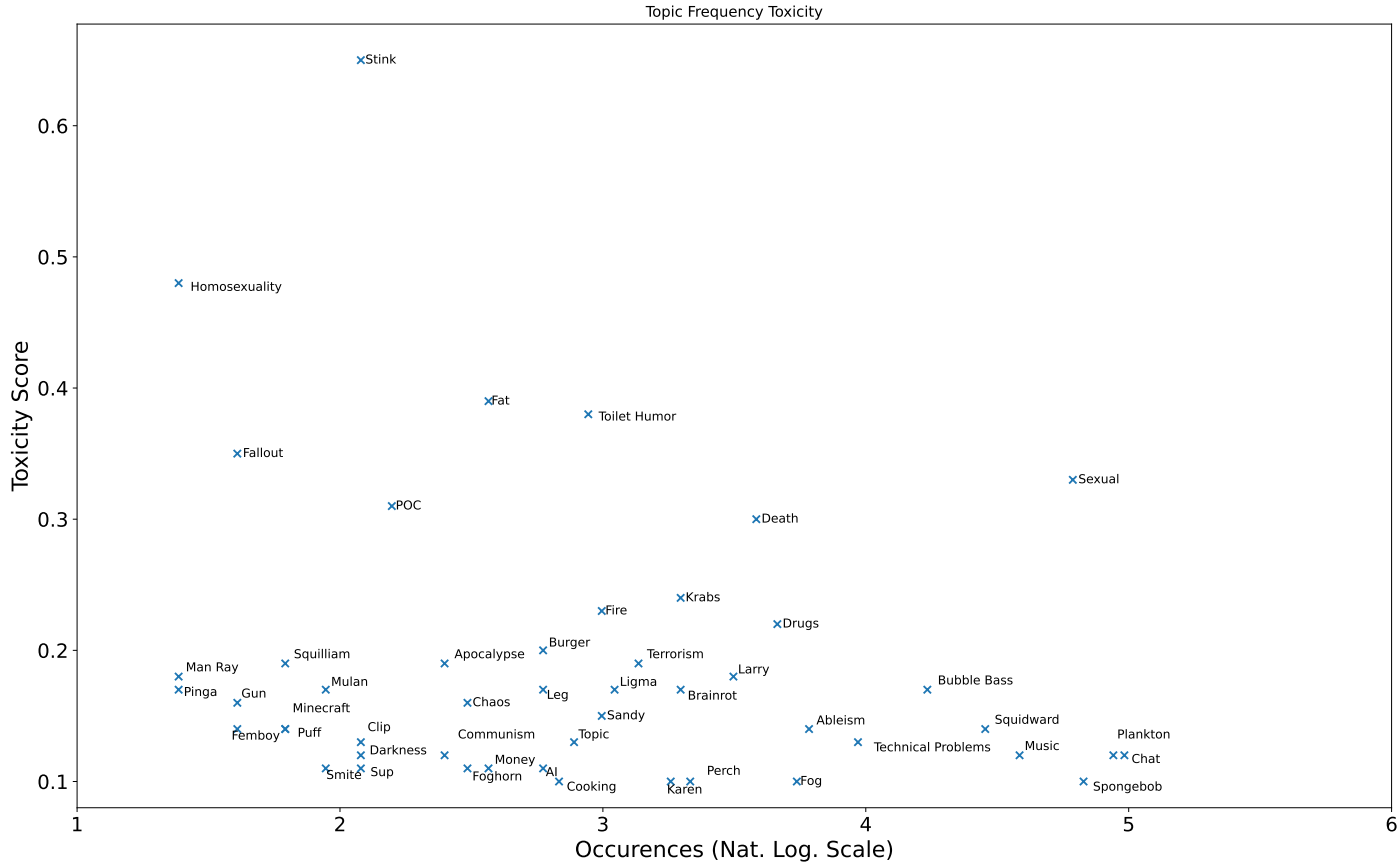


Figure 10: Toxicity associated with specific subjects. Plot has been transformed to a logarithmic scale for better legibility.

Clarifications Most subjects are self-explanatory, though a few require at least some context.

- "Stink" mostly references a specific community in-joke: "Oh brother, this guy stinks!". It is a line from an episode of the original *Spongebob Squarepants*. The titular character *Spongebob* attempts to perform a stand-up comedy routine but fails horribly. In response, one audience member shouts this line [34]. The line has since taken on its own life as an internet meme and has now found its way back to the AI parody. It was used most often in response to a bad script, or to express one's general discontent with the AI stream.

- "Clip" refers to people commenting some variation of "clip this!", a phrase which here means to save a several seconds long segment of the stream. The resulting video segment would then be referred to as a clip, hence the process of creating it is called "clipping".

- "Cooking" here is shorthand for "letting someone cook", a slang way of saying "do not interfere with or stop whatever that person is doing or saying because I think they are on to something."

- "Fallout" refers to the popular video game series. More specifically, this refers to an occurrence where characters on the show sang the song "Big Iron", which is heavily associated with one of the installments of the Fallout franchise.

- "Brainrot" and "Ligma" are both slang terms. For this project, both terms refer to people commenting with the use of different slang terms. "Brainrot" describes newer terms used primarily by Generation Alpha, while "Ligma" is a slightly older play on words. It is a way of telling someone to "lick my [genitals]".

Characters Quite a few entries in this graph refer to characters in the show. To delve into each character and their portrayal would be beyond the scope of this project, so the following points are the shortest, most condensed explanations of a few key characters [35]:

- *Spongebob*, the titular character, is a happy-go-lucky anthropomorphic sponge. He is characterized by his relentlessly positive attitude and child-like enjoyment of the world.
- *Patrick* is a starfish and Spongebob's best friend. While he is slow-witted to the point that he canonically does not possess a brain, he is fiercely loyal to his friends and is one of the most beloved characters of the show.
- *Squidward* is an octopus, and Spongebob and Patrick's neighbor. The anthropomorphic squid's defining character traits are his perpetual annoyance at anything and everything. He also believes himself to be much more talented and intelligent than he actually is.
- *Eugene Krabs*, most often referred to as "Mr. Krabs", is the owner of a fast-food restaurant and Spongebob and Squidward's boss. The restaurant is highly successful, but Mr. Krabs is also greedy to the point of absurdity.
- *Plankton* owns a rival fast-food restaurant, which is a complete failure. To fix this and attract patrons, his goal throughout the series is to steal Mr. Krabs's "secret formula". His schemes of course never succeed, though that never stops him from trying again a few episodes later.
- *Karen* is Plankton's robotic restaurant manager and wife. She assists him in his endeavors and has little role outside of that relationship.
- *Sandy* is a squirrel, and the only non-aquatic animal in the fictional town the show takes place in. She lives in a glass dome that houses her tree and a small grass lawn and survives outside of that by wearing a waterproof diving suit. She is Spongebob and Patrick's friend and is shown to be intelligent and inventive.
- *Larry* is a lobster and a bodybuilder. He works as a lifeguard at the underwater beach and is characterized by his musculature and strength.
- *Bubble Bass* is another minor character. He is a stereotypical nerd, usually depicted as unattractive, unhygienic, and generally unpleasant to be around.
- *Ms. Puff* is a pufferfish and Spongebob's driving instructor. Spongebob unfortunately is a horrendous driver, however, and is never able to get his driver's license. This places Ms. Puff in an eternal purgatory of having to endure Spongebob's terrible driving.
- *Squilliam* is another octopus. He exists as a polar opposite to Squidward, who is successful in everything that Squidward wishes he would be.
- *Perch Perkins* is a fish and local news reporter. This character only makes occasional appearances and is never directly involved in the plot of any episode.
- *Man Ray* is half man, half manta ray. He is another villain of the show, more akin to the kind of villain a superhero from a comic book would fight.

7.4 Topic Modeling

Topic	1	2	3	4	5
Average Toxicity	0.11	0.18	0.11	0.1	0.12
Topic Words +	firstname, 0.12	perch, 0.11	bomb, 0.26	pingas, 0.15	skip, 0.11
associated Toxicity	lastname, 0.12	news, 0.07	strokaren, 0.09	machine, 0.15	crowd, 0.03
	wake, 0.08	stroke, 0.11	spongebob, 0.1	crushing, 0.15	fire, 0.2
	fog, 0.09	song, 0.05	stroke, 0.11	tacos, 0.1	plankton, 0.11
	stroke, 0.11	larry, 0.19	plankton, 0.11	lag, 0.09	song, 0.05
	patrick, 0.11	plankton, 0.11	company, 0.01	eating, 0.12	stroke, 0.11
	fire, 0.2	finland, 0.01	car, 0.06	music, 0.06	peak, 0.05
	plankton, 0.11	discord, 0.04	loudton, 0.03	update, 0.07	gary, 0.12
	squirrel, 0.08	nut, 0.41	bubble, 0.17	name, 0.04	hole, 0.34
	song, 0.05	stinks, 0.68	bass, 0.12	nick, 0.07	please, 0.08

6	7	8	9	10	11
0.11	0.1	0.13	0.1	0.14	0.12
pull, 0.16	fakeyou, 0.3	ligma, 0.19	bubble, 0.17	voices, 0.05	cum, 0.44
plug, 0.14	paramount, 0.01	kinitopet, 0.2	bass, 0.12	meme, 0.07	plankton, 0.11
crowd, 0.03	march, 0.01	chat, 0.33	forces, 0.03	penis, 0.67	loudton, 0.03
gary, 0.12	free, 0.01	drive, 0.03	sonic, 0.07	perch, 0.11	fog, 0.09
leave, 0.06	heat, 0.16	karen, 0.08	squirrel, 0.08	glass, 0.06	boi, 0.02
bubble, 0.17	ai, 0.14	kids, 0.14	plankton, 0.11	crowd, 0.03	karen, 0.08
patrick, 0.11	death, 0.26	plankton, 0.11	song, 0.05	telling, 0.18	stroke, 0.11
bass, 0.12	fight, 0.01	pluh, 0.06	fist, 0.18	smells, 0.14	honda, 0.02
sandy, 0.16	lag, 0.09	stroke, 0.11	seidel, 0.12	robot, 0.02	esq, 0.02
loudton, 0.03	stream, 0.04	lore, 0.01	sam, 0.12	ahh, 0.04	snop, 0.31

Table 12: Topic Modelling for peak sections. Each column represents the top 10 words of that topic, along with their associated toxicity scores.

Referencing table 12, the process of identifying topics in the *Spongebob* streams was more difficult. Fewer streams to pull from and a much shorter overall runtime were contributing factors, along with a much smaller community of commenters. Not all resulting topics are as easily interpretable as those from the *Family Guy* stream. Nevertheless, for this stream, $k = 11$ also yielded the highest coherence score, as illustrated by Fig. 11

Topic 1: Firstname Lastname "Firstname Lastname" is the anonym given in place of a name that kept appearing over and over. There is no character of that name in the show, but there does exist a rather active member of the community going by that name. This topic references them, along with several keywords most often brought up to them, mainly about technical problems.

Topic 2: Perch Perkins, News Reporter Perch Perkins is a character, a news reporter, in *Spongebob Squarepants*. He reports on events live from the field, but seems to be suffering from many technical problems, hence the "stroke", "nuts" and "stinks" here as well.

There is also a second sub-topic here. "Song" and "Larry" refer to an instance early on in the stream where the character Larry started singing songs from Disney's *Mulan*.

Topic 3: Spongebob the Terrorist This topic references events in the stream, during which bombings were performed on Plankton's restaurant. Spongebob was the instigator of these attacks, and Karen and Plankton seemed to be the victims.

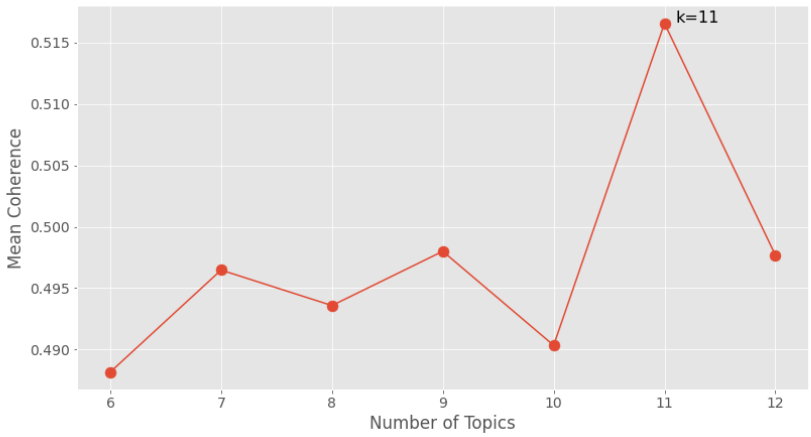


Figure 11: Coherence scores for different numbers of topics k.

Topic 4: Pingas Crushing Machine, and Tacos This topic illustrates two separate but connected instances of spam. "Pingas crushing machine", as well as "eating tacos" were commonly spammed jokes in the chat. During the second examined stream of May 5th, both the machine and tacos became big reoccurring jokes, seemingly independent of each other but nonetheless coinciding. Chatters were spamming about their love of "eating tacos" or the "crushing machine", much to the annoyance of other chatters.

Topic 5: This song is fire The "fire" here refers not to an actual fire, but to the expression of a song "being fire", another way of saying it is really good. Characters would sometimes "sing" in the stream, to which the audience responded with applause, calling for more. Not everyone seemed to enjoy these not-so-musical musical numbers though, calling for a "topic skip". It should be said that these were not actual musical numbers, merely the AI characters reading out lines of a song in their regular voices, but the audience usually didn't mind that shortcoming.

Topic 6: Pull the plug This topic mirrors a topic found in the *Family Guy* streams: Technical problems causing so much frustration that chatters began to call for a total reset, i.e. "pulling the plug". The characters mentioned in this topic were the ones exhibiting the issues, thereby getting mentioned in the demands as well.

Topic 7: The apocalypse is coming "The Heat Death of the universe is coming soon" was another running joke in the community. It was said to happen at the end of March, causing many chatters to spam that the end was nigh and everyone should live their last days free. This apocalypse was a reference to Paramount, the owner of the intellectual property rights of *Spongebob Squarepants*, who would ultimately shut down the AI version over claims of copyright infringement.

Topic 8: Karen and the kids Among other things, this topic features Plankton, Karen, and "Kids". This refers to a specific topic in the stream from April 6th, where a divorce was happening between the characters. With quite unfortunate timing, Karen also seemed to be suffering from technical issues at the exact same time, creating a rather humorous narrative among the chatters that her malfunction was causing her to behave irrationally. Or as they put it, "having a stroke".

Topic 9: Bubble Bass Bubble Bass is a character on the original show, a caricature of the stereotypical "nerd". This topic references him being compared to moderators on Discord, who have a similarly poor reputation among other internet communities. The video game Sonic Forces also seems to make an appearance here, perhaps about Bubble Bass being a big fan of said game.

Topic 10: Perch and the voices Perch Perkins made another appearance earlier, but this topic seems to focus more on the inner workings of his mind, as theorized by the chatters. Perch hearing voices inside of his head telling him to do various things was a common joke.

Topic 11: Sexual humor around Plankton and Squidward Plankton is an easy character to make fun of, seeing how his role in the original show is exactly that: A target of ridicule. This topic took that into the AI version as well, as sexual humor around Plankton's seminal fluid and relationship with his robot wife was abundant.

8 Discussion

8.1 Quantitative Analysis: General Obervations

None of the tests performed during the quantitative analysis of this project yielded any convincing evidence in favor of the research questions. This suggests that perhaps this approach was not sufficient for investigating these questions. Since the qualitative analysis did yield interesting results, the focus of this project shifted more towards investigating those, and the quantitative methods were not expanded upon much further. Future research endeavors will have to refine the admittedly rather basic methods implemented in this project.

Family Guy The experiments using the outlined quantitative methodology did not yield any conclusive results regarding a correlative connection between the two data sets. There was no strong correlation found between toxicity levels in the stream and the live chat.

Spongebob Unfortunately, the data from the *Spongebob* streams was too sparse to draw any definitive conclusions from, at least in a quantitative sense. Toxicity in these streams was at comparable levels, quantitatively speaking. One observable difference was that the toxicity levels in the *Spongebob* streams were less volatile and more stable, as the differences between peaks and non-peaks in the chat were less pronounced.

Comparison While the toxicity levels of both AI-generated versions are comparable when considering the original shows both streams are based on, that seemingly unremarkable observation does bring up a far more interesting question: Why?

When comparing the originals, no one would say that *Spongebob Squarepants* is anywhere near as transgressive or offensive a show as *Family Guy* can be. This is not to say it does not have its own flaws or problems, but it certainly does not lie with a reliance on problematic humor. Yet, both AI versions are similar, at least when evaluated in terms of how toxic their parody episodes tend to be. While it ultimately fell out of the scope of this project, this insight does imply a few things about the generative AI model that powers both streams. It lines up with other studies that suggest a homogenizing effect produced by LLM models like ChatGPT [36]. Both shows lose their unique characteristics in their AI parody, resulting in a much more homogeneous, less diverse experience for the viewer.

8.2 Research Question 1: How does toxicity develop throughout AI-generated live streams?

8.2.1 Does biased/harmful content appearing on screen invite toxicity in the live chat?

The quantitative analysis did not yield any solid evidence that toxic content appearing in the stream has a measurable impact on toxicity levels within the live chat. Although some data points do seem to point in that direction, none are strong enough to serve as conclusive evidence for this claim, whether taken individually or as together.

8.2.2 Does toxicity in the live chat lead to more biased/harmful prompts being submitted to the AI?

The cross-evaluation of both streams did not yield any conclusive results either. The complicating factor was that topic suggestions would enter a queue, so there would naturally be a substantial delay before the prompt would appear on screen. On top of that, these queues could also be entered through other means, via a website or the Discord server, neither of which could be captured in this project.

8.2.3 Do these two phenomena lead to a feedback loop of toxicity and harmful content?

As neither of the preceding subquestions could be answered with a "yes", the immediate answer to this third subquestion should also be "no". However, when comparing the two different streams of AI *Family Guy* and AI *Spongebob*, significant differences were observable. Notably, the extent and kind of toxicity present within either stream differed drastically. This could be a product of a long and slow feedback loop, where over time, the AI *Family Guy* community has devolved into a much more bigoted form of toxicity, while the AI *Spongebob* community remained fairly harmless in its toxicity.

That being said, this is merely a hypothesis, and not supported by any of the results of this project. Another equally valid explanation is that the original shows both streams were based on are significantly different, and therefore would attract significantly different kinds of communities. So, ultimately, the question of feedback loops remains unresolved, as this project could not find evidence for their existence.

8.2.4 Summary

In summary, the methods used for this project did not yield conclusive evidence for the existence of a strong causal link between toxicity levels in both stream and chat. This does not mean this link does not exist, as will be explained in the section 8.3. However, the methods used in this project were insufficient or inappropriate for measuring this correlation.

8.3 Research Question 2: What form does toxicity take in AI-generated live streams?

8.3.1 Which topics discussed within the live chat trigger more toxic reactions?

The results of the qualitative analysis laid out several patterns. These patterns very much resemble typical patterns found in online communities.

Minorities and the characters belonging to any such minorities received more toxicity. It was often difficult to determine whether a joke was in good spirits or not, or if it was a joke at all. Explicit and directed hateful messaging was rare, but not absent. Offensive humor was commonplace and received little to no pushback from other commenters, and more often celebrated or at least reciprocated.

Parallels to the original Family Guy Comparing these findings to the original *Family Guy* TV show, there are parallels to be drawn: The show is no stranger to using derogatory language [37]. This trend seems to have been replicated, both in the stream and the content it shows on screen, as well as the chat reactions. Anything from mild offensive language to posting far-right "jokes" and talking points. It should be noted here that in written form, comments on the internet lack the nuance that natural speech inherently possesses. This project cannot provide a judgment on these kinds of comments, on whether or not these are authentic and sincere, or whether these commenters are "just joking".

Nevertheless, the existence of this offensive language in the chat, sincere or not, is enough to have a noticeable impact on the overall community. Once it appeared in the 20-08 stream, it stuck around for the remainder of the stream. Despite some pushback from other commenters, offensive jokes and references to controversial political figures from the far-right became commonplace. Once again, it was impossible to determine how much of this was people merely pushing boundaries and how much was genuine alt-right dog-whistling.

Meg The character of *Megatron Harvey-Oswald Griffin*, also known simply as "Meg", was one of the biggest targets of toxicity. This came most often in the form of attacks on her identity, and physical threats. Identity attacks came in the form of misogynistic comments, both about and towards her. Although these are not real threats, Meg being a fictional character of course, so instead these "death threats" were in the form of topic suggestions involving her death. Nevertheless, the enthusiasm and apparent delight that these topic suggestions were expressing about Meg's demise was substantial. Meg was also disproportionately the target of such attacks and threats even when compared to other female characters from the stream, a phenomenon which parallels the treatment of her in the original *Family Guy* TV show.

AI Spongebob Compare this to the AI *Spongebob* stream. Not only did individual characters appear far less often as the target of toxicity, but the toxicity they did receive was neither excessive nor disproportionate. The difference between the most and least hated (if "hated" is defined by the toxicity levels they receive) was minuscule in comparison to the differences observed between *Family Guy* characters. And while there certainly still existed derogatory language and offensive humor in the *Spongebob* streams, it was far less transgressive than that observed in the *Family Guy* streams. These differences, while not immediately obvious from the quantitative analysis, did become apparent in the thematic analysis. While absolute toxicity levels were comparable, there were clear differences in who or what this toxicity was directed at.

8.3.2 Which types of toxicity appear more often than others?

Minorities As mentioned above, the *Family Guy* stream mirrored the original show in the types of toxicity that viewers engage in. Ableism, racism, misogyny, queerphobia, islamophobia, and antisemitism were all observable phenomena, both in the quantitative and qualitative analyses. Any minority group was equally valid as a target of ridicule, whether benign or malicious.

Politics Another unsurprising entry in the list of toxicity-triggering subjects was politics and anything politically contentious. Whether it was past or present elected officials, elections, geopolitical powers and the conflicts they are involved in, or social issues like climate change, equality, and discrimination; all of those and more were a reliable way to start a heated discussion that would inevitably result in toxicity. Any time these topics appeared on screen or in the chat, toxicity was sure to follow.

Even more extreme political affiliations came up, with one *Family Guy* stream quickly filling up with far-right language and imagery. As with many internet phenomena, it was difficult to tell how sincerely these beliefs were being expressed, but the effects of their presence were clear. This happened in the *Family Guy* stream of August 20th, 2023, which made that stream the most toxic one of all the examined streams.

Technical Difficulties Both the *Family Guy* and the *Spongebob* streams suffered from plenty of technical difficulties. Understandably, reactions to this were frustration and anger, often resulting in a wave of toxic reactions whenever a character's voice was too loud or when the stream would unexpectedly crash and need to be restarted. The toxicity that followed technical problems would vary depending on the severity of the problem and the character associated with it, if it was a character-specific issue. Generally speaking, this kind of toxicity was the least extreme or problematic kind though, aimed mostly (and justifiably) at the technical processes governing the stream. Jokes about and attacks on characters were still made, though mostly in rather harmless forms. These were comments like "Loudris", expressing one's frustration about Chris's loud voice; or "Strokeward" joking about Squidward's incomprehensible speech.

Sexual and toilet humor Another common occurrence in both communities was the use of sexual and toilet humor. For the most part, both of these types of toxic comments were fairly harmless though. Rarely did a comment joking about fecal matter express anything other than juvenile humor, rarely did a sexual joke be genuinely offensive. Nevertheless, invoking such language does tend to correlate with a higher toxicity score with the Perspective API. When looking at the distribution of toxic subjects in both streams, it seems to indicate that the AI *Spongebob* communities relied more heavily on such jokes, especially its use of the "stink" keyword. This could be an explanation as to why, despite scoring relatively equal average toxicity scores, the two communities still feel very different when looked at with a closer eye.

8.3.3 How does the source material affect the toxicity in the AI versions?

The most striking observation made in the execution of this project was the comparison between the *Family Guy* and *Spongebob* AI-generated streams. More specifically, how vastly different the accompanying communities were in terms of the toxicity they exhibited. While the toxicity levels of the streams themselves were comparable, and even the toxicity scores measured by the Perspective API were still similar, stark differences became visible once the chats were examined during the thematic analysis.

The toxicity found within the community for the *Family Guy* stream was far more offensive, often aimed at minority groups or members thereof. This was most obvious in the way that Meg was treated, as explained above, especially in comparison with her male peers. Comparing the female characters in the *Spongebob* streams to Meg, they received nowhere near as much toxicity, if they even did receive it in the first place. The female character with the highest associated toxicity score is Sandy, who sits right in the middle when ranking all characters and their toxicity. Even the original show's villain, Plankton, did not even reach the top 3 most hated characters in the AI stream, and that is despite his role as a villain who, similar to Meg, exists in the original mainly as a punching bag and the butt of many jokes at his expense.

While the difference in genders could be a partial explanation for this difference, Plankton being a man as opposed to Meg being a woman, this is only one example of a broader trend: The community for AI *Spongebob* treated its characters far more equally and far better in general. Jokes were still made, of course, and offensive language was still employed occasionally for humor and shock value, but nothing compared to the AI *Family Guy* community.

One other possible explanation could be community size. The AI *Family Guy* streams were more popular and active, with the ai_peter channel having about four times as many subscribers as AI Sponge Rehydrated. Chat activity was therefore higher in the AI *Family Guy* streams, and with bigger communities comes more toxicity.

Still, this alone would not explain the stark contrast between the two streams. A more likely explanation is that the AI *Family Guy* stream would attract fans of the original *Family Guy*, willing to be offensive and transgressive with their humor, while the AI *Spongebob* stream would attract fans of the original *Spongebob*, a (for the most part) very safe and respectful show that does not rely on offensive humor. In turn, the two different fanbases form two different communities with different attitudes towards humor, and different opinions on what is and is not acceptable to say or joke about.

9 Conclusion

This project examined two AI-generated parodies of popular TV shows: AI Sponge Rehydrated, inspired by the show *Spongebob Squarepants*, and ai_peter, inspired by the show *Family Guy*. Streams of these two parody shows were collected and evaluated using both quantitative and qualitative methods, aiming to answer questions about toxicity, how it develops and evolves, what forms it takes, and how toxicity is informed by the stream and its original inspiration.

Quantitative results were inconclusive. There was not enough evidence of a strong correlative link between toxicity levels in the streams and their accompanying chat logs.

The qualitative methods did yield interesting results though, identifying topics that were associated with higher toxicity scores. Politically contentious topics were unsurprisingly strong contenders for toxic subjects. Characters representing minority groups were also more often the target than their hegemonic peers, especially if the original show that the stream was based on also treated them poorly.

Lastly, there were significant differences between the two AI streams. AI *Family Guy* exhibited far more aggressive and offensive language. Even though the Perspective API did not seem to deem either stream significantly more toxic than the other, this resulted in a much more problematic atmosphere in the AI *Family Guy* community compared to its *Spongebob* counterpart, which for the most part was harmless in its toxicity. The AI *Spongebob* community relied far more heavily on sexual and toilet humor, which are far less harmful kinds of humor than jokes relying on homophobic, racist, or otherwise bigoted stereotypes.

In conclusion, while there was no strong evidence found for a direct link between streams and chats, it does appear that the style of show an AI parody is instructed to imitate does heavily influence the kinds of toxicity that will develop among the viewers of said AI parody.

References

[1] X. Fang, S. Che, M. Mao, H. Zhang, M. Zhao, and X. Zhao, *Bias of ai-generated content: An examination of news produced by large language models*, 2023. arXiv: 2309.09825 [cs.AI].

[2] N. Gross, "What chatgpt tells us about gender: A cautionary tale about performativity and gender biases in ai," *Social Sciences*, vol. 12, no. 8, 2023, ISSN: 2076-0760. DOI: 10.3390/socsci12080435. [Online]. Available: https://www.mdpi.com/2076-0760/12/8/435.

[3] D. Milmo and A. Hern, "Google chief admits 'biased' ai tool's photo diversity offended users," *The Guardian*, Feb. 28, 2024. [Online]. Available: https://www.theguardian.com/technology/2024/feb/28/google-chief-ai-tools-photo-diversity-offended-users (visited on 02/28/2024).

[4] K. D. A. Carillo and J. Marsan, ""the dose makes the poison"-exploring the toxicity phenomenon in online communities," *Thirty Seventh International Conference on Information Systems*, 2016.

[5] ai_peter. "Ai generated family guy parody," Youtube. (2024), [Online]. Available: https://www.youtube.com/watch?v=zD9wofGof80.

[6] A. S. Rehydrated. "Ai generated sponge," Youtube. (2024), [Online]. Available: https://www.youtube.com/watch?v=J3Y-Ly6H9fg.

[7] A. B. B. Rebooted. "Ai breaking bad rebooted, suggest topics on discord," Youtube. (2024), [Online]. Available: https://www.youtube.com/watch?v=wOikipG1hEo.

[8] WatchMeForever. "Nothing, forever," Twitch. (2024), [Online]. Available: https://www.twitch.tv/watchmeforever.

[9] A. F. Guy. "This is how we made ai family guy," youTube. (), [Online]. Available: https://www.youtube.com/watch?v=VoM9L-gDZtQ.

[10] "Perspective api," Jigsaw. (2021), [Online]. Available: https://perspectiveapi.com/.

[11] A. Köchling and M. C. Wehner, "Discriminated by an algorithm: A systematic review of discrimination and fairness by algorithmic decision-making in the context of hr recruitment and hr development," *Business Research*, vol. 13, no. 3, pp. 795–848, 2020.

[12] B. J. Dietvorst, J. P. Simmons, and C. Massey, "Algorithm aversion: People erroneously avoid algorithms after seeing them err.," *Journal of Experimental Psychology: General*, vol. 144, no. 1, p. 114, 2015.

[13] M. Ragot, N. Martin, and S. Cojean, "Ai-generated vs. human artworks. a perception bias towards artificial intelligence?" In *Extended abstracts of the 2020 CHI conference on human factors in computing systems*, 2020, pp. 1–10.

[14] Y. Zhang and R. Gosline, "Human favoritism, not ai aversion: People's perceptions (and bias) toward generative ai, human experts, and human–gai collaboration in persuasive content generation," *Judgment and Decision Making*, vol. 18, e41, 2023.

[15] N. Beres, J. Frommel, E. Reid, R. Mandryk, and M. Klarkowski, "Don't you know that you're toxic: Normalization of toxicity in online gaming," May 2021, pp. 1–15. DOI: 10.1145/3411764.3445157.

[16] M. J. Crockett, "Moral outrage in the digital age," *Nature human behaviour*, vol. 1, no. 11, pp. 769–771, 2017.

[17] G. Beknazar-Yuzbashev, R. Jiménez Durán, J. McCrosky, and M. Stalinski, "Toxic content and user engagement on social media: Evidence from a field experiment," *Available at SSRN 4307346*, 2022.

[18] W. A. Hamilton, O. Garretson, and A. Kerne, "Streaming on twitch: Fostering participatory communities of play within live mixed media," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '14, Toronto, Ontario, Canada: Association for Computing Machinery, 2014, pp. 1315–1324, ISBN: 9781450324731. DOI: 10.1145/2556288.2557048. [Online]. Available: https://doi.org/10.1145/2556288.2557048.

[19] L. Dreier and J. Pirker, "Toxicity in twitch live stream chats: Towards understanding the impact of gender, size of community and game genre," in *2023 IEEE Conference on Games (CoG)*, IEEE, 2023, pp. 1–4.

[20] H. Almerekhi, H. Kwak, J. Jansen, and J. Salminen, "Detecting toxicity triggers in online discussions," Sep. 2019, pp. 291–292, ISBN: 978-1-4503-6885-8. DOI: 10.1145/3342220.3344933.

[21] M. R. Porsgaard, "Semitic stereotypes: How being a jewish stereotype, speaking jewish english and being a negative character are correlated on south park and family guy," *Leviathan: Interdisciplinary Journal in English*, vol. 5, pp. 33–45, 2019.

[22] V. Nagy, "Motherhood, stereotypes, and south park," *Women's Studies*, vol. 39, no. 1, pp. 1–17, 2010.

[23] M. Sienkiewicz and N. Marx, "Click culture: The perils and possibilities of family guy and convergence-era television," *Communication and Critical/Cultural Studies*, vol. 11, no. 2, pp. 103–119, 2014.

[24] E.-C. Gross, "Artificial intelligence for the generation of satirical articles-an exploratory approach," *Bulletin of the Transilvania University of Braşov, Series VII: Social Sciences and Law*, vol. 15, no. 2, pp. 231–240, 2022.

[25] O. H. Hamid, "Chatgpt and the chinese room argument: An eloquent ai conversationalist lacking true understanding and consciousness," in *2023 9th International Conference on Information Technology Trends (ITT)*, 2023, pp. 238–241. DOI: 10.1109/ITT59889.2023.10184233.

[26] J. R. Searle, "Minds, brains, and programs," *Behavioral and brain sciences*, vol. 3, no. 3, pp. 417–424, 1980.

[27] J. Searle, *The chinese room*, 1999.

[28] L. Hauser, "Searle's chinese box: Debunking the chinese room argument," *Minds and Machines*, vol. 7, no. 2, pp. 199–226, 1997.

[29] R. I. Damper, "The logic of searle's chinese room argument," *Minds and Machines*, vol. 16, pp. 163–183, 2006.

[30] YCS. "Cs - youtube comment search chrome browser extension." (Oct. 4, 2024), [Online]. Available: https://chromewebstore.google.com/detail/ycs-youtube-comment-searc/pmfhcilikeembgbiadjiojgfgcfbcoaa.

[31] R. Řehůřek and P. Sojka, "Software Framework for Topic Modelling with Large Corpora," English, in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, http://is.muni.cz/publication/884893/en, Valletta, Malta: ELRA, May 2010, pp. 45–50.

[32] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.", 2009.

[33] Wikipedia, *List of characters in the Family Guy franchise — Wikipedia, the free encyclopedia*, http://en.wikipedia.org/w/index.php?title=List%20of%20characters%20in%20the%20Family%20Guy%20franchise&oldid=1272628535, [Online; accessed 01-February-2025], 2025.

[34] sakshi. "Crazy? i was crazy once [...]," Know Your Meme. (), [Online]. Available: https://knowyourmeme.com/memes/oh-brother-this-guy-stinks.

[35] Wikipedia, *List of SpongeBob SquarePants characters — Wikipedia, the free encyclopedia*, http://en.wikipedia.org/w/index.php?title=List%20of%20SpongeBob%20SquarePants%20characters&oldid=1269497312, [Online; accessed 01-February-2025], 2025.

[36] B. R. Anderson, J. H. Shah, and M. Kreminski, "Homogenization effects of large language models on human creative ideation," in *Proceedings of the 16th conference on creativity & cognition*, 2024, pp. 413–425.

[37] L. D. Ricke, "Funny or harmful?: Derogatory speech on fox's family guy," *Communication Studies*, vol. 63, no. 2, pp. 119–135, 2012.