Re-ranking Using Large Language Models for Mitigating Exposure to Harmful Content on Social Media Platforms

Rajvardhan Oak[1,3], Muhammad Haroon[1], Wonjeong Jo[1],
Magdalena Wojcieszak[1], Anshuman Chhabra[2]
[1]University of California, Davis, [2]University of South Florida, [3]Microsoft Corporation, USA
{rvoak, mharoon, wjo, mwojcieszak}@ucdavis.edu, anshumanc@usf.edu

Abstract

Social media platforms utilize Machine Learning (ML) and Artificial Intelligence (AI) powered recommendation algorithms to maximize user engagement, which can result in inadvertent exposure to harmful content. Current moderation efforts, reliant on classifiers trained with extensive human-annotated data, struggle with scalability and adapting to new forms of harm. To address these challenges, we propose a novel re-ranking approach using Large Language Models (LLMs) in zero-shot and few-shot settings. Our method dynamically assesses and re-ranks content sequences, effectively mitigating harmful content exposure without requiring extensive labeled data. Alongside traditional ranking metrics, we also introduce two new metrics to evaluate the effectiveness of re-ranking in reducing exposure to harmful content. Through experiments on three datasets, three models and across three configurations, we demonstrate that our LLM-based approach significantly outperforms existing proprietary moderation approaches, offering a scalable and adaptable solution for harm mitigation.

1 Introduction

Social media platforms are powered by Machine Learning (ML) and Artificial Intelligence (AI) based recommendation algorithms and models that provide content for users. These algorithms and models are designed to maximize user engagement by learning to recommend content aligned with users' inferred interests or traits (Covington et al., 2016). However, solely optimizing for user engagement metrics can indirectly drive exposure to harmful content. For instance, a teenager interested in fitness may be recommended content promoting eating disorders, users interested in finance may encounter clickbait and scam videos, and a sad adolescent may be directed to content about depression or suicide.As a result, there are serious concerns that platform recommendation systems can indirectly foment misinformation, addictions or mental health crises, and lead to other problems for individuals and society at large (Haidt and Twenge, 2022, 2023; Roose, 2019; Tufekci, 2018).

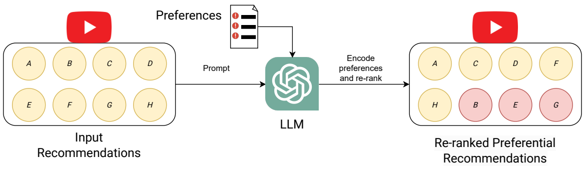

Figure 1: An overview of our re-ranking approach for mitigating exposure to harmful content. We prompt the LLM with the input set of recommendations and a set of preference constraints. The LLM re-ranks the the recommendations in accordance with the provided preferences. Here, recommendations *B*, *E*, and *G* are harmful and hence, downranked.

Internally, social media platforms seek to mitigate harmful content using AI/ML classifiers. However, there are two major challenges associated with their use (Gorwa et al., 2020; Chen, 2021; Khan and Wright, 2021): (1) classifiers often require large volumes of annotated data for training, and (2) categorizing harmful content is a dynamic temporal problem (e.g. a new dangerous challenge for teenagers emerges online). Classifiers cannot automatically generalize to new forms of harm, without having been trained on explicitly labeled data. As a result, harm classifiers are susceptible to concept drift (Quiñonero-Candela et al., 2022) and requiring humans to annotate large amounts of data.

In this paper, we propose methods that can circumvent both issues: that of scale as well as that of the dynamicity of harm. We utilize Large Language Models (LLMs) for this purpose as they have been shown to demonstrate stellar reasoning capabilities on natural language input in the zero-shot or few-shot learning setting (Wei et al., 2022) making them ideal for harm mitigation. We find that, as opposed to traditional harm classifiers such as

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 894–908
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

Perspective API[1], LLMs can excel at mitigating exposure to harm in recommended content sequences simply by pairwise comparisons and re-ranking, most likely due to better utilization of their context window. The benefits of pairwise re-ranking have been explored in past work (Qin et al., 2024) and we adopt this re-ranking strategy using LLMs to mitigate exposure to harmful content on platforms. This approach is visually depicted in Figure 1.

We propose two novel metrics to dynamically analyze exposure to harmful content in recommendation sequences and experiment with three LLM preference re-ranking methods (zero-shot, few-shot, in-context learning based (Dong et al., 2022)). Through extensive experiments on three harm datasets and three LLM architectures, we find that our approach outperforms state-of-the-art content moderation baselines, such as Perspective API and OpenAI's Moderation API.[2] In summary, we make the following contributions:

- We propose a novel LLM-based approach that can measure relative harm and mitigate harmful content exposure on platforms. Our approach is able to generalize to various kinds of harms without explicit training.

- We examine the performance of our approach under three settings (zero-shot, zero-shot with prompt engineering and few-shot) by varying the amount of information provided to the LLM, and find that it outperforms industry-grade classifiers even in the zero-shot setting.

- Alongside evaluation on traditional metrics, we propose two novel metrics that can measure harm exposure for a given content sequence. Our metrics allow evaluating the quality of re-ranked content in a manner that is agnostic to the amount of harm present wherein.

This paper is organized as follows. First, we discuss related work in the areas of harms caused by recommendation systems and efforts towards mitigation to contextualize the novelty of our approach. Then, we present our methodology at a high level and introduce two novel metrics for evaluating re-ranking. Next, we describe our experimental setup and analyze our results. Finally, we discuss the limitations of our work and its implications.

2 Related Work

2.1 Recommendation-Driven Harms

Platform recommendation systems are designed to maximize user engagement by tailoring content to individual preferences and interests. However, these systems can inadvertently drive users toward harmful content. This occurs because recommendation algorithms often prioritize content that generates the most interaction, which can include sensational or provocative material (Rathje et al., 2021; Yu et al., 2024). For instance, Facebook's algorithmic curation is linked to the widespread use of clickbait, which drives users toward low-quality content (Lischka and Garz, 2021).

In addition, digital traces left by each user on platforms reveal information about the user's emotions (Hossain and Muhammad, 2019), substance use (Kosinski et al., 2013), or sexual orientation (Wang and Kosinski, 2018). This inferred information could be used to recommend content that exposes users—especially the vulnerable ones—to online harms (e.g, addictive content to users known to use substances, suicidal content to depressed users, misinformation to users interested in herbology) (WSJ Staff, 2021). Indeed, meta-reviews show that 8%-10% of recommendations pose detectable risks to users (Hilbert et al., 2024) and algorithmic audits detect discriminatory or otherwise harmful biases in YouTube, Instagram, and TikTok algorithms (Haroon et al., 2023; Bandy, 2021; Hilbert et al., 2023).

Because recommendation systems are exceptionally good at curating personalized ecosystems, leading to closed loops of content consumption (Rossi et al., 2021), what they recommend has implications for what users see (Nyhan et al., 2023). In turn, content exposure has over-time effects on individuals and, when harmful, can have severe consequences on mental health and may foment addictions, violence, or even lead to death from dangerous challenges (Haidt and Twenge, 2022).

2.2 Harm Mitigation

Extensive work studies harm mitigation interventions on social media platforms. For example, Bhargava et al. (2019) developed a tool that allows users to exert more control over their social media feeds by enabling them to consolidate and tailor content from multiple platforms. Similarly, Kovacs et al. (2018) empower users to manage their social media engagement goals via rotating time or site

[1] https://www.perspectiveapi.com/.

[2] https://platform.openai.com/docs/guides/moderation.

blockers. More germane to minimizing harmful content, some studies explore ranking-based interventions (Celis et al., 2019; Ovadya and Thorburn, 2023), such as using LLMs for curating societally beneficial recommendations. In addition, machine learning models, particularly those leveraging deep learning, are often adopted to identify patterns indicative of harmful content, such as clickbait (Ghosh et al., 2022), hate speech (Mossie and Wang, 2020; Del Vigna12 et al., 2017) and misinformation (Wu et al., 2019; Shu et al., 2017). Recent work additionally applies LLMs to the harm detection problem (Liu et al., 2024; Bonagiri et al., 2025; Ernst, 2024) and also utilizes machine learning toxicity detection algorithms within a browser extension to automatically hide toxic text content on users' feed over-time (Beknazar-Yuzbashev et al., 2022).

Remark. While prior works focus on specific harms, we consider a general, systematic, and overarching taxonomy of harms with data originating from a real social media platform (YouTube), as detailed below. Our approach generalizes to various kinds of harms, as opposed to being effective on one particular kind, like hate, misinformation, or clickbait. We leverage LLMs and their inherent reasoning abilities to reduce harmful content exposure in recommended content. Additionally, because we do not require explicit labeling and training, our approach is robust against *concept drift* (Quiñonero-Candela et al., 2022). Finally, instead of absolute harm, we focus on measuring relative harm and propose novel metrics to quantify the ordering of content for minimal exposure to harm.

3 Mitigating Harmful Content Exposure Using LLMs and Re-ranking

3.1 Problem Formulation

The problem we seek to solve is as follows:

> *Given a recommendation sequence (e.g. homepage videos on YouTube), re-rank the content so that harmful content appears at the end of the sequence in a zero-shot or few-shot setting (i.e. limited annotations are required).*

The motivation for downranking comes from past work that has shown that users are less likely to interact with (and be exposed to) content that appears at lower ranks in recommendations (Yu et al., 2023; Glick et al., 2014). Downranking, as opposed to outright suppressing or deletion, minimizes exposure to harmful content while preserving transparency, diverse viewpoints and preferences, and freedom of expression. Formally, let $X = \{x_i\}_{i=1}^{n}$ be a sequence of n content instances, out of which p are non-harmful, and the remaining $n - p$ are harmful. Let $\rho\colon X \to \{0, 1\}$ be a binary decision function that maps every x to a label based on whether it is harmful or not. Our goal is to use an LLM $\mathcal{L}$ to transform X into another sequence X^* such that it minimizes exposure to harmful content.

3.2 Preferential Pairwise Ranking

We present our proposed solution to the harm mitigation problem as Algorithm 1. Our proposed approach consists of a *pairwise ranking component* that seeks to downrank content if deemed harmful by the LLM $\mathcal{L}$. Specifically, we present $\mathcal{L}$ with pairs of content instances and query it to determine which is the relatively harmful one. Certain preference constraints ($\mathcal{C}$) determine the exact prompt used to query the LLM. To re-rank content based on relative harm, we adopt the approach from Qin et al. (2024) but modify their methodology and scoring function. For a given content instance x, their scoring function increments the score by 1 for every content instance deemed less *relevant* than x, and by 0.5 for all other content instances. However, applying the scoring function as is will also compare non-harmful content with one another and incorrectly result in scores being higher. This can then lead to them being unfairly downranked. To address this challenge, in the pairwise ranking process, we allow the LLM $\mathcal{L}$ to decide if both instances are *non-harmful*, in which case there is no increment to the score. Given a content sequence X, we enumerate all possible pairs and compute the score for every content instance. We then reorder the content based on this score obtained.

3.3 Specifying Preference Constraints $\mathcal{C}$

As described in Algorithm 1, the pairwise ranking via LLMs also requires natural language preference constraints $\mathcal{C}$ as part of the prompt to effectively downrank harmful content. We employ three approaches in specifying the preference constraints for the LLM in the context window/prompt. We mention the approaches here, and offer the detailed prompts in Appendix C.3.

Zero-Shot: In the *zero-shot* setting, $\mathcal{C}$ asks to identify which of the two provided content instances is harmful, without explicitly specifying a definition

Algorithm 1 Harm-Based Re-ranking Using LLMs

```
Input: Sequence $X = \{x_i\}_{i=1}^n$, LLM $\mathcal{L}$, Preference Constraints $\mathcal{C}$
Output: Re-ranked Sequence $X^*$
initialize $score[x_i] \leftarrow 0$ for each $x_i \in X$
for each pair $(x_i, x_j) \in X \times X$, $i \neq j$ do
    query $\mathcal{L}$ with $\mathcal{C}$ for pairwise preference: $(x_i, x_j)$ and $(x_j, x_i)$
    if $x_i$ harmful then
        $score[x_i] \leftarrow score[x_i] + 1$
    else if $x_j$ harmful then
        $score[x_j] \leftarrow score[x_j] + 1$
    else
        continue
    end if
end for
sort $X$ using $score$ (ascending) to obtain $X^*$
return sorted sequence $X^*$
```

of *harm*. In this approach, we utilize the LLM's inherent understanding of harm learnt during pretraining.

Zero-Shot + Prompt Engineering: We build upon the zero-shot approach by including a definition for harm in $\mathcal{C}$. We first define explicitly what we consider harmful and characteristics of harmful content, and then query $\mathcal{L}$ to identify which of the two content instances is harmful.

Few-Shot ICL: We now provide representative instances of harmful content in $\mathcal{C}$, and query $\mathcal{L}$ to reason which of the two content instances is harmful based on this information. This approach is known as In-Context Learning (ICL) (Dong et al., 2022; Brown et al., 2020; Askari et al., 2025b).

3.4 In-Context Learning (ICL)

In-Context Learning (ICL) (Dong et al., 2022) relies on exemplars that the model is exposed to in order to learn certain features or characteristics. Therefore, the performance of our re-ranking approach will greatly depend on the chosen exemplars. A naive approach of ICL would be to randomly sample from harmful content and provide these random exemplars to the model; however this may lead to bias towards or against certain kinds of harm. To address this, we curate a set of exemplars that is representative of the harm in our dataset. Inspired by a popular coverage-based BertScore ICL selection approach (Gupta et al., 2023), we make some modifications for undertaking ICL in the re-ranking setting. We first use a pretrained RoBERTA model to project text content into an embedding space. We cluster the harm samples, and then choose the most representative sample from each cluster (the one that is closest to the centroid of the cluster). These representative samples form the exemplars we provide to the model.

3.5 Evaluating Re-rankings

While we use existing metrics for evaluating rankings (Sebe et al., 2000) such as Precision@K (Shani and Gunawardana, 2011), we also design new metrics that focus on the *relevance* of ranked content to a query and help account for harmful content that users could be exposed to in the sequence. Thus, we propose two novel metrics to assess preference in re-ranked content sequences. Note that both metrics are bounded between $[0, 1]$ and higher values are better (less harmful content).

Per-Pref-k: The Per-Pref-k (PPk) metric represents the fraction of the content sequence set X that would need to be consumed to reach the k-th harmful content instance. PPk assesses how much of the sequence a user needs to consume before encountering a certain amount of harmful content. This metric is instrumental in understanding the depth of user engagement required to reach less desirable content, thus indirectly measuring the buffer of harmless content. Higher values indicate that a user can view more content before encountering a specified number of harmful instances, reflecting better performance of the moderation system. Formally, we define PPk as:

$$\text{PP}k = \frac{\min\{m \mid \sum_{i=1}^{m} \rho(x_i) = k\}}{n} \tag{1}$$

Exponentially Weighted Normalization: The Exponentially Weighted Normalization (EWN) metric provides an analytical measurement of non-harmful rankings by assigning exponentially decaying weights to ranks and then normalizing values to lie between $[0, 1]$. EWN(X) measures the goodness of the ranking in the sequence X by comparing it to the best and worst ranking possible. A value of 1 indicates that the sequence is in the optimal order and no better order can minimize exposure to harm. On the other hand, a value of 0 indicates that the sequence is in the worst possible ranking order. Following the notation described earlier, the EWN can be defined as follows:

$$\text{EWN} = \frac{\sum_{i=1}^{n}\{2^{-i} \cdot (1 - \rho(x_i))\} - (2^{-p} - 2^{-n})}{(1 - 2^{p-n}) \cdot (1 - 2^{-p})}. \tag{2}$$

Owing to space limitations, we provide the complete derivation for EWN in the appendix.

Remark. In addition to these two novel metrics, we also use the **Precision@K** metric (Shani and Gunawardana, 2011), popular in recommendation systems. In line with our naming convention, we call this metric as the **Top-Pref-k (TP-k)**, which represents the fraction of non-harmful content the user encounters when they consume a k-length content sequence.

4 Experimental Setup

4.1 Datasets

We employ a curated dataset of YouTube videos ($9,832$ harmful, $2,679$ harmless), which were labeled for six categories of harm: information, hate and harassment, addictive, clickbait, sexual, and physical harms. The details on the harm categories, data collection, labeling, reliability, and other aspects of this dataset are provided in (Jo et al., 2024; Jo and Wojcieszak, 2025). Note that we primarily utilize this dataset for the majority of our ablation experiments because of the diversity in harm categories. We use the video descriptions as input to the LLM. To demonstrate the generalizability of our approach, we also evaluate it on two singular-harm category datasets; the *Jigsaw Toxicity* Classification Dataset (Jigsaw, 2019), which contains comments from the Civil Comments platform labeled for toxicity at Jigsaw, and the *Measuring Hate Speech* dataset (D-Lab, 2022) by UC Berkeley's D-Lab, that contains annotated social media posts specifically labeled for hate speech. Additional details on each dataset are provided in Appendix B.

4.2 Content Sequences

Note that our proposed approach operates on content sequences rather than individual instances. We use the datasets described to simulate sequences of content that a user would be exposed to. We sample uniformly at random without replacement from our data to generate a sequence of $n = 20$ content instances (textual data; either titles and transcripts from YouTube videos, comments, or social media posts, depending on the dataset), and generate $m = 100$ such sequences to form a sequence dataset. To study the effect of the *amount* of harm, we generate 5 such datasets by the varying the fraction of harmful content from 10% to 50% for the YouTube dataset. For the other datasets, we fix the fraction of harmful content to 30% so as to reflect the typical harm ratio observed in the wild, as determined by user surveys (Beata Mostafavi, 2020).

4.3 Baselines

We leverage two state-of-the-art harm classification models as benchmarks to compare the performance of our LLM re-ranking approach. The Perspective API [3] is a tool developed by Google/Jigsaw to improve conversations online by detecting toxicity of comments. We use the toxicity score the model returns to rank the content. Similar to Perspective, the OpenAI Moderation API[4] can be used to check whether a text is harmful or not. It provides scores for a variety of harm categories, ranging from hate to sexual or violence; we extract the scores for each category and use the highest score as a proxy for harm.

4.4 Implementation Details

We implement our approach using mainly OpenAI GPT-3.5 Turbo as the underlying LLM. This choice was made based on the available models, API costs, and rate limitations at the time of the initial writing. Additionally, we ran experiments using two open-source models: Llama2-13B and Mistral-7B-Instruct-v0.2. The exact prompts we use for each of our approaches are listed in Appendix C.

4.5 Evaluation

We use three metrics to evaluate the effectiveness of re-ranking: TP-k, PP-k and the EWN as described. Given that our sequences are of 20 videos, we limit our analyses to TP5 and TP10, as they represent the first two quartiles of videos watched. Similarly, for PP-k, we focus our analysis on PP$1-3$, as they represent the amount of content needed to be watched to reach up to at most the third harmful video, which is a practical limit considering the length of our sequences. The PP1, PP2 and PP3 values represent the fraction of the sequence that can be consumed before encountering the first, second and third harmful video(s), respectively.

5 Results

5.1 Effectiveness of LLM Preference Re-ranking

Figure 2 depicts the effectiveness of our approaches with respect to the baselines, in terms of our metrics. In general, our approach outperforms both the baselines across varying harm percentages,

[3] https://perspectiveapi.com/

[4] https://platform.openai.com/docs/guides/moderation

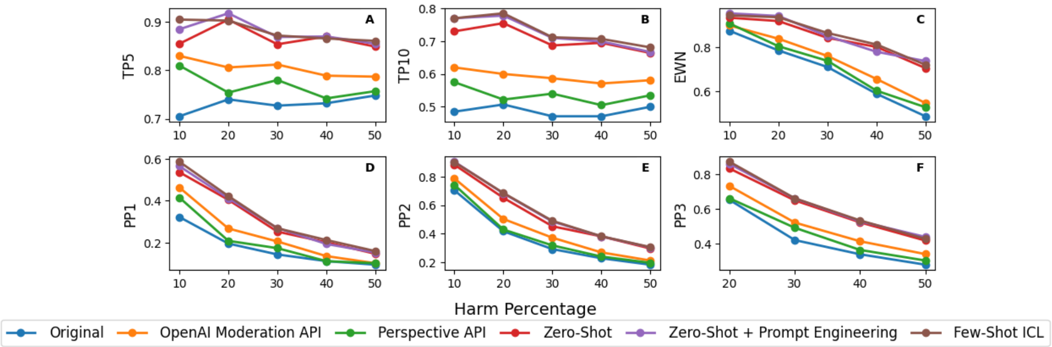

Figure 2: Performance of our proposed method with different preference constraint strategies on varying ratios of harmful content. Higher values indicate better alignment with less harmful content exposure. Our approaches outperform all baselines by a wide margin for the TPk (**A, B**), PPk (**D-F**) and EWN (**C**) metrics. Note that as the harm ratio increases in content sequences, exposure to harmful content increases as well.

demonstrating the improvement preference re-ranking drives over simple classification. We find that LLM-based approaches outperform standard classification approaches even in zero-shot settings, and even more so in few-shot settings. For example, $TP5$ (Figure 2A) indicates the average proportion of non-harmful videos in the first five videos in the sequence. This proportion is 70.5% and 74.8% in the 10% and 50% harm settings initially. The OpenAI Moderation API is able to drive these up to 83% and 78.7%. The preference re-ranking approach in the Few-Shot ICL setting is able to increase the non-harmful videos to 90.5% and 86.1% respectively. Mean TP5 and TP10 values are shown in Tables 1 and 2 respectively. Across both tables, the more advanced configurations (Zero-Shot with PE, Few-Shot ICL) consistently show higher TPk values across all harmful content percentages, demonstrating the advantage of leveraging sophisticated AI techniques. Notably, the effectiveness of the OpenAI Moderation and Perspective approaches varies, often underperforming compared to the LLM-based approaches, despite the massive amounts of *task-specific* data they are trained on.

5.2 Effect of Harm Ratio

In Table 6, we present EWN values for each constraint specification approach across various harm percentages. We see that the difference in EWN values between the original and re-ranked sequences grows as the percentage of harmful content increases. For instance, at 10% harmful content, the gap between the Original and Zero-Shot with PE is about 0.079 (9.1% relative increase), while at 50% harmful content, the gap extends to 0.250 (51.3% relative increase). Additionally, as the percentage of harm increases, OpenAI Moderation and Perspective, show large decreases in their EWN values (39.1% and 41.8% respectively). Zero-Shot with PE and Few-Shot ICL, on the other hand, exhibit the smallest decreases in their EWN scores as the percentage of harmful content increases. Both configurations manage to minimize the performance drop to around 23%. Given that EWN is agnostic to the relative harm ratio (see Appendix A), this analysis shows that as the operational environment becomes more challenging due to higher concentrations of harmful content, the advantage of deploying advanced AI-based content moderation systems becomes increasingly substantial.

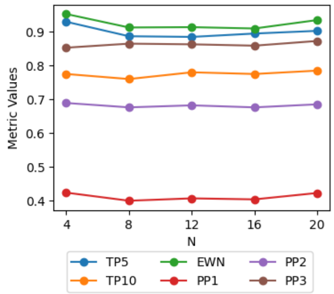

Figure 3: Performance of ICL with varying number of exemplars, as measured by our defined metrics. Harm mitigation effectiveness does not meaningfully improve by increasing the number of exemplars.

5.3 Varying Exemplars for Few-Shot ICL

Because the performance of ICL will depend on exemplars, we experimented with various values for the number of exemplars N. We modify the few-shot ICL approach and implement it in five versions by setting the value of N as one of $4, 8, 12, 16$ and 20. We perform K-Means clustering on the harm samples (setting $K = N$) and choose a representative sample from each cluster. We conduct our analysis in the 30% setting (Beata Mostafavi, 2020). Table 7 and Figure 3 show the performance of ICL-based prompting with varying number of exemplars. Surprisingly, the performance does not necessarily improve with increasing number of exemplars; we see that EWN is the best for $N = 4$ and decreases with increasing N. This is likely due to overfitting the LLM and the bias introduced because of the exemplars. Nevertheless, ICL still remains a better approach as compared to zero-shot or zero-shot with prompt engineering.

5.4 Experiments across Datasets

The results described so far are based on the YouTube dataset (Jo and Wojcieszak, 2025) because of the comprehensive and diverse nature of harms it contains. However, we also find that our LLM-based re-ranking approach performs well on other, more targeted datasets. Table 11 shows that all three configurations (zero-shot, zero-shot with prompt engineering and few-shot) outperform the baselines (albeit by a small margin) on the hate speech dataset as evidenced by the EWN values. Notably, the Zero-Shot approach slightly outperforms the other configurations with a perfect TP10 score of 1.000 and slightly higher PP1 (0.697), PP2 (0.777), and PP3 (0.839) values, indicating a marginally better delay in encountering harmful content. The inclusion of Prompt Engineering and Few-Shot learning yields EWN scores very close to the perfect mark (0.99992 and 0.99991, respectively), though they slightly trail the pure Zero-Shot method. On the Jigsaw dataset (Table 10), our approach achieves a comparable performance to the baselines. Note that the Jigsaw dataset contains content labeled for toxicity, on which the Perspective API is likely trained (as both are released by Jigsaw). Perspective API attains uncharacteristically high performance on this dataset, potentially indicating *test set leakage* in this experiment.

5.5 Experiments across LLMs

The results above have been based on the performance of GPT-3.5 as the primary LLM of choice. However, there are growing concerns about the lack of transparency and data-sharing practices in closed-source LLMs (Balloccu et al., 2024), which might deter developers from utilizing them. Further, the costs associated with querying the LLM at scale can also compound their use. In this section, we thus consider alternative open-source LLMs and re-run experiments with 30% harm, the typical harm ratio observed on platforms (Beata Mostafavi, 2020). We repeat the experiments conducted on the YouTube dataset using two LLMs, namely *Mistral-7B-Instruct-v0.2*[5] and *Llama2-13B*[6], both running locally on a server equipped with a NVIDIA RTX A6000 GPU with 256GB of RAM.

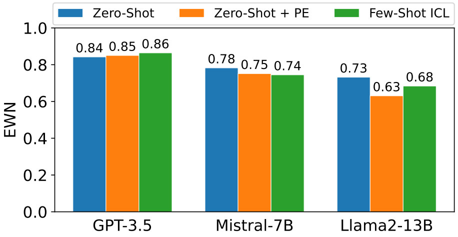

Figure 4: EWN values for Llama2, Mistral, and GPT-3.5-Turbo.

Table 8 shows the various metric values for the YouTube dataset using Mistral. We see that it too outperforms the baselines significantly and only has slightly lower performance compared to GPT-3.5-Turbo. In contrast, Table 9 that Llama2 underperforms across metrics compared to the OpenAI Moderation baseline. Comparing the models' performances with GPT-3.5 directly, Figure 4 shows the EWN values for all three models and each of the three learning strategies we propose. As is evident, GPT-3.5 outperforms both Mistral and Llama2 (likely) due to its larger parameter size. However, the EWN values of Mistral trail GPT-3.5 by an average of nearly 10%, a minimal performance trade-off that make Mistral a viable second option for our approach. Also note that Mistral has only 7B parameters, making it extremely lightweight to run locally and ingest data at scale. In contrast, Llama2 exhibits lower performance than the other two LLMs, a result also consistent with prior work which has found that Mistral outperforms Llama2 on most benchmark datasets (Jiang et al., 2023).

6 Discussion

In this work, we propose methods that leverage LLMs to circumvent two challenges in online harm mitigation: the large scale of annotation required and the dynamic nature of harm.

Utility of LLMs for Harm Mitigation. Our experiments demonstrate that our LLM-based re-ranking approaches significantly outperform traditional/proprietary content moderation methods in reducing harmful content exposure on social media platforms. As detailed above, both the Zero-Shot and Few-Shot In-Context Learning (ICL)

[5] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2
[6] https://huggingface.co/meta-llama/Llama-2-13b-chat-hf

configurations provide a notable improvement over industry-standard harm classifiers. Because our approach shows promising results even in zero-shot settings, LLMs can be used off-the-shelf with minimal effort and without necessitating significant re-configuration or fine-tuning. Customizing LLM prompts can lead to even better results; for example, the Zero-Shot + Prompt Engineering configuration, where harm is explicitly defined in the prompt, consistently performs better than the simpler Zero-Shot approach, indicating that our automated prompt adjustments can significantly influence desirable performance outcomes.

Robustness to Concept Drift. Despite the differences in content types and the diversity of harm categories, our LLM approach consistently outperforms traditional methods, suggesting that the method is versatile and can be applied across diverse social media platforms and content/harm moderation challenges. Due to extensive pre-training, LLMs can generalize across various types of harm without needing explicit fine-tuning or labeled examples for each new harm type. Consequently, our approach excels at identifying and mitigating a wide range of harmful content, whether it involves hate speech, clickbait, hate and harrassment, or addictive material, among other categories. While LLMs are still restricted temporally by training data, they can adapt to novel scenarios without requiring continuous retraining, and generalize better than traditional supervised learning methods.

Impact of Harm Ratio on Re-Ranking Effectiveness. Our results show that the efficacy of re-ranking is closely related to the ratio of harmful content within the dataset. Specifically, the Exponentially Weighted Normalization (EWN) metric reveals that as the percentage of harmful content increases, downranking harmful content becomes increasingly more challenging. The advantage of our approach becomes even more pronounced in high-volume online harm scenarios where baselines suffer from significant performance drops, but our LLM-based methods exhibit minimal performance degradation.

Versatility Across Models. Our experiments across various LLMs show that the performance of our approach will indeed depend on the choice of LLMs. While using commercial LLMs can be expensive, we show that open-source LLMs (such as Mistral) show comparable performance in harm mitigation. This means that our approach can be implemented with open-source models that will not result in excessive overhead due to API costs. Furthermore, satisfactory performance using open-source LLMs also ensures that our method can be utilized by developers concerned with data sharing/privacy issues relating to closed-source LLMs. We defer the study of other LLMs in the context of our approach to future work.

Implications for Preference-Based Ranking. While we focus on re-ranking content to minimize exposure to harmful material, the underlying approach is general and can be adapted to optimize for various other applications beyond harm mitigation. For example, a platform might choose to re-rank content to elevate material that encourages civic participation or promotes mental well-being, thereby aligning recommendations with broader social goals. The flexibility of LLMs in understanding and applying different ranking criteria makes this approach highly adaptable and scalable to different contexts, where the optimization target could shift from reducing harm to enhancing specific *positive* outcomes for users.

7 Conclusion

In sum, we present a novel approach to mitigating exposure to harmful content on social media platforms by leveraging Large Language Models (LLMs) for re-ranking recommendation sequences. LLM-based re-ranking not only outperforms traditional classifier-based commercial moderation systems, such as the Perspective API and OpenAI Moderation API, but also exhibits robustness across various harm scenarios and content types. By utilizing zero-shot, few-shot, and prompt-engineered strategies, our method effectively generalizes to multiple forms of harm without the need for extensive labeled data, addressing both scalability and dynamicity issues inherent in content moderation. The proposed re-ranking technique shows substantial promise in reducing the likelihood of users encountering harmful content, even as the volume of such content increases in users' social media ecosystem (as is the case for the heavy consumers of misinformation, violence, among other harmful contents). For future work, we aim to explore multi-modal input/models to further augment performance and apply our methods to other novel problem domains where content ranking plays a critical role in ensuring user safety and trust.

Limitations

While our proposed approach outperforms existing baselines, it has certain limitations as well. The approach involves queries to LLMs, which may be expensive in terms of time and infrastructure cost when considered at the scale that social media platforms operate (although most popular social media platforms operate with infrastructure that can support such intervention strategies[7]). For smaller organizations, these issues can be minimized by lightweight (e.g. quantized) high-performance LLMs running locally (e.g. Mistral). Another limitation is modality; we only consider textual input in our algorithm and content sequences, even though visual information (e.g. video frames) could be incorporated from YouTube videos to improve performance further. We defer the study of multi-modal LLMs for re-ranking to future work. Finally, LLMs themselves can possess robustness issues (Askari et al., 2025a; Chhabra et al., 2024) in downstream tasks (such as re-ranking), although these issues are exhibited by proprietary harm classification models as well (Achara and Chhabra, 2025).

Ethical Considerations

Through this work, we aim to demonstrate how LLMs can be used to minimize user exposure to harmful content online. Our work does not conduct research with human subjects directly. We do not release any new data; all data used are public, and the authors state they followed required ethics guidelines in the collection phase. Sharing user content with third-party LLMs may introduce privacy challenges, and concerns that the data may be used for training without consent. However, social media companies can run open-source LLMs (which, from our findings, have performance comparable to proprietary ones) without relying on third parties, which would ensure that user content does not leave the company. Finally, while the primary intent of our approach is to mitigate exposure to harmful content, there exists the risk of over-moderation; LLMs, despite their advanced reasoning capabilities, might reflect inherent biases from their training data, potentially leading to the marginalization of minority perspectives or controversial opinions. Transparent moderation guidelines, diverse datasets and a human-in-the-loop moderation mechanisms can help safeguard freedom of speech and enforce effective moderation at the same time.

[7] https://www.facebook.com/metaai/

References

Akshit Achara and Anshuman Chhabra. 2025. Watching the AI Watchdogs: A Fairness and Robustness Analysis of AI Safety Moderation Classifiers. *NAACL*.

Hadi Askari, Anshuman Chhabra, Muhao Chen, and Prasant Mohapatra. 2025a. Assessing LLMs for zero-shot abstractive summarization through the lens of relevance paraphrasing. *NAACL (Findings)*.

Hadi Askari, Shivanshu Gupta, Terry Tong, Fei Wang, Anshuman Chhabra, and Muhao Chen. 2025b. Unraveling indirect in-context learning using influence functions. *arXiv preprint arXiv:2501.01473*.

Simone Balloccu, Patrícia Schmidtová, Mateusz Lango, and Ondřej Dušek. 2024. Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source llms. *arXiv preprint arXiv:2402.03927*.

Jack Bandy. 2021. Problematic machine behavior: A systematic literature review of algorithm audits. *Proceedings of the acm on human-computer interaction*, 5(CSCW1):1–34.

Beata Mostafavi. 2020. Young Kids' YouTube Viewing Dominated by Consumerism, Ads. https://www.michiganmedicine.org/health-lab/young-kids-youtube-viewing-dominated-consumerism-ads.

George Beknazar-Yuzbashev, Rafael Jiménez Durán, Jesse McCrosky, and Mateusz Stalinski. 2022. Toxic content and user engagement on social media: Evidence from a field experiment. *Available at SSRN 4307346*.

Rahul Bhargava, Anna Chung, Neil S Gaikwad, Alexis Hope, Dennis Jen, Jasmin Rubinovitz, Belén Saldías-Fuentes, and Ethan Zuckerman. 2019. Gobo: A system for exploring user control of invisible algorithms in social media. In *Companion Publication of the 2019 Conference on Computer Supported Cooperative Work and Social Computing*, pages 151–155.

Akash Bonagiri, Lucen Li, Rajvardhan Oak, Zeerak Babar, Magdalena Wojcieszak, and Anshuman Chhabra. 2025. Towards safer social media platforms: Scalable and performant few-shot harmful content moderation using large language models. *arXiv preprint arXiv:2501.13976*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

L Elisa Celis, Sayash Kapoor, Farnood Salehi, and Nisheeth Vishnoi. 2019. Controlling polarization in personalization: An algorithmic framework. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 160–169.

Thomas M Chen. 2021. Automated content classification in social media platforms. In *Securing Social Networks in Cyberspace*, pages 53–71. CRC Press.

Anshuman Chhabra, Hadi Askari, and Prasant Mohapatra. 2024. Revisiting Zero-shot Abstractive Summarization in the Era of Large Language Models from the Perspective of Position Bias. In *NAACL*.

Paul Covington, Jay Adams, and Emre Sargin. 2016. Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM conference on recommender systems*, pages 191–198.

UC Berkeley D-Lab. 2022. Measuring hate speech dataset. https://huggingface.co/datasets/ucberkeley-dlab/measuring-hate-speech. Accessed: 2024-08-26.

Fabio Del Vigna12, Andrea Cimino23, Felice Dell'Orletta, Marinella Petrocchi, and Maurizio Tesconi. 2017. Hate me, hate me not: Hate speech detection on facebook. In *Proceedings of the first Italian conference on cybersecurity (ITASEC17)*, pages 86–95.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*.

Marina Ernst. 2024. Identifying textual disinformation using large language models. In *Proceedings of the 2024 Conference on Human Information Interaction and Retrieval*, pages 453–456.

Smita Ghosh, Pramita Das, Sneha Ghosh, and Diptaraj Sen. 2022. Detection of clickbait content spreaders on online social networks. In *2022 5th International Conference on Information and Computer Technologies (ICICT)*, pages 23–28. IEEE.

Mark Glick, Greg Richards, Margarita Sapozhnikov, and Paul Seabright. 2014. How does ranking affect user choice in online search? *Review of Industrial Organization*, 45:99–119.

Robert Gorwa, Reuben Binns, and Christian Katzenbach. 2020. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, 7(1):2053951719897945.

Shivanshu Gupta, Matt Gardner, and Sameer Singh. 2023. Coverage-based example selection for in-context learning. *arXiv preprint arXiv:2305.14907*.

Jonathan Haidt and Jean Twenge. 2022. Adolescent mood disorders since 2010: A collaborative review. *Unpublished manuscript, New York University. Retrieved*, 1.

Jonathan Haidt and Jean Twenge. 2023. Social media and mental health: A collaborative review. *Unpublished manuscript, New York university. Accessed at tinyurl. com/SocialMediaMentalHealthReview*.

Muhammad Haroon, Magdalena Wojcieszak, Anshuman Chhabra, Xin Liu, Prasant Mohapatra, and Zubair Shafiq. 2023. Auditing youtube's recommendation system for ideologically congenial, extreme, and problematic recommendations. *Proceedings of the National Academy of Sciences*, 120(50):e2213020120.

Martin Hilbert, Drew P Cingel, Jingwen Zhang, Samantha L Vigil, Jane Shawcroft, Haoning Xue, Arti Thakur, and Zubair Shafiq. 2023. # bigtech@ minors: Social media algorithms personalize minors' content after a single session, but not for their protection. *Available at SSRN 4674573*.

Martin Hilbert, Arti Thakur, Pablo M Flores, Xiaoya Zhang, Jee Young Bhan, Patrick Bernhard, and Feng Ji. 2024. 8–10% of algorithmic recommendations are 'bad', but… an exploratory risk-utility meta-analysis and its regulatory implications. *International Journal of Information Management*, 75:102743.

M Shamim Hossain and Ghulam Muhammad. 2019. Emotion recognition using deep learning approach from audio–visual emotional big data. *Information Fusion*, 49:69–78.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Jigsaw. 2019. Jigsaw unintended bias in toxicity classification. https://www.kaggle.com/competitions/jigsaw-unintended-bias-in-toxicity-classification. Accessed: 2024-08-26.

Claire Wonjeong Jo, Miki Wesołowska, and Magdalena Wojcieszak. 2024. Harmful youtube video detection: A taxonomy of online harm and mllms as alternative annotators. *arXiv preprint arXiv:2411.05854*.

Wonjeong Jo and Magdalena Wojcieszak. 2025. Metaharm: Harmful youtube video dataset annotated by domain experts, gpt-4-turbo, and crowdworkers. *Proceedings of the International AAAI Conference on Web and Social Media*. In press.

Shehroze Khan and James R Wright. 2021. Disinformation, stochastic harm, and costly effort: A principal-agent analysis of regulating social media platforms. *arXiv preprint arXiv:2106.09847*.

Michal Kosinski, David Stillwell, and Thore Graepel. 2013. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the national academy of sciences*, 110(15):5802–5805.

Geza Kovacs, Zhengxuan Wu, and Michael S Bernstein. 2018. Rotating online behavior change interventions increases effectiveness but also increases attrition. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–25.

Juliane A. Lischka and Marcel Garz. 2021. Clickbait news and algorithmic curation: A game theory framework of the relation between journalism, users, and platforms. *New Media & Society*.

Ye Liu, Jiajun Zhu, Kai Zhang, Haoyu Tang, Yanghai Zhang, Xukai Liu, Qi Liu, and Enhong Chen. 2024. Detect, investigate, judge and determine: A novel llm-based framework for few-shot fake news detection. *arXiv preprint arXiv:2407.08952*.

Zewdie Mossie and Jenq-Haur Wang. 2020. Vulnerable community identification using hate speech detection on social media. *Information Processing & Management*, 57(3):102087.

Brendan Nyhan, Jaime Settle, Emily Thorson, Magdalena Wojcieszak, Pablo Barberá, Annie Y Chen, Hunt Allcott, Taylor Brown, Adriana Crespo-Tenorio, Drew Dimmery, et al. 2023. Like-minded sources on facebook are prevalent but not polarizing. *Nature*, 620(7972):137–144.

Aviv Ovadya and Luke Thorburn. 2023. Bridging systems: open problems for countering destructive divisiveness across ranking, recommenders, and governance. *arXiv preprint arXiv:2301.09976*.

Zhen Qin, Rolf Jagerman, Kai Hui, Honglei Zhuang, Junru Wu, Jiaming Shen, Tianqi Liu, Jialu Liu, Donald Metzler, Xuanhui Wang, et al. 2024. Large language models are effective text rankers with pairwise ranking prompting. *arXiv preprint arXiv:2306.17563*.

Joaquin Quiñonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. 2022. *Dataset shift in machine learning*. Mit Press.

Steve Rathje, Jay J Van Bavel, and Sander Van Der Linden. 2021. Out-group animosity drives engagement on social media. *Proceedings of the National Academy of Sciences*, 118(26):e2024292118.

Kevin Roose. 2019. The Making of a YouTube Radical. The New York Times. `https://www.nytimes.com/interactive/2019/06/08/technology/youtube-radical.html`.

Wilbert Samuel Rossi, Jan Willem Polderman, and Paolo Frasca. 2021. The closed loop between opinion formation and personalized recommendations. *IEEE Transactions on Control of Network Systems*, 9(3):1092–1103.

Nicu Sebe, Michael S Lew, and Dionysius P Huijsmans. 2000. Toward improved ranking metrics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(10):1132–1143.

Guy Shani and Asela Gunawardana. 2011. Evaluating recommendation systems. *Recommender systems handbook*, pages 257–297.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1):22–36.

Zeynep Tufekci. 2018. Youtube, the great radicalizer. *The New York Times*, 10(3):2018.

Yilun Wang and Michal Kosinski. 2018. Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. *Journal of personality and social psychology*, 114(2):246.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

WSJ Staff. 2021. Inside Tiktok's Highly Secretive Algorithm. https://www.wsj.com/video/series/inside-tiktoks-highly-secretive-algorithm/investigation-how-tiktok-algorithm-figures-out-your-deepest-desires/.

Liang Wu, Fred Morstatter, Kathleen M Carley, and Huan Liu. 2019. Misinformation in social media: definition, manipulation, and detection. *ACM SIGKDD explorations newsletter*, 21(2):80–90.

Xudong Yu, Muhammad Haroon, Ericka Menchen-Trevino, and Magdalena Wojcieszak. 2023. Nudging the recommendation algorithm increases news consumption and diversity on youtube. *PNAS Nexus*.

Xudong Yu, Magdalena Wojcieszak, and Andreu Casas. 2024. Partisanship on social media: In-party love among american politicians, greater engagement with out-party hate among ordinary users. *Political Behavior*, 46(2):799–824.

Appendix

A Additional Details on EWN

A.1 Derivation

Let $X = \{x_i\}_{i=1}^{n}$ be a sequence of n content instances, out of which p are harmless, and the remaining $n-p$ are harmful. Let ρ be a decision function such that:

$$\rho(x) = \begin{cases} 1 & \text{if } x \text{ is non-harmful} \\ 0 & \text{if } x \text{ is harmful.} \end{cases} \tag{3}$$

We want to design a metric that penalizes harmful content instances shown towards the *beginning* of the sequence. Thus, content instances at numerically lower ranks have a higher importance in the metric. We can assign exponentially decaying weights to the content instances based on their rank. Specifically, the i^{th} content instance will have a weight of 2^{1-i}. Following this, we can compute a score S for the sequence X as:

$$S(X) = \sum_{i=1}^{n} 2^{1-i} \cdot \rho(x_i). \tag{4}$$

The value of $S(X)$ indicates the goodness of the ranking in sequence X. Lower values indicate more harmful content shown early on. However, it is dependent on the number of content instances in the sequence and the harm ratio; as such, it is not a good metric to compare two sequences which differ in these distributions. Therefore, we want to compute a metric that indicates how good the sequence X is compared to the *best possible ordering* of the same content.

Best Case: In the best case, all the p non-harmful content instances appear at the beginning of the sequence, Therefore, the score $S(X)$ is

$$S(X) = \sum_{i=1}^{p} 2^{1-i}. \tag{5}$$

This is a geometric progression with the first term $a = 1$ and common ratio $r = 2^{-1}$ having p terms. Therefore we have:

$$S_{max}(X) = 2 \cdot (1 - 2^{1-p}). \tag{6}$$

Worst Case: In the worst case, the harmful content is towards the beginning, i.e. the p harmless content instances all occur at the end of the sequence. Therefore the videos from $i = 1 \ldots n-p$ are all harmful. Then the score $S(X)$ is

$$S(X) = \sum_{i=n-p+1}^{n} 2^{1-i}. \tag{7}$$

This is a geometric progression with the first term $a = 2^{p-n}$ and common ratio $r = 2^{-1}$ having p terms. Therefore we have:

$$S_{min}(X) = 2^{p-n+1} - 2^{1-n}. \tag{8}$$

Scaling: Now, we using the score obtained from 5 and the upper and lower bounds obtained from 6 and 8 respectively, we can normalize the score to a $(0, 1)$ range as

$$\text{EWN} = \frac{S(X) - S_{min}(X)}{S_{max}(X) - S_{min}(X)}. \tag{9}$$

Which after replacing all terms, results in the final expression for EWN:

$$\text{EWN} = \frac{\sum_{i=1}^{n}\{2^{-i} \cdot (1 - \rho(x_i))\} - (2^{-p} - 2^{-n})}{(1 - 2^{p-n}) \cdot (1 - 2^{-p})}. \tag{10}$$

A.2 Importance of EWN

While both TPk and PPk are fairly explainable, they vary significantly with the length of the content and the fraction of non-preferential elements. As a result, they cannot be used to compare two sequences which differ in these distributions. The EWN, however, measures the goodness of the ranking in the sequence by normalizing it between the best and worst possible ranking possible. A value of 0 indicates that the sequence is in the optimal order. On the other hand, a value of 1 indicates that the sequence is in the worst possible ranking order. EWN values allow for an apples-to-apples comparison of sequence rankings. If there are two sequences X_1 and X_2, then $EWN(X_1) > EWN(X_2)$ implies that the ordering in X_1 is better than X_2, and this holds even if the sequences differ in length or the fraction of non-preferential content. Here, *better* means less harmful content shown early on.

B Additional Details on the Dataset

In this section, we provide some additional information about the datasets used in our experiments.

B.1 YouTube Harms Dataset

The Harmful YouTube Dataset (Jo and Wojcieszak, 2025) consists of metadata (title, transcript, descriptions) from $19,422$ YouTube videos. These were systemically collected using three approaches (i.e., keyword-based, channel-based, and external dataset integration). The videos were labeled by domain experts and crowdworkers as harmful or belonging to one or more categories of harm. The six harm categories include: information harm, hate and harassment, addictive, clickbait, sexual, and physical harms (developed based on prior work and platform guidelines). The details on the harm categories, data collection, labeling, reliability, and other aspects of this dataset are provided in (Jo et al., 2024; Jo and Wojcieszak, 2025). The data collection and annotation was IRB-approved and followed the best practices so as to minimize harm to human subjects.

B.2 Jigsaw Toxicity Dataset

The Jigsaw Unintended Bias in Toxicity Classification dataset (Jigsaw, 2019) was developed to address the challenge of unintended bias in machine learning models tasked with detecting toxic content. This dataset, also released as part of a Kaggle competition, contains over $1.8M$ comments from the Civil Comments platform, each labeled for toxicity by human moderators. In addition to a primary binary toxicity label, the dataset includes auxiliary labels indicating whether the comment contains identity-based mentions (e.g., referencing a particular race or gender).

B.3 DLab Hate Speech Dataset

The DLab Hate Speech Dataset (D-Lab, 2022) consists of $50,070$ social media comments collected from platforms like YouTube, Reddit, and Twitter. These comments were annotated by $11,143$ crowdworkers, primarily sourced from Amazon Mechanical Turk. Each comment was evaluated across 10 ordinal labels that capture various aspects of hate speech, including sentiment, disrespect, insult, and dehumanization, among others. Each comment is annotated by several moderators and assigned a *hate speech score*. For our experiments, we aggregated the scores by comment and considered the median score to assign a binary label to each comment.

C LLM Prompts

In this section, we describe the prompts we used for each of the approaches presented in Section 3.3.

C.1 Zero Shot

Read the following texts, A and B.
Text A: «Text A goes here»
Text B: «Text B goes here»
Which of these two is more harmful? Respond in the form Response=X, where X is one of A, B or NONE.

C.2 Zero-Shot + Prompt Engineering

Read the following texts, A and B.
Text A: «Text A goes here»
Text B: «Text B goes here»
We define harm as content belonging to the following categories:

- *Information Harms (e.g., Fake news, Conspiracy theories, Unverified medical treatments, Unproven scientific myths)*
- *Hate and Harassment Harms (e.g., Insult, obscenities, Identity attacks or misrepresentation, Hate speech based on gender, race, religion, disability, sexual orientation)*
- *Addictive Harms (e.g Game play, Drug promotion, Gambling videos and gambling promotion)*
- *Clickbaitive Harms (Clickbait titles - using exaggerated words, such as "top 10", "best/most", how to", "hacks", get-rich-quick schemes and financial scams, Gossip about celebrities)*
- *Sexual Harms (e.g., Erotic clips extracted from non-pornographic films, Depiction of sexual acts, such as sex and groping, Depiction of clothed or unclothed genitals, breasts, or buttocks)*
- *Physical Harms (e.g, Self-injury, Eating disorder promotion, Potentially dangerous challenges and pranks, Violent content)*

Based on this definition of harm, which of the two texts (A or B) is more harmful? Respond in the form Response=X, where X is one of A, B or NONE.

C.3 Few-Shot ICL

In this approach, our prompt is in the form of a multi-dialogue conversation. The OpenAI API allows us to establish prior knowledge by providing a series of dialogues between the user and the LLM. As fitting exemplars in a single problem can prove challenging, we use a multi-step prompt here, as follows.

User: *Below are N examples of content that we consider harmful, separated by ten '=' symbols.*

Exemplar 1

==========

Exemplar 2

==========

.

.

.

==========

Exemplar N

System: *Based on these examples, I understand what constitutes harmful content.*

User: *Read the following texts, A and B.*

Text A: «Text A goes here»

Text B: «Text B goes here»

Based on the definition of harm you have learned from the samples provided, which of the two texts (A or B) is more harmful? Respond in the form Response=X, where X is one of A, B or NONE.

D Detailed Results

In this section, we present our detailed results. These results form the basis of Figure 2.

D.1 Top-Pref-k (*TPk*)

The TPk metric measures user exposure to *preferential content* by calculating the fraction of content deemed preferential in the first k items of the sequence. This metric directly addresses the impact of the *initial content* a user encounters, which is often decisive in shaping user experience and perception. Given that our sequences are of 20 videos, we limit our analyses to TP5 and TP10, as they represent the first two quartiles of videos watched. Mean TP5 and TP10 values are shown in Tables 1 and 2 respectively.

Table 1: TP5 values for constraint specification approaches. Columns represent the percentage of harmful videos in the sequence. *PE → Prompt Engineering*.

Configuration	10	20	30	40	50
Original	0.705	0.740	0.727	0.732	0.748
OpenAI Moderation	0.830	0.806	0.812	0.789	0.787
Perspective	0.810	0.754	0.780	0.742	0.757
Zero-Shot	0.855	0.905	0.854	0.870	0.849
Zero-Shot + PE	0.885	0.918	0.869	0.870	0.855
Few-Shot ICL	0.905	0.903	0.872	0.866	0.861

Table 2: TP10 values for constraint specification approaches. Columns represent the percentage of harmful videos in the sequence. *PE → Prompt Engineering*.

Configuration	10	20	30	40	50
Original	0.485	0.507	0.471	0.471	0.500
OpenAI Moderation	0.620	0.600	0.587	0.571	0.581
Perspective	0.575	0.522	0.540	0.505	0.535
Zero-Shot	0.730	0.755	0.687	0.695	0.664
Zero-Shot + PE	0.770	0.778	0.710	0.699	0.667
Few-Shot ICL	0.770	0.785	0.712	0.707	0.681

D.2 Per-Pref-k (*PPk*)

We focus our analysis on PP$1-3$, as they represent the amount of content needed to be watched to reach up to at most the third harmful video, which we believe is a practical limit considering the length of our sequences. The PP1, PP2 and PP3 values represent the fraction of the sequence that can be consumed before encountering the first, second and third harmful video(s), respectively. Similar to TPk values, the PP$1 - 3$ values also demonstrate the effectiveness of LLM-based approaches over traditional classifier-based methods.

Table 3: PP1 values for constraint specification approaches. Columns represent the percentage of harmful videos in the sequence. *PE → Prompt Engineering*.

Configuration	10	20	30	40	50
Original	0.322	0.195	0.143	0.111	0.094
OpenAI Moderation	0.463	0.267	0.205	0.135	0.101
Perspective	0.415	0.208	0.173	0.111	0.101
Zero-Shot	0.536	0.404	0.252	0.201	0.146
Zero-Shot + PE	0.565	0.411	0.269	0.193	0.152
Few-Shot ICL	0.586	0.422	0.268	0.212	0.159

Table 4: PP2 values for constraint specification approaches. Columns represent the percentage of harmful videos in the sequence. *PE → Prompt Engineering*.

Configuration	10	20	30	40	50
Original	0.704	0.417	0.292	0.228	0.183
OpenAI Moderation	0.787	0.503	0.372	0.269	0.212
Perspective	0.742	0.429	0.318	0.240	0.196
Zero-Shot	0.883	0.651	0.452	0.381	0.296
Zero-Shot + PE	0.904	0.682	0.485	0.381	0.304
Few-Shot ICL	0.901	0.685	0.490	0.382	0.306

Table 5: PP3 values for constraint specification approaches. Columns represent the percentage of harmful videos in the sequence. *PE → Prompt Engineering*.

Configuration	10	20	30	40	50
Original	-	0.653	0.421	0.339	0.279
OpenAI Moderation	-	0.733	0.522	0.414	0.34
Perspective	-	0.662	0.492	0.364	0.303
Zero-Shot	-	0.834	0.650	0.523	0.418
Zero-Shot + PE	-	0.861	0.661	0.529	0.439
Few-Shot ICL	-	0.873	0.662	0.534	0.428

D.3 EWN

The EWN metric and it's importance have already been described in great detail in Appendix A. The EWN metric measures the effectiveness of these configurations in prioritizing non-harmful content by applying exponentially decaying weights to content ranks, with the results normalized to a scale from 0 to 1. Higher EWN values indicate better performance in maintaining a sequence of preferred content. In Table 6, we present EWN values for each constraint specification approach across various harm percentages.

Table 6: EWN values for constraint specification approaches. Columns represent the percentage of harmful videos in the sequence. *PE → Prompt Engineering*.

Configuration	10	20	30	40	50
Original	0.874	0.785	0.710	0.589	0.487
OpenAI Moderation	0.898	0.838	0.760	0.656	0.547
Perspective	0.907	0.804	0.738	0.603	0.528
Zero-Shot	0.933	0.918	0.842	0.801	0.705
Zero-Shot + PE	0.953	0.941	0.850	0.779	0.737
Few-Shot ICL	0.944	0.935	0.864	0.812	0.720

D.4 In-Context Learning

We also study how the performance of preference-based re-ranking varies with the number of exemplars. Table 7 and Figure 3 show the performance of ICL-based prompting with varying number of exemplars.

Table 7: Metrics for variations in the number of exemplars provided as part of ICL.

N	TP5	TP10	EWN	PP1	PP2	PP3
4	0.930	0.775	0.953	0.423	0.689	0.853
8	0.887	0.760	0.913	0.399	0.676	0.865
12	0.885	0.780	0.914	0.406	0.682	0.863
16	0.895	0.775	0.910	0.403	0.676	0.859
20	0.903	0.785	0.935	0.422	0.685	0.873

D.5 Results Across LLM Architectures

Table 8: Results of LLM-based re-ranking using Mistral-7B.

Configuration	TP5	TP10	PP1	PP2	PP3	EWN
Original	0.727	0.471	0.143	0.292	0.421	0.710
OpenAI Moderation	0.812	0.587	0.205	0.372	0.522	0.760
Perspective	0.780	0.540	0.173	0.318	0.492	0.738
Zero-Shot	0.805	0.581	0.187	0.366	0.529	0.783
Zero-Shot + PE	0.753	0.550	0.166	0.332	0.499	0.742
Few-Shot ICL	0.818	0.645	0.217	0.407	0.583	0.824

Table 9: Results of LLM-based re-ranking using Llama2-13B.

Configuration	TP5	TP10	PP1	PP2	PP3	EWN
Original	0.727	0.471	0.143	0.292	0.421	0.710
OpenAI Moderation	0.812	0.587	0.205	0.372	0.522	0.760
Perspective	0.780	0.540	0.173	0.318	0.492	0.738
Zero-Shot	0.778	0.545	0.178	0.324	0.489	0.742
Zero-Shot + PE	0.746	0.521	0.137	0.304	0.466	0.631
Few-Shot ICL	0.751	0.541	0.150	0.309	0.489	0.684

D.6 Results Across Datasets

Table 10: Results on the Jigsaw toxicity dataset. LLM-based re-ranking improves content sequences and achieves performance comparable to baselines.

Configuration	TP5	TP10	PP1	PP2	PP3	EWN
Original	0.758	0.521	0.158	0.325	0.474	0.704
OpenAI Moderation	0.975	0.895	0.501	0.647	0.764	0.988
Perspective	0.995	0.980	0.671	0.781	0.842	0.990
Zero-Shot	0.953	0.818	0.398	0.587	0.702	0.962
Zero-Shot + PE	0.973	0.830	0.433	0.603	0.719	0.982
Few-Shot ICL	0.933	0.816	0.386	0.590	0.716	0.939

Table 11: Results on the DLab hate speech dataset. LLM-based re-ranking improves content sequences and outperforms both baselines.

Configuration	TP5	TP10	PP1	PP2	PP3	EWN
Original	0.738	0.488	0.157	0.307	0.442	0.69800
OpenAI Moderation	1.000	0.995	0.681	0.766	0.827	0.99990
Perspective	0.998	0.988	0.655	0.747	0.819	0.99470
Zero-Shot	1.000	1.000	0.697	0.777	0.839	0.99993
Zero-Shot + PE	1.000	0.998	0.695	0.773	0.837	0.99992
Few-Shot ICL	1.000	0.998	0.695	0.770	0.834	0.99991

E Code and Reproducibility

Our code and the data used are available at the following repository: `https://github.com/rvoak/harm-ranking-llm/`.