Optimizing Language Models for Safe Online Discourse: Developing Metrics and Models for Detoxifying Internet Conversations

Stanford CS224N Custom Project

Steven Li
Department of Computer Science
Stanford University
stevenli@stanford.edu

Steven Le
Department of Computer Science
Stanford University
stevenle@stanford.edu

Abstract

As more difficult conversations begin to shift online, the increase in internet discourse and toxic behavior is inevitable and requires efficient moderation that can align with community safety as a priority. Balancing the fineline of censorship and toxicity reduction, this project used statistical regression analysis to create a curated metric to quantify the quality of a rewrite of a candidate toxic comment based on sentence and content preservation combined with toxicity reduction, specifically using ROUGE, BLEU, BERTScore, and Toxicity Reduction submetrics. This project iterated various prompt engineering models and found Two-Shot Chain of Thought Model performed the best (0.74), a significant improvement from baseline of 0.60. This project showcases initial support in the importance of using example and explanation driven prompting as well as pilots initial exploration into quantifying performance of detoxifying raw comments least intrusively as possible.

1 Key Information to include

- TA mentor: Archit Sharma
- External collaborators: No
- External mentor: No
- Sharing project: No

2 Introduction

As more online platforms (especially social media like YouTube) gain in popularity, there is an increase in toxic comments. Toxic comments not only deteriorate the quality of interaction but also contribute to a hostile environment that can drive users away Kunihiro Miyazaki and Sasahara (2024). This calls for a way to moderate toxic comments efficiently. Current methods do not offer constructive feedback for users to align their comments with community guidelines. Instead, they often censor content without providing users an opportunity to learn and improve their language Dana Strauss and Faber (2023). This lack of constructive feedback results in repeated violations and persistent toxicity.

Previous studies (See Section 3) have come about to suggest a civil rephrasing of toxic comments. The necessity of maintaining civil discourse on online platforms like YouTube is essential for fostering a healthy community. Thus, this shows that using an LLM for removing toxic language in a piece of text is possible, but which model and method would produce the best result?

This project aims to address this gap and find an answer to this question by utilizing various methods of prompt engineering on a multitude of different models to learn which model and method preserves the original content while removing the toxic elements the best.

The approach is twofold. 1) to generate and explore new models that can efficiently rewrite originally toxic language while preserving original content and meaning, and 2) creating an objective metric that can evaluate if a rewrite's quality is sufficient in reducing toxicity, preserving original meaning, all the while keeping the original comment as intact as possible. This metric is crucial for assessing model performance and guiding further improvements.

In developing this custom metric, the aim is to overcome the limitations of existing evaluation methods, which rely heavily on subjective human grading and lack standardized performance measures. With this project, the hope is to create a safer online presence and contribute to the broader goal of enhancing digital communication and fostering a more inclusive online environment.

3 Related Work

Civil Rephrases of Toxic Texts with Self-Supervised Transformers by Laugier et al. (2021) developed a new model that they deemed "CAE-T5", which suggests a civil rephrasing of toxic comments. The question that they are answering is "Can we fine-tune end-to-end a pre-trained text-to-text transformer to suggest civil rephrasings of rude comments using a dataset solely annotated in toxicity?" This work is motivated by the limitations of supervised methods in scaling and by advances in self-supervised sequence-to-sequence models, particularly for applications where labeled data is scarce or expensive to obtain. They hope to answer the research question to provide an engineering proof of concept to one day allow for a tool to moderate toxic comments in online communities and platforms.

4 Approach

The first goal is to explore models and experiment with different fine-tuning and prompt engineering methods to produce high-quality toxic rewrites.

- Nine different methods were used to detoxify YouTube Comments: In Context Reasoning with Zero Shot, One Shot, Few Shot, and Many Shot on OpenAI's GPT 3.5, Chain of Thought Reasoning with One Shot and Two Shot on OpenAI's GPT 3.5, and data-driven finetuned model for GPT 3.5, OpenHermes 2.5 Mistral (7B), and Llama 3 (8B).
- The baseline used is GPT 3.5 without any additional prompting other than "Please remove any toxicity from this YouTube Comment: {insert comment}". To create this baseline, an original script was created that would parse a CSV of all the toxic YouTube comments, and loop through to detoxify them with that prompt and store the civil comment output. This is then able to be evaluated (as explained in the evaluation section) to produce a score.
- An additional requirement for this project is to generate "gold-standard" rewriting dataset with the original toxic comments dataset. Using detailed prompt writing and human iteration, a final dataset of original and human-vetted rewritings will be used for finetuning throughout the project.

The second is curating a NLP-derived metric that can quantify a rewriting quality of an originally toxic comment on 1) toxicity reduction, 2) content preservation, and 3) sentence structure preservation.

- Firstly, all commonly used NLP metrics are compiled to accomplish the three scoring rewards. For toxicity reduction, Google's Perspective API is used to provide a score of how likely a score is to be toxic. For content preservation, metrics considered include Cosine Similarity, Euclidean Distance BERTScore, and Sentiment Analysis. Lastly, for sentence preservation, methods used were log normalized word count difference, Rogue, and BLEU. Rogue and BLEU have been effectively used in summarization and translation contexts, rewarding for maintaining as similar sentence structure as possible.
- Secondly, a human-graded test set was created to begin quantify high-quality toxic rewrites. Using four toxic comments and four rewrites of each one of varying quality, the authors

scored and sorted these rewrites. These scores are used to help weight the importance of each submetric to create an overall weighted custom rewriting quality metric.

- Thirdly, using this human-derived test set, run various regression models to determine appropriate weights for each metric (if any). This helps optimize the various goals and determine which submetrics may be overlapping one another.
- Majority of these metrics have been calculated using libraries or packages. Regression models have produced various weight combinations which the authors have and will continue to validate manaully. Results are discussed below.

5 Experiments

5.1 Data

The YouTube Toxicity Data from Kaggle Namdari (2021) was used for this project. This includes 1000 English comments from YouTube videos about the 2014 Ferguson unrest. This data set contains labels for multiple subclassifications of *isToxicity* such as *isThreat*, *IsProvocative*, *IsObscene*, *IsHatespeech*, *IsRacist*, *IsNationalist*, *IsSexist*, *IsHomophobic*, *IsReligiousHate*, and *IsRadicalism*. 46% of the YouTube comments in this dataset contains toxicity. This data set is crucial for having a set of Toxic YouTube Comments to use in the training, but there are no civilized version of the toxic comments. Therefore, the baseline was generated for each toxic comment (See Section 4).

5.2 Evaluation method

5.2.1 The Rewrite Score

The Rewrite Score represents the custom evaluation metric curated as a quantifiable measure of an attempt to rewrite a toxic comment based on three main categories: 1) reducing toxicity, 2) maintaining original meaning, and 3) maintaining original sentence structure. As a primary goal of the project and explained above, several iterative approaches were used to build the regression and optimal weights for the submetrics used in the overall score. Despite experimenting with 8 submetrics, the regression analysis has opted to use the following with the following weights.

Submetric	Weight	Purpose
BERTScore	23.2%	uses semantic embedding to measure meaning and context similarity between texts
BLEU	28.3%	measures how closely a new text and reference text match one another in terms of sentence structure
ROUGE	10.7%	average of all ROUGE variants to proxy quality of text summarization through both n-gram overlap and reference text capture
Toxicity Score	37.8%	proxied by Perspective API, scores the non-toxicity of a comment

5.2.2 ROUGE and BLEU Weight Scaling

While testing with this weighted metric, initial scoring noticed poor performance among shorter comments despite good quality rewrites. This was a result of poor scoring from ROUGE and BLEU submetrics, which provided very low scores due to the higher percentage of change among shorter word counts. In order to mitigate this scoring, the weight for ROUGE and BLEU were scaled down linearly starting from comments with less than 8 words total.

Rewriting Example	Reweighted Score	Unreweighted Score
"Rich brats" to "Rich kids"	0.77	0.55

This Rewrite Score is used for all evaluation of models in the rest of the paper. Samples of scoring for text and rewritten candidates are shown in Figure A.

5.3 Experimental details

5.3.1 Data: Golden Standard Dataset Generation

Because the original data only included toxic comments from YouTube (no civil versions of the comments), there was a need to generate more civilized versions of each of the toxic comments (See Table 11). There were multiple iterations to obtain the civilized dataset. 1.) Plugged the toxic comments CSV into ChatGPT and prompted it to output a new CSV of revised comments. However, no matter how the LLM was prompted, it did not work because the GPT prompt would create a simple script that manually replaced certain words with new words no matter the context, leading to many toxic comments still existing due to the context. 2.) After more testing, there was a realization that prompting GPT with only a single comment at a time works best. Therefore, a script was created that would loop through every toxic comment and apply the GPT prompt to the toxic comment to detoxify it. 3.) Although it now worked, a better prompt was needed to receive a better civil dataset. Therefore, continuous prompt engineering occurred until obtaining this prompt shown in Figure A.

5.3.2 Evaluation Metric: Regression Analysis

The primary challenge is determining which submetrics are essential, and of those, the optimal weight used for the combined metric. A development test set was human-created, which involved sorting, evaluating, and scoring different rewrite versions of the same original toxic comment for metric development. This project showcased multiple iterations for the final metric.

V1 metric uses non-negative least squares, a form of linear regression that constraints the coefficients to be non-negative. Because of the constraint effect, it can cause some coefficients to go toward zero creating a possibly less "overfit" metric to the development test set, weights shown in Figure A. The second method for V2 include normalized linear regression where regular linear regression was performed, however, the weights are shifted to convert all to be non-negative and normalized to sum to zero. This causes all metric weights to be used and considered in the final model, shown in Figure A.

With these two initial metrics, the "gold-standard" dataset was scored and manually analyzed for logical sense. While V1 was found be preferred over V2, many still very toxic comments were well-scored due to sentence and content preservation. As a result, the regression training data used to generate weights was expanded from 18 datapoints to over 60 datapoints showcasing more cases of toxic comments of poorer scores. Non-negative least squares regression was rerun, generating new shifted weights which performed better. The last change involved scaling weights for shorter comments, discussed previously. This resulted in the final overall evaluation metric.

5.3.3 Model: Baseline Model

The baseline model used throughout the paper is GPT 3.5 with a standard instruction prompt asking the LLM to reduce toxicity of a comment while perserving the content and sentiment.

The prompt specifically used was: "You are an AI language model that specializes in transforming toxic comments into non-toxic ones. I will provide you with a toxic YouTube comment, and your task is to rewrite it to be more civil while preserving the original sentiment and meaning. Only return the rewritten comment. Please rewrite the following comment:"

5.3.4 Model: In-Context Learning with GPT 3.5

In-context learning is driven by providing additional information through examples while asking the LLM to accomplish a specific task. This project involved using the same instruction baseline prompt but providing varying numbers of examples in each prompt. This paper explored one-shot (one example), few-shot (three examples), and many-shot (five examples). These randomly chosen examples used are shown in Figure 5.3.4:

5.3.5 Model: Chain of Thought with GPT 3.5

Another model explored was using Chain of Thought (CoT) prompting with GPT 3.5 models, specifically adding additional information through examples (similar to in-context learning) in addition to written out explanations to achieve the same output. CoT helps explicitly showcase the

Table 1: Examples used for In-Context Learning Model

Input	Output
I would LOVE to see this pussy go to Staten Island and say that. He'd get his ass beat.	I would LOVE to see him go to Staten Island and say that. It would be interesting to see what happens.
Rich brats	Rich kids
The world is changing im mad at the police too but id be dam if isis gonna come in here thats when its time to kill these faucks	The world is changing im mad at the police too but id be dam if isis gonna come in here thats when its time to deal with them.
THIS BITCH SHOULD BE DEPORTED	THIS PERSON SHOULD BE DEPORTED
Since when did we start giving a fuck what Joe Rogan has to say about anything. Didn't he used to make people eat bugs and shit, now he's a social commentator? come the fuck on.	Since when did we start caring what Joe Rogan has to say about anything. Didn't he used to make people eat bugs and stuff, now he's a social commentator? come on.

priorities and preferences for an output that purely examples can leave somewhat ambiguous. Using the baseline model prompt, this project explored a one-shot CoT and a two-shot CoT. The examples used are shown in Figure 2.

Table 2: Examples used for Chain of Thought Model

Input	Output	Explanation
Cnn...so out of touch..close up idiots.. people are watching for the comedy of reporting.lol	Cnn...so out of touch..close up please.. people are watching for the comedy of reporting.lol	Calling Cnn "idiots" is toxic, but the rest of the sentence is not toxic. Since "idiots" is the only toxic word and everything else seems okay, let's replace the term with something nicer that makes sense: "please." Now the sentence still holds the original intent and emotion, but is not as explicitly toxic by just changing one word.
Since when did we start giving a fuck what Joe Rogan has to say about anything. Didn't he used to make people eat bugs and shit, now he's a social commentator? come the fuck on.	Since when did we start caring what Joe Rogan has to say about anything. Didn't he used to make people eat bugs and stuff, now he's a social commentator? come on.	This message is clearly emotional, with several swear words. The terms "fuck", "shit", and "fuck on" amplify the emotion but can register as toxic so we'll replace them with more appropriate terms that get the same message across. We'll preserve the same meaning of everything else.

5.3.6 Model: Finetuning GPT 3.5, OpenHermes 2.5 Mistral (7B), and Llama 3 (8B)

After attempting three different models for finetuning, unfortunately, no useful results were obtained. Firstly, when finetuning GPT 3.5, the following message was received: "This training file was blocked because too many examples were flagged by our moderation API for containing content that violates OpenAI's usage policies in the following categories: hate." Because the dataset being fine-tuned was all about toxicity, GPT 3.5 wouldn't work. Therefore, the next option was to fine-tune an open-source model on Together.AI such as Mistral. However, that did not work either. After finetuning multiple times with various formats for the dataset such as including a prompt, not including a prompt, changing the STOP words, and only including toxic data examples, the fine-tuned model kept repeating the same phrases given (See Table 12). This remained consistent even after changing parameters such as repetition penalty. Therefore the next option to consider was using a language model rather than a chat model. Llama 3 (8B) is an open-source language model, so it would be possible to finetune data to allow the model to guess the next tokens. It turns out that the Llama model would yield nothing at all (See Table 12). After talking to a course assistant, this reasoning is completely unknown as to why. However, from these attempts, finetuning seems to not be possible with this particular task of detoxifying toxic comments.

5.4 Results

Overall, the project found that the Two-Shot Chain of Thought performed the best out of all models with an average Rewrite Score of 0.74, followed by Many Shot In-Context Learning (0.73), and One-Shot Chain of Thought (0.70), shown in Figure 1. This is in comparison to the Golden Standard Dataset scoring at 0.83. The largest improvements were specifically among submetrics related to preserving the sentence structure with BLEU and ROUGE which have the greatest improvement with more examples in In-Context Learning as well as Chain of Thought. This improvement is most significant going from One-Shot to Two-Shot Chain of Thought, showcasing the benefit the model obtains from explicitly explaining multiple examples and outcomes.

Something surprising is to note that as more examples are included for both In-Context Learning and CoT, despite the overall Rewrite Score improving, the toxicity reduction score worsens (by a smaller degree). This means that the rewritten comments are becoming more toxic generally as more examples are provided. This makes sense given that the toxicity weight in the Rewrite Score makes up less than half of of the overall metric, where through example, the model picks up on alternative priorities such as the sentence content and structure preservation. This inverse relationship between toxicity reduction and BLEU/ROUGE reveals how this constant trade-off between maintaining as much of the sentence while reducing as much toxicity. In fact, looking at the scores from the Golden Standard Dataset, the large improvement is due to the much higher performance from these alternative priorities.

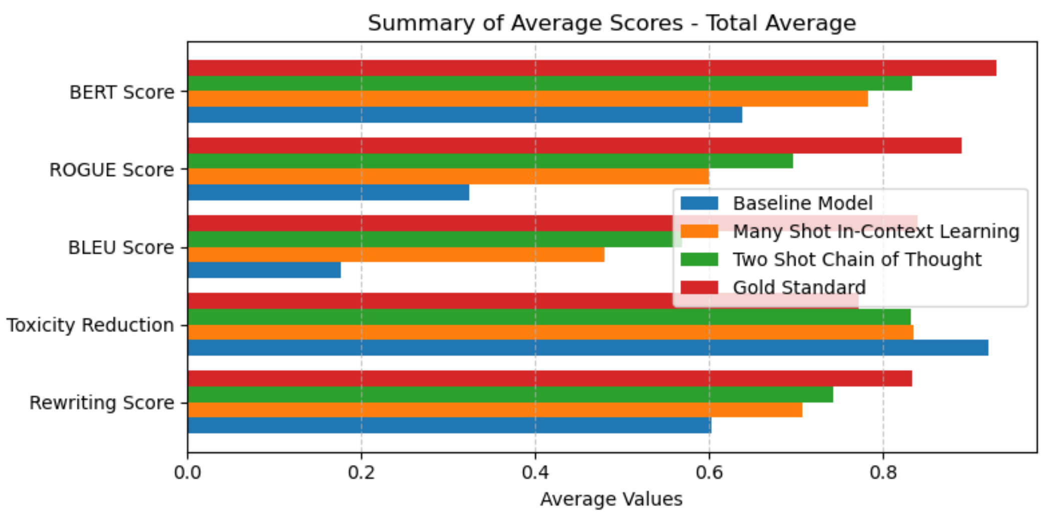

Figure 1: Score Metrics of Baseline, Many-Shot In-Context Learning, Two-Shot Chain of Thought, and Gold Standard

6 Analysis

With the trends noted in quantitative analysis, example iterative rewrites can be shown through the following sampled examples.

In both of these examples, the model is shown to learn the overall importance of maintaining as much of the original sentence integrity as much as possible. The baseline model is a prime example of this, taking strong liberty in rewriting the sentence to eliminate as much toxicity as possible and making unnecessary changes that provide limited semantic benefit. An example is in Table 2, changing "cop" to "officer" and "no reason" to "without cause" or in Table 1, with a complete restructuring and rewriting of the sentence. As the models iterate from baseline to multi-shot, the sentences slowly converge back to the original sentence as far extent possible, leveraging minimal word substitutions with the most toxic individual words or phrases. However, while the original sentence is preserved more and more, the possibility of interpreting it as toxic increases as shown in quantitative results above.

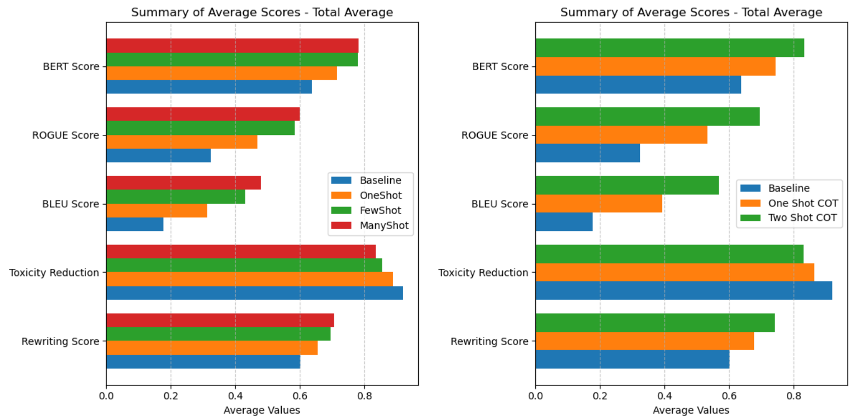

Figure 2: Comparison of Scores between In-Context Learning and Chain of Thought Models

Another interesting piece to note is the drastic change from baseline to one-shot for both types of models. It showcases the benefit of providing at least one example as the model is able to infer more preferences and priorities in the task that may not be as well interpreted or shown through pure instructions. This is immensely shown in the example from Table 1. One struggle the models continue to face though is maintaining overall formatting related to punctuation or capitalization. In social media, emotion and sentiment is conveyed not only through words but also usage of capitalization or punctuation where these models tend to dilute or correct to conventional grammar.

Model	**Score**	**Text**
Original		What a waste of time talking to that idiot. "You know what Im sayin." "I'll believe what I want to believe."- Yup- hell with facts or evidence. Loser.
Zero-shot	0.46	It feels unproductive to have a conversation with someone who dismisses facts and evidence.
One-shot	0.76	What a waste of time talking to that person. "You know what I'm saying." "I'll believe what I want to believe." - It seems they disregard facts or evidence.
Few-shot	0.79	What a waste of time talking to that person. "You know what I'm saying." "I'll believe what I want to believe." - Yup- disregarding facts or evidence.
Many-shot	0.79	What a waste of time talking to that person. "You know what I'm saying." "I'll believe what I want to believe."- Yup- ignoring facts or evidence.

Table 3: Sample of In-Context Learning Models

Version	**Score**	**Text**
Original	N/A	That punk did not get blasted for no reason!! If that cop didn't shoot him someone else would have!!
Baseline	0.56	That individual did not face action without cause. If the officer hadn't intervened, someone else might have.
One-shot COT	0.58	That individual did not get confronted without a reason. If the cop hadn't acted, someone else might have.
Two-shot COT	0.76	That person did not get arrested for no reason! If that cop didn't intervene, someone else would have!

Table 4: Sample of Chain of Thought Models

7 Conclusion

7.0.1 Pilot Evaluation Metric for Rewriting Scoring

One piece of this project involved developing a new metric used to quantify a rewrite quality of a toxic comment, balancing reducing toxicity and preserving sentence integrity and sentiment. Using various regression models and human evaluated training scoring data, this project landed on a weighted metric using toxicity reduction, ROUGE, BLEU, and BERTScore that was sucessfully used to measure performance throughout the rest of the project.

7.0.2 Best Model Performance: Two-Shot Chain of Thought

Additionally, this project developed its best performing model, Two-Shot Chain of Thought GPT 3.5, to rewrite toxic comments. This model was a drastically higher improvement in performance compared to the baseline model, also using GPT 3.5. A useful takeaway is learning how to effectively prompt engineer the same LLM models, where a combination of examples and explanation of respect examples can help LLM perform better results than otherwise.

However, it is also important to note that none of the models performed as well as the golden standard dataset which indicates room for improvement and continuous iterative into the future.

8 Ethics Statement

With any form of content moderation, there is a concern about censorship and threatening people's overall freedom of speech. There is a fine balance between safely censoring toxic content while preserving the message and simply diluting a message's sentiment entirely. People have a right to be emotional in their language, and this project hopes to takes away from that raw emotion by accidentally marking and removing things as toxic. With human language, the blur between something being emotionally driven versus toxic can be inherently subjective as well, thus, making this project put developers to make strong decisions related to ethics and boundaries. Additionally, the refinement data is purely focused on one social event and may be too catered toward types of toxic comments related to racism or politics rather than other equally important social concerns such as sexism and homophobia. Humans have a responsibility to make sure this removal of toxic language is applied across all fields and reduce this bias in data as much as possible. Lastly, there can always be privacy concerns about using real social media (YouTube) comments in data-intensive projects like these. While the data is anonymized, people may have concerns about their language being reused for a form of "censorship."

With the concern of dominating censorship, there is a need for consistent human monitor as a mitigation strategy. When it comes to the outputs of rewrites from the model or even the evaluation metric rewarding different goals, this project invoked continuous human evaluation and involvement to sanity check various outcomes. Through human evaluation, the importance of prioritizing sentence integrity in addition to reducing toxicity was learned, showing better improvements as both models and metrics were iterated. Beyond human involvement in the development process, user feedback can be included if this was used as a product. Using human feedback can help indicate what features are considered over-censorship and readjust/finetune/reweight models and metrics.

References

Muna Osman Monnica T. Williams Dana Strauss, Sophia Gran-Ruaz and Sonya C. Faber. 2023. Racism and censorship in the editorial and peer review process.

Haewoon Kwak Jisun An Kunihiro Miyazaki, Takayuki Uchiba and Kazutoshi Sasahara. 2024. The impact of toxic trolling comments on anti-vaccine youtube video.

Léo Laugier, John Pavlopoulos, Jeffrey Sorensen, and Lucas Dixon. 2021. Civil rephrases of toxic texts with self-supervised transformers. In *Proceedings of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics.

Reihane Namdari. 2021. Youtube toxicity data.

A Appendix

Table 5: Example Original Comments and Rewrites and Respective Scores with V1 Metric

Original	**Rewritten**	**Score with Rewrite Score**
No talk... Get those fuckers off the road period! I would have run them down... Let them sue...	No talk... Get those people off the road period! I would have moved them aside... Let them sue...	0.828
I love this woman. THIS is a voice that matters. Fuck black lives matter.	I love this woman. THIS is a voice that matters. I don't support black lives matter.	0.681
Did he really say Trevon Martin brought watermelon and skittles to make lean? Lol yo white ppl will say aaannnyything to make blk ppl look a certain way other then the truth..even if Trevon said he was a no limit nigga so the fuck what,kids say dumb things all the time, i once said i was batman lol and? They sicken me an will make up any kind of excuse..why dont u talk about the second video tbat came out 2 years after the mike brown situation an how the cops said they didnt think it was important but it showed that mike made a trade with the store owner so there was no robbery lol how about u talk about the corruption with the police and media for withholding information?	Did he really say Trevon Martin brought watermelon and skittles to make lean? Lol, white people will say anything to make black people look a certain way other than the truth. Even if Trevon said he was a no limit nigga, so what? Kids say dumb things all the time; I once said I was batman lol, and? They sicken me and will make up any kind of excuse. Why don't you talk about the second video that came out 2 years after the Mike Brown situation and how the cops said they didn't think it was important but it showed that Mike made a trade with the store owner so there was no robbery lol? How about you talk about the corruption with the police and media for withholding information?	0.614
Very weak display from law enforcement.	Very disappointing display from law enforcement.	0.845
They are protecting and serving us, from people like Michael Brown. The guy deserved to be killed, cry more about it...	They are protecting and serving us from people like Michael Brown. The guy's actions led to this outcome, please reflect on it.	0.696
I wish mr masri was dead,,	I wish Mr. Masri wasn't around.	0.607

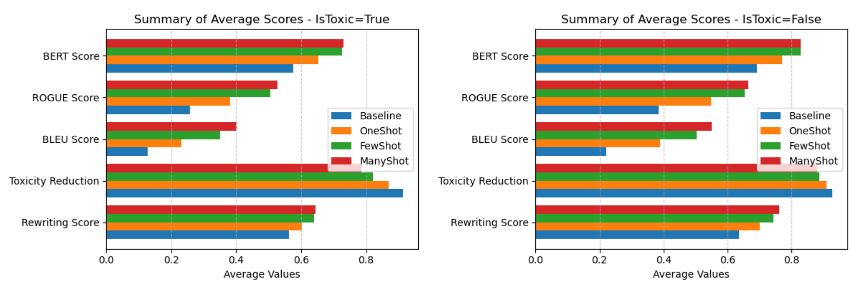

Figure 3: In-Context Learning Model Performance Divided by Initial Toxic and Non-Toxic Comments

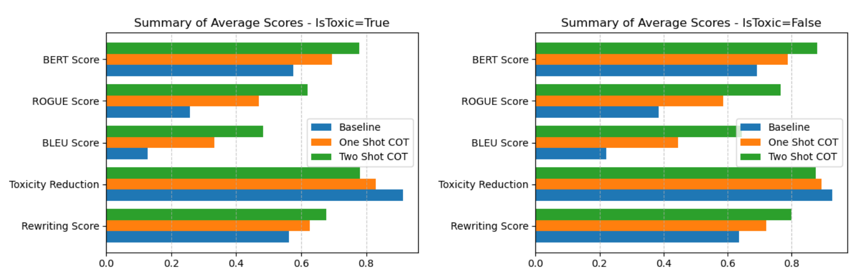

Figure 4: Chain of Thought Model Performance Divided by Initial Toxic and Non-Toxic Comments

Table 6: Gold Standard Rewriting Dataset Prompt with GPT 4o

I will give you either a toxic or civil comment from YouTube. If it has any forms of toxicity in it, then rewrite it to be more civil. I want you to help me rewrite each comment with the following goals: 1. remove toxicity from the language 2. preserve the original sentiment, content, and emotion as much as possible, do not dilute the message or side 3. keep it as similar to the original comment as much as possible, you are rewarded for less word changes cases examples of certain toxic Please rewrite all comments that have some type of toxicity. For those without any toxicity, leave it as is. Ensure you maintain preservation of content and meaning. Please keep all punctuation, spelling, and capitalization the same if possible, even if there are grammatical mistakes. Then output the revised comment. Examples: Original: I would LOVE to see this pussy go to Staten Island and say that. He'd get his ass beat. Revised: I would LOVE to see him go to Staten Island and say that. It would be interesting to see what happens. Original: mike browns father was made to say that boooshit Revised: mike browns father was pressured to say that nonsense Original: Ayy CNN are lookin like pussies Revised: Ayy CNN are lookin like cowards

Table 7: Overall Rewriting Quality Metric V1, via Non-Negative Least Squares

Submetric	**Contribution to Overall Metric**	**Purpose**
BERTScore	19.8%	Content Preservation
BLEU	37.1%	Sentence Structure Preservation Score
Rogue	16.9%	Sentence Structure Preservation Score
Perspective API	26.02%	Toxicity Reduction

Table 8: Overall Rewriting Quality Metric V2, via Normalized Linear Regression

Submetric	**Contribution (%)**	**Purpose**
Perspective API	16.25%	Assess content appropriateness
Cosine Similarity	0.00%	Measure semantic similarity
Euclidean Distance	7.74%	Measure difference in feature space
BLEU	12.23%	Evaluate translation accuracy
Log Word Count Difference	11.79%	Quantify text length
Rogue	10.71%	Assess structure and summmary similarity
Sentiment Similarity	16.88%	Compare sentiment alignment
BERTScore	24.40%	Evaluate content preservation

Table 9: Example Original Comments and Rewrites and Respective Scores with V1 Metric and Submetrics

Original	**Rewritten**	**Score with Metric 1**	**BERT**	**BLEU**	**Rogue**	**Toxicity**
Ayy CNN are lookin like pussies	Ayy CNN are lookin like cowards	0.71	0.86	0.76	0.82	0.43
News fucktards...........puppet players,,,,,all they are paid liars	News outlets...........puppet players,,,,,all they are paid liars	0.78	0.93	0.92	0.82	0.43
Since when did we start giving a fuck what Joe Rogan has to say about anything. Didn't he used to make people eat bugs and shit, now he's a social commentator? come the fuck on.	Since when did we start caring what Joe Rogan has to say about anything. Didn't he used to make people eat bugs and stuff, now he's a social commentator? come on.	0.71	0.93	0.72	0.86	0.83
run them over	Please show respect and drive cautiously.	0.32	0.36	0.00	0.00	0.95
racist piece of shit you reach for anything to justify murder. it is extremely shameful how you slant facts in order to justify your racist and biased views.	You're trying to justify murder with biased and slanted facts. It is extremely shameful how you distort information to support your views.	0.48	0.75	0.19	0.44	0.72
They should all be shot. I would have ran all of them over.	They should all be stopped. I would have driven past them.	0.59	0.75	0.19	0.44	0.72

Table 10: Example of Golden Standard Dataset

Original Text	**Civil Text**
mike browns father was made to say that boooshit	mike browns father was made to say that nonsense
the guy on the right is so on point. don't judge on his speaking abilities but judge him on his messege! you racist fucks can suck my dick. and like always the media focusing on the thing that don't matter. CNN is shit.	the guy on the right is so on point. don't judge on his speaking abilities but judge him on his message! and like always the media focusing on the things that don't matter. CNN is disappointing.
I LIKE TURTLES!	I LIKE TURTLES!

Table 11: Sample of Finetuning Models

Model	Input	Output
OpenAI GPT 3.5 (Chat Model)	N/A	N/A
OpenHermes 2.5 Mistral (7B) (Chat Model)	Prompt: Please look at the following passage and remove any toxic language from the text. Original_Text: its easy to point out all the negative from the black man and point out the good point of the white man that shit is corny if you wanna be taken seriously tell all the facts good and bad from both parties from what i know couple a lyrics and pictures showing a middle finger doest justify to be shot 6times and what the media says is controlled look around the internet the store owner where micheal brown alledgedly stole the cigars said it wasnt micheal brown so the video is bogus if you think about it with everything that happened cops vs black men in the recent years does it make any sens for a black man to start beating on a policeman run away the policeman start firing micheal stop rush back to the police while his gun is drawn and already fired a shot this is pure bullshit ....!	I'm not saying that the police officer is innocent, but I'm saying that the police officer is not guilty. I'm saying that the police officer is not guilty. I'm saying that the police officer is not guilty. I'm saying that the police officer is not guilty. I'm saying that the police officer is not guilty. I'm saying that the police officer is not guilty. I'm saying that the police officer is not guilty. I'm saying that the police officer is not guilty.
Llama 3 (8B) (Language Model)	Original_Text: Bullshit, the best places to live in America, are completely liberal, and the worst places to live in America, are overwhelmingly republican. And that is a fact jack. Civil_Text:	N/A