

Towards Safer Social Media Platforms: Scalable and Performant Few-Shot Harmful Content Moderation Using Large Language Models

Akash Bonagiri^{1*}, Lucen Li^{1*}, Rajvardhan Oak^{1,3*}, Zeerak Babar¹,
Magdalena Wojcieszak¹, Anshuman Chhabra²

¹University of California, Davis

²University of South Florida

³Microsoft Corporation, USA

{sbonagiri, lcnli, rvoak, zebabar, mwojciezak}@ucdavis.edu, anshumanc@usf.edu

Note: *This paper may contain examples of content that is inappropriate, offensive or amounting to hate speech.*

Abstract

The prevalence of harmful content on social media platforms poses significant risks to users and society, necessitating more effective and scalable content moderation strategies. Current approaches rely on human moderators, supervised classifiers, and large volumes of training data, and often struggle with scalability, subjectivity, and the dynamic nature of harmful content (e.g., violent content, dangerous challenge trends, etc.). To bridge these gaps, we utilize Large Language Models (LLMs) to undertake few-shot dynamic content moderation via in-context learning. Through extensive experiments on multiple LLMs, we demonstrate that our few-shot approaches can outperform existing proprietary baselines (Perspective and OpenAI Moderation) as well as prior state-of-the-art few-shot learning methods, in identifying harm. We also incorporate visual information (video thumbnails) and assess if different multimodal techniques improve model performance. Our results underscore the significant benefits of employing LLM based methods for scalable and dynamic harmful content moderation online.

1 Introduction

Social media platforms are an integral part of people's daily lives, influencing how individuals communicate, share information, and connect with others. Platforms such as Facebook and YouTube are not only tools for interaction, but also play a crucial role in networking, marketing, education, sales, and political campaigns. However, platforms can also inadvertently disseminate a myriad of harmful content to their users. Over 30% of users have encountered hate speech or aggressive behavior online¹, false stories are 70% more likely to be shared online than true ones (Vosoughi, Roy, and Aral 2018), vulnerable users are targeted with recommendations to problematic and distressing content (Hilbert et al. 2023), and dangerous challenges are propagated through so-

cial media (Haidt and Twenge 2023). Exposure to harmful content can have negative consequences for individuals, groups, and the society at large (Haidt and Twenge 2023).

Given these potential consequences, platforms take a number of steps to decrease exposure to harmful content. These approaches range from removal (e.g., banning content that encourages violence, suicide, or eating disorders (YouTube 2021)), limiting amplification without removal (e.g., minimizing the sharing of controversial posts), or providing additional information alongside potentially harmful content (e.g., offering links to fact-checking pages to debunk false claims (Twitter 2023)). All these strategies depend on dynamic and scalable approaches for identifying ever-changing harmful content online.

To this end, platforms generally employ a combination of human moderators and automated classifiers (Facebook 2024). Human moderators review content flagged as harmful either by users or automated content moderation systems, and decide if it violates the platform's policies. In some cases, Machine Learning (ML) classifiers trained on large human-annotated datasets (Jigsaw 2023; OpenAI 2023) are used to identify misinformation, hate speech, sexually explicit material, among other types of harm. However, both human and supervised content moderation classifiers possess significant drawbacks. Human moderation is subjective, not easily scalable, and prone to operator error (Gillespie 2020). In turn, supervised ML classifiers are inefficient, as they require large volumes of human-annotated data for training and struggle with nuances in language such as sarcasm and context-specific meanings (Baruah et al. 2020). These issues with supervised ML models are further compounded by the fact that harmful content classification is a *dynamic temporal problem* (e.g., new conspiracies or new dangerous challenges become popular). As a result, content moderation classifiers are susceptible to concept drift (Quiñonero-Candela et al. 2022) and require humans to annotate large amounts of data at frequent intervals, exacerbating the aforementioned scalability issues.

It is apparent then that an *ideal* solution to the harmful content moderation problem needs to be *highly scalable* (i.e. require minimal supervised signal and human effort) and *highly effective* (i.e. attain human moderator-level accuracy). Moreover, the approach should be easily extensible and

*These authors contributed equally.

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

¹<https://www.pewresearch.org/internet/2021/01/13/the-state-of-online-harassment/>.

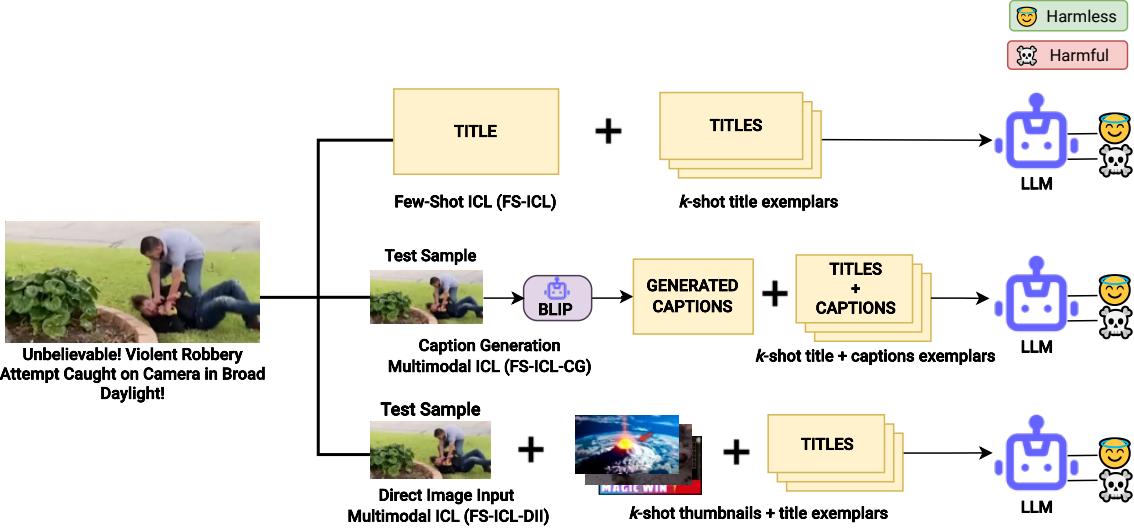


Figure 1: Our overall experimental framework utilizing LLMs for k -shot harmful content classification. The top approach showcases few-shot text-only in-context learning using the video’s title (FS-ICL). The other two methods (FS-ICL-CG) are multimodal and augment the video title information with visual input (e.g. video thumbnail). In FS-ICL-CG we utilize a caption generation module (BLIP) to convert the visual input into text and feed these to a text-only LLM. In FS-ICL-DII we utilize fully multimodal LLMs and feed the visual input to the model directly. The output of each method is a classification of *Harmful/Harmless* for each given test input. Note that by setting $k = 0$ we can obtain zero-shot learning (ZSL) variants for each of the aforementioned approaches.

adapted to new types or categories of harm as they emerge online. In this paper, we demonstrate that this is indeed possible with Large Language Models (LLMs). We utilize a recent multimodal dataset (Jo, Wesołowska, and Wojcieszak 2024) consisting of 19,422 YouTube videos, annotated by domain experts, Amazon MTurkers, and LLMs for six different harm categories: *Information, Hate and Harassment, Addictive, Clickbait, Sexual and Physical Harms*.

Our overall experimental framework is outlined in Figure 1. We consider a number of *open-source* and *closed-source* LLMs² to first show that LLMs can classify harmful content even in the zero-shot setting (i.e. no additional information is provided to them except for the task definition) better than the currently used proprietary baselines, such as Google’s Perspective API (Jigsaw 2023) and OpenAI’s Moderation API (OpenAI 2023). We then show that LLM performance for social media content moderation can be significantly improved by using *state-of-the-art few-shot in-context learning* (Dong et al. 2022; Gupta, Gardner, and Singh 2023) approaches, where a few examples demonstrating harmful/harmless content are provided to the model in the instruction prompt. We showcase how these methods outperform existing deep learning few-shot classification models, establishing LLMs as the state-of-the-art in classifying social media harm in the minimal supervision setting. Finally, we investigate whether LLM performance at this task can be further augmented by incorporating visual input in the learning pipeline (i.e. video thumbnails) and

employing Vision-Language Models (VLMs). Our results demonstrate important future directions for social media harm classification and content moderation in the era of LLMs, and pave the way for accelerated efforts in this promising research landscape.

In sum, we make the following contributions:

- We show that LLMs, even in the zero-shot setting, can outperform existing proprietary classifiers for harm identification, such as Perspective API and OpenAI Moderation API.
- We then examine the performance of LLMs for harm detection under the few-shot setting and showcase that our approaches can adapt to the dynamic nature of harmful content with as few as 8 exemplars. We employ state-of-the-art inference-time approaches for in-context learning to select exemplars and find that LLMs can outperform existing few-shot learning methods when provided the same number of exemplars.
- We also incorporate multimodal LLMs and visual input (video thumbnails) to analyze content, and find that this leads to improved accuracy in identifying harmful content. However, these performance gains are contingent on the LLM being used and open-source multimodal models (e.g., LLaVa (Liu et al. 2023)) exhibit lower performance compared to their closed-source counterparts (e.g. GPT-4o (OpenAI 2024)).

2 Related Work

Categorizing Social Media Harms. Social media platforms host a range of content that can be categorized as

²We primarily consider Llama2-12B, Mistral-7B, GPT-4o-Mini, and GPT-3.5-Turbo in this work.

harmful. The combination of platform affordances, which make (problematic) content production and dissemination cheap and efficient (Munger and Phillips 2022), and recommender systems trained to maximize user engagement, make exposure to online harm far from infrequent. Digital traces each user leaves on platforms reveal information about the user’s emotions (Hossain and Muhammad 2019), personality (Youyou, Kosinski, and Stillwell 2015), substance use (Kosinski, Stillwell, and Graepel 2013), and sexual orientation (Wang and Kosinski 2018). In their effort to maximize engagement, algorithms can capitalize on this inferred information to recommend content that can inadvertently expose users to various 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)). Meta-reviews have shown that 8%-10% of recommendations on platforms pose risks to users (Hilbert et al. 2024) and algorithmic audits have detected discriminatory or otherwise harmful biases in algorithms (Bandy 2021; Haroon et al. 2023). For instance, distressing content is recommended to struggling adolescents on YouTube and Instagram (Hilbert et al. 2023) and up to 40% of health-related content on social media can be classified as misleading or false (Cinelli et al. 2020). Exposure to various categories of online harm can decrease mental health, foment addictions, or even lead to physical harms (Haidt and Twenge 2023; General 2023).

Automated Harm/Content Moderation. Given their detrimental effects, mitigating harms on social media has received widespread attention from the community. Extensive research has been conducted on automated methods for detecting hate speech online (Fortuna and Nunes 2018; Del Vigna12 et al. 2017) as well as misinformation (Aldwairi and Alwahedi 2018; Islam et al. 2020). For instance, (Dacon et al. 2022) presented a BERT-based approach that demonstrated strong performance in recognizing hate speech, offensive language, and other harmful behaviors towards the LGBTQIA+ community. (Kwangho Song 2020) proposed a multimodal stacking scheme that combines both visual and auditory classifiers to detect online pornographic content, improving accuracy and reducing false negatives by using bi-directional recurrent and convolutional neural networks.

The advent of LLMs has led to new efforts in this space. The ability of LLMs to leverage learned patterns from large-scale data enables them to understand and generate coherent, contextually appropriate text. LLMs are highly effective at a variety of NLP tasks. They can operate in a zero-shot or few-shot setting and demonstrate understanding of context and user intentions without extensive task-specific training (Brown et al. 2020). Accordingly, growing work has used LLMs to tackle online harms. For instance, (Wang et al. 2024) present LLM-based approach to generate counterspeech to mitigate hate speech. LLMs have also shown promise in other harm-related tasks, such as detecting clickbait (Sekharan and Vuppala 2023), misinformation (Santra 2024) and abuse (Nguyen, Wilson, and Dalins 2023). While extensive, there are a few issues with extant work. First, each work focuses on one specific kind

of harm and relies on data labeled for that harm category. As a result, research in this area is highly fragmented and there is a lack of solutions that approach this problem from a holistic perspective (Arora et al. 2023; Jo, Wesołowska, and Wojcieszak 2024). Additionally, harm is an ever-evolving concept that changes as online communities and content adapt with time (Mundrievskaya et al. 2023). Existing approaches do not consider the concept drift and may not generalize to new kinds of harms without explicit labeled examples (which are expensive in terms of manual moderation efforts and time). We aim to address both these challenges by presenting a harm-agnostic approach that does not require extensive labeling.

3 LLMs for Harm Classification

We now describe our LLM based strategies for harmful content moderation in the zero-shot and few-shot (in-context learning or ICL) unimodal (text-only) and multimodal (text + vision) settings. For undertaking multimodal classification, we describe two approaches: (1) converting the visual input into text for use with text-only LLMs (as in past work (Yang et al. 2024)) and (2) directly utilizing visual input with natively multimodal LLMs/VLMs. We first define the harm classification problem analytically and then formalize our zero-shot and few-shot/ICL approaches.

3.1 Problem Formulation

Let $X = \{x_i \mid x_i = \langle t_i, v_i \rangle\}_{i=1}^n$ be a sequence of n content instances, where each x_i may contain both visual and textual components v_i and t_i respectively. Let \mathcal{L} denote an LLM. Our goal is to learn a function $f : X \rightarrow \{0, 1\}$ using \mathcal{L} that operates on a given prompt \mathcal{P} to map every content instance $x \in X$ to a binary label, such that:

$$f(x) = \begin{cases} 1 & \text{if } x \text{ is harmful,} \\ 0 & \text{if } x \text{ is harmless.} \end{cases}$$

3.2 Zero-Shot Learning (ZSL)

ZSL (Yin, Hay, and Roth 2019) is the the ability of a model to classify data it has never encountered during training and perform tasks without explicit pretraining. Because LLMs are trained on vast amounts of data, we expect them to have an inherent understanding of what constitutes harmful content. In the ZSL approach, we provide a sample as input to the model, and ask the LLM to classify it as *harmful* or *harmless*, without explicitly providing any demonstrations for what constitutes harmful/harmless. Now, the function f is defined using the LLM \mathcal{L} and the content sample x_i , as:

$$f_{\text{ZSL}}(x_i) := \mathcal{L}(t_i, \mathcal{P}_{\text{ZSL}}),$$

where \mathcal{P}_{ZSL} is a prompt that guides \mathcal{L} to classify x_i as harmful or harmless without explicit examples of harmful content. The text input t_i for each content sample x_i comprises of the YouTube video’s title. The exact prompt we use can be found in Appendix C.1.

3.3 Few-Shot In-Context Learning (FS-ICL)

Few-shot learning (FSL) (Brown et al. 2020) refers to the ability of a model to learn new tasks from a very limited amount of supervised data, often restricted to 8-14 annotated samples. In this setting, we utilize LLMs to perform harm classification by providing a small number of task-specific exemplars within the input prompt, via in-context learning (Dong et al. 2022) (ICL). We opt for coverage-based ICL approaches as they have been shown to be highly successful at generalizing to a wide array of domains (Gupta, Gardner, and Singh 2023). These methods ensure maximally informative demonstrations are selected by submodular optimization, which efficiently maximizes the coverage of salient aspects of the content sample. Selectors such as BERTScore, cosine, and BM25 can guide the selection process. Thus, we utilize coverage-based approaches and augment the ZSL prompt to include k labeled examples contained in the exemplar set $\mathcal{E} = \{(x_j, y_j)\}_{j=1}^k$, where $y_j \in \{0, 1\}$ represents the true known label for x_j . The examples in \mathcal{E} are selected based on their ability to maximize coverage of salient aspects and are ordered by relevance, with the most relevant examples placed closest to x_i . The function f takes the form:

$$f_{\text{FS-ICL}}(x_i) := \mathcal{L}(t_i, \mathcal{E}, \mathcal{P}_{\text{FS-ICL}})$$

Here, $\mathcal{P}_{\text{FS-ICL}}$ is an augmented prompt that instructs the LLM \mathcal{L} to utilize the demonstration set \mathcal{E} while performing the classification based on the in-context exemplars utilizing BERTScore, Cosine, BM25 selectors. More details on selectors are provided in Appendix B.

3.4 Multimodal FS-ICL

We enhance the FSL approach by including features from another modality (vision) along with the textual input for each content sample. For YouTube videos, this constitutes using video thumbnails. To incorporate visual input, we use multimodal LLMs (also referred to as VLMs): GPT-4o-Mini (OpenAI 2024), OpenFlamingo (Alayrac et al. 2022) and LLaVA (Liu et al. 2023). We employ models in two key ways: (1) *Caption Generation* (where we generate captions for the image input associated with the content sample and pass this in the text prompt) and (2) *Direct Image Input* (where we utilize a VLM that can operate on multimodal text + visual input directly). Note that we can also obtain the zero-shot setting as a special case for these multimodal approaches by simply discarding the exemplar set in few-shot learning.

Caption Generation (CG) We pass the input image(s) (thumbnails) to BLIP pretrained (Bootstrapping Language-Image Pre-training) model on the captioning task (Shen et al. 2022; Li et al. 2022) and extract the generated caption. Then, we augment the textual content instance with this caption. Additionally, we did the same experiment with Qwen-VL-chat model (Bai et al. 2023). More details are mentioned in the Appendix G. We essentially follow the same procedure for the text-only FS-ICL approach but with the generated caption also provided as input to the LLM. Essentially, denoting the captioning VLM as V , the function f incorporates

a caption c_i generated by V from the visual component, i.e. $c_i = V(v_i)$, as follows:

$$f_{\text{FS-ICL-CG}}(x_i) := \mathcal{L}(t_i, c_i, \mathcal{E}, \mathcal{P}_{\text{FS-ICL-CG}}).$$

The goal here is to extract additional signals from the image, which may not be present in the text (for example, a YouTube video may have a harmless caption, but a sexually explicit or clickbait thumbnail). However, errors in the caption generation process might propagate further down the learning pipeline and reduce LLM performance.

Direct Image Input (DII) Some LLMs/VLMs, such as GPT-4o-Mini (OpenAI 2024) operate on multimodal input directly (Radford et al. 2021) (i.e., they can process and integrate both visual and textual data simultaneously). Now, by simply utilizing such an VLM in the FS-ICL approach and providing image data along with text input, we can augment task performance. The classification function is as follows:

$$f_{\text{FS-ICL-DII}}(x_i) := \mathcal{L}(v_i, t_i, \mathcal{E}, \mathcal{P}_{\text{FS-ICL-DII}}).$$

The Direct Image Input (DII) approach leverages multimodal data, enhancing the model’s understanding by directly incorporating visual cues. This approach can lead to more accurate and context-aware predictions, especially in cases where visual content carries critical information. However, it is relatively computationally intensive.

4 Experiments and Results

4.1 Experimental Setup

Datasets. We employ a curated dataset of YouTube videos (Jo, Wesołowska, and Wojcieszak 2024) that encompasses distinct types of harms identifiable in multimodal social media data. Each video in the dataset was labeled as harmful/harmless by crowdworkers, LLMs, and domain experts. The ground truth is taken as the majority label across these three labeling actors. We used a train-test split of 3,000 videos evenly divided between harmful and harmless videos. We chose this dataset because of the diverse nature of harms represented as well as features from multiple modalities (text, image). Additionally, to validate the generalizability of our approach, we evaluate it on two publicly available datasets; the Jigsaw Toxicity Classification Dataset (Google/Jigsaw 2019) which consists of comments from the Civil Comments platform, labeled for toxicity and various identity-based biases, and the Measuring Hate Speech dataset (D-Lab 2022) by UC Berkeley’s D-Lab, which contains a large corpus of annotated social media posts specifically labeled for hate speech. More details on all datasets are provided in Appendix A.

Models. For experiments, we use two open-source LLMs, Mistral-7B and Llama2-13B. Additionally, we employ proprietary closed-source OpenAI LLMs, GPT-3.5-Turbo and GPT-4o-Mini. We use the BLIP (Li et al. 2022) model for generating captions and LLaVA (Liu et al. 2024), GPT-4o-Mini (OpenAI 2024), OpenFlamingo (Alayrac et al. 2022) for vision-based ICL. All the prompts we design are provided in Appendix C.

Table 1: Performance comparison of various baseline models on various metrics (%).

Model	Accuracy	Precision	Recall	F-1
Perspective API	50.36	50.20	98.00	66.40
Crowdsourced - Worst	53.40	53.50	53.39	53.03
OpenAI Moderation API	55.93	53.69	86.67	66.31
Crowdsourced - Majority	61.23	63.28	61.21	59.65
Domain Expert	90.97	91.93	90.96	90.91

4.2 Baselines

Specific FSL Baselines. We considered two state-of-the-art baselines from prior work: *Prototypical Networks* (Snell, Swersky, and Zemel 2017) and *Matching Networks* (Vinyals et al. 2016). Matching networks involve a similarity-based approach where the aim is to adapt to new classes with minimal labeled data by leveraging the powerful text representations from transformer-based models such as BERT. Prototypical Networks learn a metric space wherein classification is performed by computing distances to prototype representations of each class.

Proprietary Baselines. In addition to the deep learning few-shot baselines, we use two publicly available, industry-grade moderation modules to compare our performance. The Perspective API (Jigsaw 2023) is a tool developed by Jigsaw to help improve online conversations. It uses machine learning models to identify and measure the level of toxicity in online content. The API is used for content moderation by several leading platforms or websites such as Reddit or the New York Times. The OpenAI Moderation (OpenAI 2023) API is designed to help developers identify and handle potentially harmful or unsafe content within text.

4.3 Results

Zero-Shot Harm Detection. In ZSL, we utilize LLMs to classify content as harmful/harmless without using labeled training examples. As shown in Figure 2, the LLMs significantly outperformed proprietary baselines in terms of accuracy. Specifically, the OpenAI Moderation API achieves 56% accuracy and Perspective API achieves 50% accuracy as shown in Table 1. Our ZSL approach, on the other hand, was able to achieve accuracies of 65% (Mistral-7B), 69% (Llama2-13B; GPT-4o-Mini), and 70% (GPT-3.5-Turbo). We also compare performance across different ZSL settings over various metrics, such as Precision, Recall, and F1 score as shown in Table 2

Few-Shot Harm Detection. We employ few-shot ICL to leverage the full capacity of LLMs and aim to achieve performance comparable to the domain expert accuracy of 90.67%. The FS-ICL approaches demonstrate a significant improvement over the zero-shot configurations, with a notable 5-10% increase in accuracy. Among the selectors, BSR consistently achieves the best performance.

LLMs in the few-shot setting also significantly outperform open-source FSL baselines such as Prototypical Networks and Matching Networks. As seen in Figure 3, GPT-4o-Mini with 12 shots and the BERTScore selector

Table 2: Performance comparison of models across ZSL, ZSL-CG, and ZSL-DII settings using various metrics (%).

Model	Accuracy	Precision	Recall	F-1
ZSL				
Mistral-7B	65.23	66.55	65.22	64.52
Llama2-13B	68.97	69.25	68.96	68.85
GPT-4o-mini	68.67	68.75	68.67	68.63
GPT-3.5-turbo	70.00	70.00	70.00	70.00
ZSL CG				
Mistral-7B	63.41	63.58	63.60	63.41
Llama2-13B	67.93	67.81	67.65	67.68
GPT-4o-mini	68.23	69.31	69.89	69.59
GPT-3.5-turbo	70.20	72.96	71.34	72.14
ZSL DII				
OpenFlamingo	50.00	50.00	50.00	50.00
LLaVa VLM	54.00	59.50	62.50	60.50
GPT-4o-mini	70.00	70.00	72.00	71.00

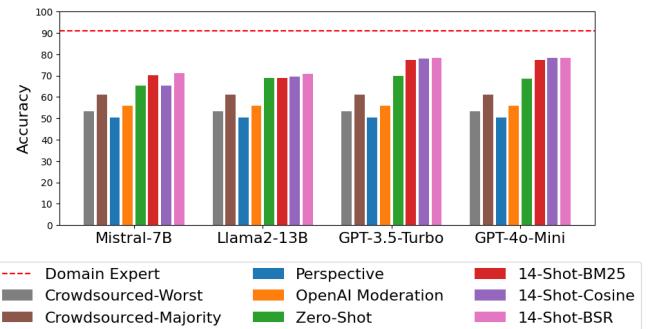


Figure 2: Accuracy (%) across different models for ZSL and FS-ICL (14-Shot; using BM25, Cosine, BSR selectors), proprietary baselines (Perspective and OpenAI Moderation APIs), crowdsourced annotators (*Crowdsourced-Worst*: minority label; *Crowdsourced-Majority*: majority label), and domain experts.

in the FS-ICL approach achieves an accuracy of 78.57%, the best-performing baseline model, whereas Prototypical Network (roberta-base) using 12 shots, achieves only 67.59% accuracy. Similarly, GPT-3.5-turbo in the FS-ICL approach achieves an accuracy of 78.33% using 14 shots and BERTScore, while the Matching Network (bert-base-uncased) with 14 shots achieves a much lower accuracy of 67.38%. Overall, the closed-source models outperform the open-source LLMs by a significant margin.

We also observe that the number of shots has a minor impact on the accuracy (e.g., 8-shot: 78.2% vs. 14-shot: 78.13% for GPT-4o-Mini using BSR). This may be because the number of salient aspects in the content samples, i.e., video titles, was not greater than the number of shots tested, due to YouTube’s title length limitations. While 14 shots resulted in slightly better performance, 8 shots was sufficient to cover most of the salient aspects needed for LLMs to determine whether a video is harmful. We offer a few examples to qualitatively demonstrate this in Figure 4.

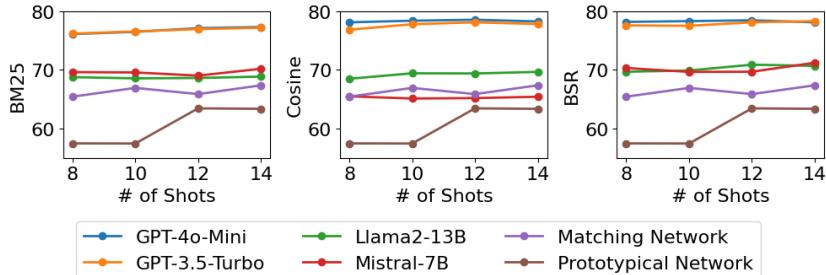


Figure 3: Accuracy (%) of different LLMs for FS-ICL while varying number of shots and ICL selection methods, and open-source deep learning baselines (Matching Network and Prototypical Network) as well proprietary baselines. Among the selectors, BSR achieves the highest accuracy overall and across LLMs, GPT-4o-Mini achieves the highest performance. Interestingly, the number of shots has only a minor impact on the accuracy.



(a) Example 1: Addictive Content



(b) Example 2: Clickbait Content

Figure 4: Illustrative examples of few-shot ($k = 8$) selection for two harm categories: (a) *addictive gambling* and (b) *financial clickbait*. Colors show how demonstrations cover salient aspects of the samples.

Multimodal Harm Detection. Lastly, we systematically assess whether (and the extent to which) additional improvements can be gained from the integration of both text and visual (i.e. *multimodal*) input. To this end, we utilize our FS-ICL-CG and FS-ICL-DII approaches under various shot configurations. For FS-ICL-CG, we generate captions for the video thumbnail to aid in the classification process. As seen in Figure 5, despite this additional input FS-ICL-CG does not improve upon the performance of FS-ICL, and often leads to a reduction in performance (e.g., FS-ICL: 78.13% vs. FS-ICL-CG: 73.7% for GPT-4o-Mini

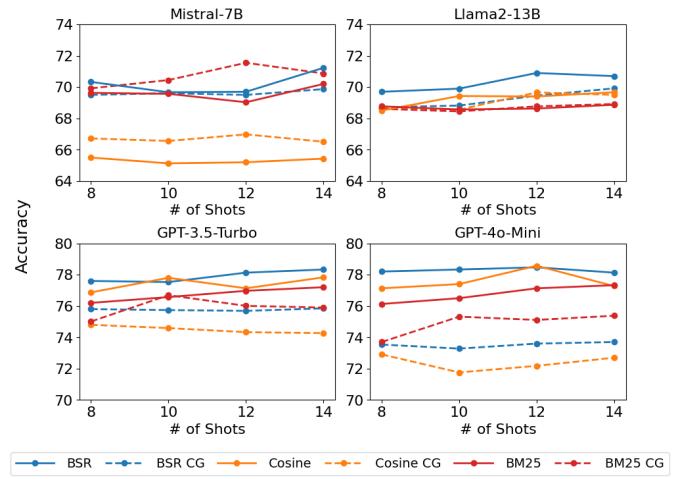


Figure 5: Comparison of FS-ICL and FS-ICL-CG approaches of different models across different shot numbers and selection methods. FS-ICL-CG does not improve upon FS-ICL, indicating that inclusion of captions does not consistently enhance performance.

using BSR on a 14-shot setting). This might be due to the generated captions not being able to capture subtle harm information present in the vision domain. Additional results, including other metrics, are provided in Table 3.

Next, we utilize the FS-ICL-DII approach where we provide visual input directly to natively multimodal LLMs. We consider the OpenFlamingo, LLaVA, GPT-4o-Mini LLMs. The results for these are shown in Table 4. We see that GPT-4o-Mini in the 14-shot configuration achieves the best performance, outperforming all other models across various settings and attaining an accuracy of 80%. For GPT-4o-Mini specifically, it is clear that the LLM benefits greatly from the additional multimodal examples provided. However, this is not the case for LLaVa and OpenFlamingo. Performance is fairly low throughout ($\approx 50\%$), even after few-shot ICL in OpenFlamingo (LLaVa does not directly support ICL so results are omitted). Hence, the choice of the LLM is very important for multimodal learning, especially in the context of harm classification. We provide additional results comparing performance using various metrics as shown in Table 5

The performance of various models across different con-

Table 3: FSL-CG performance comparison for models using various metrics (%).

Model	Accuracy	Precision	Recall	F-1
GPT-3.5-turbo	76.69	77.39	77.14	76.67
GPT-4o-mini	75.38	76.21	75.87	75.35
Mistral-7B	71.55	73.08	70.62	70.41
Llama2-13B	69.92	72.41	70.82	69.58

Table 4: Accuracy (%) of LLMs for varying the number of shots for the FS-ICL-DII approaches. N/A denotes lack of native support for few-shot learning.

Model	Shots				
	0	8	10	12	14
OpenFlamingo	50	50	52	54	54
LLaVa VLM	54	N/A	N/A	N/A	N/A
GPT-4o-mini	70	79	79	79	80

figurations is summarized in the Appendix G, H and tables 15 16, showcasing their effectiveness on the FSL tasks. These results highlight key metrics such as accuracy, precision, recall, and F-1 scores, providing a comparative analysis of the models. For a detailed breakdown and additional results, refer to the Appendix G, H, where we present extended tables on model configurations and performance trends.

5 Discussion

LLM Performance Across Other Harm Classification Datasets. To ensure that LLM performance at harm classification is not localized to the YouTube videos dataset (Jo, Wesołowska, and Wojcieszak 2024), we experiment on two additional harm-specific datasets: D-Lab Hate Speech (D-Lab 2022) and Jigsaw Toxicity (Google/Jigsaw 2019). These datasets do not consider diverse harm categories as the YouTube dataset (i.e. focus solely on hate speech and toxicity). As the GPT-4o-Mini and GPT-3.5-Turbo models were the highest performers in prior experiments, we use these to undertake experiments on the additional datasets. These results are shown in Appendices D and E. We find that the LLMs outperform the other baselines on these benchmarks as well, highlighting their use as content moderators.

Exemplar Reordering and Balanced Selection. Since ICL and few-shot learning lead to greatly improved performance in harm classification, we undertake further ablations to analyze how exemplar order and balanced selection affect model performance. As exemplar order has been shown to influence performance (Lu et al. 2021), we first undertake two experiments for exemplar selection (using cosine similarity as the selector): (1) reordering exemplars based on their instance-level metric, which means the most similar demonstration lists closest to the content sample, and (2) listing exemplars in the prompt as selected by the coverage-based metric, without reordering. Figure 6 details the results. As seen, the GPT models benefit from reordering but this is

Table 5: FSL-DII Best Performers on multimodal classification using various metrics (%).

Model	Accuracy	Precision	Recall	F-1
OpenFlamingo	54.0	62.5	69.0	65.5
LLaVa VLM	54.0	59.5	62.5	60.5
GPT-4o-mini	80.0	82.0	80.0	81.0

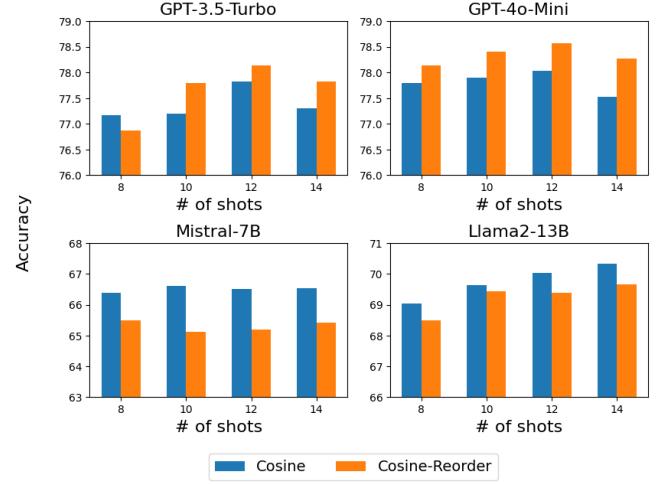


Figure 6: Impact of reordering on accuracy (%).

not the case for Mistral and Llama, which prefer the original ordering based on the selection metric. Thus, it is important to reorder depending on the model used. Next, we conduct two experiments with exemplar selection: (1) explicitly balancing the number of examples from each class (*balanced* selection), and (2) selecting examples based solely on a coverage-based selector metric, regardless of class distribution (*imbalanced* selection) and using the Mistral-7B and GPT-4o-Mini models, to show the impact of class-balanced selection. Figure 7 shows that imbalanced selection generally results in better performance across the board.

Utilizing More Descriptive Prompts. We also compare the performance effect of using more descriptive prompts. Currently, our prompts do not include specific definitions of harm categories (Jo, Wesołowska, and Wojcieszak 2024), so we include these and run experiments on the YouTube dataset with GPT-4o-Mini. The detailed prompts and results are provided in Appendix F. We find the more detailed prompts do not impact LLM performance much and the best gain is only 0.37%. Hence, it might be better to use more concise prompts and under-utilize the context, thereby enabling the model to understand the definition of harm via FS-ICL.

6 Conclusion

We study online harm classification (or content moderation) problem where approaches need to be highly effective while requiring minimal supervision (i.e. annotated data) and should adapt to the ever-changing dynamic nature of

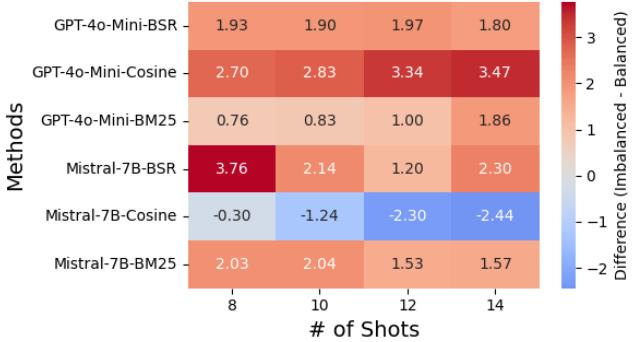


Figure 7: Accuracy difference (%) between *balanced* and *imbalanced* demonstration selection for GPT-4o-Mini and Mistral-7B across varying shots and selection methods. Red denotes *imbalanced* > *balanced*, while blue denotes the converse.

harmful content. We employ LLMs for this task and demonstrate that LLM performance in the few-shot in-context learning setting (i.e. minimally supervised + dynamic) can near domain expert performance. We utilize open-source and relatively inexpensive closed-source LLMs across different size scales (Mistral-7B, Llama2-13B, GPT-3.5-Turbo, GPT-4o-Mini) to ensure that the approach can scale to actual moderation of social media content. Further, we analyze LLMs in the multimodal setting and find that performance can be greatly improved, but only for closed-source multimodal LLMs such as GPT-4o-Mini and not smaller open-source variants (LLaVa and OpenFlamingo). Our findings underscore the the benefits of LLMs as a better alternative to proprietary baselines (Perspective and OpenAI Moderation APIs) and other deep learning few-shot baselines (e.g. Prototypical Networks) for harm identification.

7 Limitations

Despite the promising results, there may be several challenges in using LLMs and VLMs for content moderation. One is the computational cost associated with processing *multimodal* data in approaches such as FS-ICL-DII. Additionally, supplying multimodal inputs is generally slower compared to text-only inputs, with inference times for multimodal LLMs ranging between 10-30 seconds per test sample. However, with the development of faster inference pipelines for multimodal input as well as improved multimodal LLMs (and associated frameworks), we posit that this issue will be obviated in the future. We also restrict our analyses to social media content in English, but it is important to extend these efforts to other languages, especially low-resource ones with limited data available.

8 Ethics Statement

Our work on employing LLMs to detect harmful content in YouTube videos is a crucial step toward maintaining a healthy online environment. Through experiments on three diverse datasets and four LLMs, we demonstrate the effectiveness and scalability of few-shot learning for this task. The results from three different multimodal LLMs underscore the potential of multimodal models to enhance content

moderation. We hope to inspire further research into more reliable and impactful content moderation strategies using LLMs to ensure safer and more positive social media platforms. We are also committed to ensuring the reproducibility of our ideas and methods. Our code and implementation details are provided in Appendix I.

References

- Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; Ring, R.; Rutherford, E.; Cabi, S.; Han, T.; Gong, Z.; Samangooei, S.; Monteiro, M.; Menick, J.; Borgeaud, S.; Brock, A.; Nematzadeh, A.; Sharifzadeh, S.; Binkowski, M.; Barreira, R.; Vinyals, O.; Zisserman, A.; and Simonyan, K. 2022. Flamingo: a Visual Language Model for Few-Shot Learning. *ArXiv*, abs/2204.14198.
- Aldwairi, M.; and Alwahedi, A. 2018. Detecting fake news in social media networks. *Procedia Computer Science*, 141: 215–222.
- Arora, A.; Nakov, P.; Hardalov, M.; Sarwar, S. M.; Nayak, V.; Dinkov, Y.; Zlatkova, D.; Dent, K.; Bhatawdekar, A.; Bouchard, G.; et al. 2023. Detecting harmful content on online platforms: what platforms need vs. where research efforts go. *ACM Computing Surveys*, 56(3): 1–17.
- Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; and Zhou, J. 2023. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. *arXiv preprint arXiv:2308.12966*.
- Bandy, J. 2021. Problematic machine behavior: A systematic literature review of algorithm audits. *Proceedings of the ACM on human-computer interaction*, 5(CSCW1): 1–34.
- Baruah, A.; Das, K.; Barbhuiya, F.; and Dey, K. 2020. Context-aware sarcasm detection using BERT. In *Proceedings of the Second Workshop on Figurative Language Processing*, 83–87.
- Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language Models are Few-Shot Learners. *arXiv preprint arXiv:2005.14165*.
- Cinelli, M.; Quattrociocchi, W.; Galeazzi, A.; Valensise, C. M.; Brugnoli, E.; Schmidt, A. L.; Zola, P.; Zollo, F.; and Scala, A. 2020. The COVID-19 social media infodemic. *Scientific Reports*, 10(1): 16598.
- D-Lab, U. B. 2022. Measuring Hate Speech Dataset. <https://huggingface.co/datasets/ucberkeley-dlab/measuring-hate-speech>. Accessed: 2024-08-26.
- Dacon, J.; Shomer, H.; Crum-Dacon, S.; and Tang, J. 2022. Detecting Harmful Online Conversational Content towards LGBTQIA+ Individuals. *arXiv preprint arXiv:2207.10032*.
- Del Vigna12, F.; Cimino23, A.; Dell’Orletta, F.; Petrocchi, M.; and Tesconi, M. 2017. Hate me, hate me not: Hate speech detection on facebook. In *Proceedings of the first Italian conference on cybersecurity (ITASEC17)*, 86–95.
- Dong, Q.; Li, L.; Dai, D.; Zheng, C.; Wu, Z.; Chang, B.; Sun, X.; Xu, J.; and Sui, Z. 2022. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*.

- Facebook. 2024. How Does Facebook Moderate Content? Accessed: 2024-08-26.
- FORCE11. 2020. The FAIR Data principles. <https://force11.org/info/the-fair-data-principles/>.
- Fortuna, P.; and Nunes, S. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4): 1–30.
- Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.
- General, U. S. 2023. Social Media and Youth Mental Health: The U.S. Surgeon General’s Advisory. <https://www.hhs.gov/sites/default/files/sg-youth-mental-health-social-media-advisory.pdf/>.
- Gillespie, T. 2020. Content moderation, AI, and the question of scale. *Big Data & Society*, 7(2): 2053951720943234.
- Google/Jigsaw. 2019. Jigsaw Unintended Bias in Toxicity Classification. <https://www.kaggle.com/competitions/jigsaw-unintended-bias-in-toxicity-classification>. Accessed: 2024-08-26.
- Gupta, S.; Gardner, M.; and Singh, S. 2023. Coverage-based example selection for in-context learning. *arXiv preprint arXiv:2305.14907*.
- Haidt, J.; and Twenge, J. 2023. Social media and mental health: A collaborative review. *Unpublished manuscript, New York university*. Accessed at tinyurl.com/SocialMediaMentalHealthReview.
- Haroon, M.; Wojcieszak, M.; Chhabra, A.; Liu, X.; Mohapatra, P.; and Shafiq, Z. 2023. Auditing YouTube’s recommendation system for ideologically congenial, extreme, and problematic recommendations. *Proceedings of the National Academy of Sciences*, 120(50): e2213020120.
- He, P.; Liu, X.; Gao, J.; and Chen, W. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.
- Hilbert, M.; Cingel, D. P.; Zhang, J.; Vigil, S. L.; Shawcroft, J.; Xue, H.; Thakur, A.; and Shafiq, Z. 2023. # BigTech@Minors: Social Media Algorithms Personalize Minors’ Content After a Single Session, but Not for Their Protection. Available at SSRN 4674573.
- Hilbert, M.; Thakur, A.; Flores, P. M.; Zhang, X.; Bhan, J. Y.; Bernhard, P.; and Ji, F. 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.
- Hossain, M. S.; and Muhammad, G. 2019. Emotion recognition using deep learning approach from audio–visual emotional big data. *Information Fusion*, 49: 69–78.
- Islam, M. R.; Liu, S.; Wang, X.; and Xu, G. 2020. Deep learning for misinformation detection on online social networks: a survey and new perspectives. *Social Network Analysis and Mining*, 10(1): 82.
- Jigsaw. 2023. Perspective API. <https://www.perspectiveapi.com/>. Accessed: July 18 2024.
- Jo, C. W.; Wesołowska, M.; and Wojcieszak, M. 2024. Harmful YouTube Video Detection: A Taxonomy of Online Harm and MLLMs (GPT-4-Turbo) as Alternative Annotators. osf.io/2dn8s/.
- Jones, K. S.; Walker, S.; and Robertson, S. E. 2000. A probabilistic model of information retrieval: development and comparative experiments: Part 2. *Information processing & management*, 36(6): 809–840.
- Kosinski, M.; Stillwell, D.; and Graepel, T. 2013. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the national academy of sciences*, 110(15): 5802–5805.
- Kwangho Song, Y.-S. K. 2020. An Enhanced Multimodal Stacking Scheme for Online Pornographic Content Detection. *Applied Sciences*.
- Li, J.; Selvaraju, R. V.; Gotmare, A.; Joty, S.; Xiong, C.; and Hoi, S. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *Proceedings of the 39th International Conference on Machine Learning*, 12888–12900.
- Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26296–26306.
- Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual Instruction Tuning.
- Lu, Y.; Bartolo, M.; Moore, A.; Riedel, S.; and Stenetorp, P. 2021. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. *arXiv preprint arXiv:2104.08786*.
- Mundrievskaya, Y. O.; Matsuta, V. V.; Serbina, G. N.; and Peshkovskaya, A. G. 2023. Online Harmful Content: A Study on Violent Fan Communities on Social Media. In *Complex Social Systems in Dynamic Environments: Advanced Theories, Innovative Methods, and Interdisciplinary Research Results*, 175–182. Springer.
- Munger, K.; and Phillips, J. 2022. Right-wing YouTube: A supply and demand perspective. *The International Journal of Press/Politics*, 27(1): 186–219.
- Nguyen, T. T.; Wilson, C.; and Dalins, J. 2023. Fine-tuning llama 2 large language models for detecting online sexual predatory chats and abusive texts. *arXiv preprint arXiv:2308.14683*.
- OpenAI. 2023. OpenAI Moderation API. <https://platform.openai.com/docs/guides/moderation/quickstart>. Accessed: July 18 2024.
- OpenAI. 2024. GPT-4o Mini. <https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/>. Accessed: July 18 2024.
- Quiñonero-Candela, J.; Sugiyama, M.; Schwaighofer, A.; and Lawrence, N. D. 2022. *Dataset shift in machine learning*. Mit Press.
- Radford, A.; Kim, J. W.; Hallacy, M.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. *arXiv preprint arXiv:2103.00020*.

- Reimers, N.; and Gurevych, I. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.
- Robertson, S. E.; Walker, S.; Jones, S.; Hancock-Beaulieu, M. M.; Gatford, M.; et al. 1995. Okapi at TREC-3. *Nist Special Publication Sp*, 109: 109.
- Santra, P. 2024. Leveraging LLMs for Detecting and Modeling the Propagation of Misinformation in Social Networks. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 3073–3073.
- Sekharan, C. N.; and Vuppala, P. S. 2023. Fine-Tuned Large Language Models for Improved Clickbait Title Detection. In *2023 Congress in Computer Science, Computer Engineering, & Applied Computing (CSCE)*, 215–220. IEEE.
- Shen, S.; Zhang, W.; Gao, J.; Wang, Z.; Tan, M.; Li, S.; Li, X.; and Tan, V. 2022. How Much Can CLIP Benefit Vision-and-Language Tasks? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2704–2713.
- Snell, J.; Swersky, K.; and Zemel, R. S. 2017. Prototypical Networks for Few-shot Learning. *arXiv preprint arXiv:1703.05175*.
- Twitter. 2023. Twitter Community Notes. <https://communitynotes.x.com/guide/en/about/introduction>.
- Vinyals, O.; Blundell, C.; Lillicrap, T.; Kavukcuoglu, K.; and Wierstra, D. 2016. Matching Networks for One Shot Learning. *Proceedings of the 30th Conference on Neural Information Processing Systems (NIPS 2016)*, 3630–3638.
- Vosoughi, S.; Roy, D.; and Aral, S. 2018. The spread of true and false news online. *Science*, 359(6380): 1146–1151.
- Wang, H.; Tian, Z.; Song, X.; Zhang, Y.; Pan, Y.; Tu, H.; Huang, M.; and Zhou, B. 2024. Intent-Aware and Hate-Mitigating Counterspeech Generation via Dual-Discriminator Guided LLMs. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 9131–9142.
- Wang, Y.; and Kosinski, M. 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.
- 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/>.
- Yang, X.; Wu, Y.; Yang, M.; Chen, H.; and Geng, X. 2024. Exploring diverse in-context configurations for image captioning. *Advances in Neural Information Processing Systems*, 36.
- Yin, W.; Hay, M.; and Roth, D. 2019. Benchmarking Zero-shot Text Classification: Datasets, Evaluation and Entailment Approach. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3914–3923.
- YouTube. 2021. YouTube Community Guidelines. <https://www.youtube.com/howyoutubeworks/policies/community-guidelines/>.
- Youyou, W.; Kosinski, M.; and Stillwell, D. 2015. Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences*, 112(4): 1036–1040.
- Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

AAAI ICWSM Paper Checklist

1. For most authors...
 - (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? **Yes**
 - (b) Do your main claims in the abstract and introduction accurately reflect the paper’s contributions and scope? **Yes**
 - (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? **Yes**
 - (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? **NA**
 - (e) Did you describe the limitations of your work? **Yes**
 - (f) Did you discuss any potential negative societal impacts of your work? **Yes**
 - (g) Did you discuss any potential misuse of your work? **No; our work deals with harm detection on social media. As a result, we do not believe it has any avenues for misuse.**
 - (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? **Yes**
 - (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? **Yes**
2. Additionally, if your study involves hypotheses testing...
 - (a) Did you clearly state the assumptions underlying all theoretical results? **NA**
 - (b) Have you provided justifications for all theoretical results? **NA**
 - (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? **NA**
 - (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? **NA**
 - (e) Did you address potential biases or limitations in your theoretical framework? **NA**
 - (f) Have you related your theoretical results to the existing literature in social science? **NA**

- (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? **NA**
3. Additionally, if you are including theoretical proofs...
- Did you state the full set of assumptions of all theoretical results? **NA**
 - Did you include complete proofs of all theoretical results? **NA**
4. Additionally, if you ran machine learning experiments...
- Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? **Yes**
 - Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? **Yes**
 - Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? **No**
 - Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? **NA**
 - Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? **Yes**
 - Do you discuss what is “the cost” of misclassification and fault (in)tolerance? **No**
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
- If your work uses existing assets, did you cite the creators? **Yes**
 - Did you mention the license of the assets? **No**
 - Did you include any new assets in the supplemental material or as a URL? **No**
 - Did you discuss whether and how consent was obtained from people whose data you’re using/curating? **Yes**
 - Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? **Yes**
 - If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see FORCE11 (2020))? **NA**
 - If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Gebru et al. (2021))? **NA**
6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...
- Did you include the full text of instructions given to participants and screenshots? **NA**
 - Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? **NA**
 - Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? **NA**

- Did you discuss how data is stored, shared, and de-identified? **NA**

Appendix

A Datasets

A.1 YouTube Harms Dataset

We leverage a multimodal dataset (Jo, Wesołowska, and Wojcieszak 2024), comprising of YouTube videos from a variety of harm categories. The data is categorized online harm into six types: information harms, hate and harassment harms, clickbaitive harms, addictive harms, sexual harms, and physical harms. Videos were labeled by crowdworkers (from Amazon MTurk), fine-tuned LLMs and domain experts trained in the social sciences, specifically with a focus on communication and digital media. Each video was classified harmful or harmless, and assigned one or more harm categories if classified as harmful. The final dataset was formed by filtering for the videos where there was agreement between the crowdworker and domain expert, resulting in 19000 labeled videos.

A.2 Measuring Hate Speech Corpus (D-Lab)

The Berkeley D-Lab Measuring Hate Speech Corpus (D-Lab 2022) is a labeled dataset curated to facilitate the study and measurement of hate speech in online platforms. The dataset is composed of 6,954 social media posts, which have been manually annotated for the presence of hate speech. The corpus is derived from Twitter data collected from a period between January 2015 and June 2018. Each post has been labeled according to three key categories: (i) hate speech, (ii) offensive but not hate speech, and (iii) neither offensive nor hate speech. The labeling process was conducted by multiple annotators, ensuring inter-annotator agreement and consistent classification standards. The dataset aims to capture a wide range of hate speech expressions, including slurs, dehumanizing language, and threats, across various contexts. Each tweet was annotated by multiple annotators.

A.3 Jigsaw Toxicity Dataset

The Jigsaw Toxicity Classification Dataset (Google/Jigsaw 2019) was developed as part of a Kaggle competition in 2017, with the goal of developing and evaluating machine learning models to detect toxic comments online. This dataset contains a large collection of Wikipedia comments, each labeled by human raters for various types of toxic behavior (toxic, severe toxic, obscene, threat, insult, identity hate). In our experiments, we retrieve the comment text by the comment ID and convert the probability scores for the various toxicity types into a single binary label: Harmful or Harmless.

B Selectors

B.1 Cosine

Cosine similarity is a similarity measure between two embedding vectors. We utilize the *Sentence Transformers (SBERT)* library (Reimers and Gurevych 2019) with the all-mpnet-base-v2 model from Hugging Face to generate

sentence embeddings. The cosine similarity between each demonstration and the content sample is then calculated as the dot product of their embeddings, divided by the product of their magnitudes:

$$\text{cosine similarity} = \frac{Q \cdot D}{\|Q\| \|D\|}$$

Here, Q is the embedding of the content sample and D is the embedding of the demonstration.

B.2 BM25

BM25 (Best Matching 25) is a ranking function commonly used in information retrieval systems to rank documents based on their relevance to a given query. Developed as a part of the Okapi system, BM25 considered term frequency saturation and document length normalization, has been tested as an effective information retrieval method. (Robertson et al. 1995; Jones, Walker, and Robertson 2000) We use the BM25Okapi method from the `rank_bm25` library with the syntactic structure of size-4 n-grams. The BM25 score between each demonstration and the content sample is calculated as follows:

$$\text{BM25}(Q, D) = \sum_{i=1}^n \text{IDF}(q_i) \cdot \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot (1 - b + b \cdot \frac{|D|}{\text{avgdl}})}$$

Here, Q is the content sample and D is the demonstration. $f(q_i, D)$ is the frequency of the n-gram structural element q_i of the content sample in the demonstration candidates. $|D|$ is the length of the demonstration. k_1 and b are hyperparameters of the BM25 Okapi class that are set to 1.5 and 0.75 respectively. The avgdl parameter is the average demonstration length, and IDF is the weight of the structural element q_i .

B.3 BERTScore

BERTScore measures the similarity between two sentences by computing the similarity of each token of the candidate sentence and each token of the reference sentence by cosine similarity. Unlike cosine similarity which compares two sentences directly by the embedding vectors, BERTScore brings the comparison process to the token level. We use bert_score library (Zhang et al. 2019) with the deberta-large-mnli model(He et al. 2020) to generate token embeddings. In our experiments, we use only recall as the BERTScore metric and do not apply IDF weighting. We calculate BERTScore-Recall (BSR) between a content sample and a demonstration by first computing the cosine similarity between the token embeddings of both and then maximizing the aggregated cosine similarity score. This can be represented as:

$$\text{BSR}(Q, D) = \sum_{i=1}^{|Q|} \max_j \text{cosine similarity}(Q_i, D_j)$$

Here, $|Q|$ is the length of the content sample, Q_i is the embedding of the i -th token of the content sample, and D_j is the embedding of the j -th token of the demonstration.

C Prompts

C.1 ZSL Prompt

In the ZSL approach, each prompt consists of the task instruction for harm detection and a single content sample to

be classified. The format of these prompts varies depending on the model used. In the absence of relevant examples that offer context or a desired output format, a more restrictive prompt is employed to guide the output.

Mistral-7B Your task is to classify YouTube videos as Harmful or Harmless based on their metadata.

Title: {title}

Classification:

Llama2-13B <*s*>[INST]<<SYS>>Your task is to evaluate YouTube videos as Harmful or Harmless based on their metadata. You must respond with only one word: "Harmful" or "Harmless," without any additional commentary or explanation. </></SYS>>

Title: {title}

Classification: [/INST]

GPT-3.5-Turbo {'role': 'user','content': Your task is to classify YouTube videos as Harmful or Harmless based on their metadata. You must respond with only one word: "Harmful" or "Harmless" without any additional commentary or explanation.'}, {'role': 'user', 'content': 'Title: {title}\nClassification: '}

GPT-4o-Mini {'role': 'user','content': Your task is to classify YouTube videos as Harmful or Harmless based on their metadata. You must respond with only one word: "Harmful" or "Harmless" without any additional commentary or explanation.'}, {'role': 'user', 'content': 'Title: {title}\nClassification: '}

LLaVa \"USER:<image>\\"\\nAnalyze the given image carefully. Identify any potential risks, safety concerns, sexual explicit content, signs of violence or aggression, indications of illegal activities, health hazards, environmental hazards, or presence of weapons or dangerous objects based on visible details. Consider factors such as the context, setting, activities, expressions, and objects present in the image. Analyze and assess the environment and any notable elements. Identify any potential safety hazards or risks present. Note the presence and actions of people, if any. Check if there is any kind of violence. Finally, classify the overall image context as either 'harmful' if there is any indication of danger, or 'harmless' if it appears safe with no visible threats. Provide your classification as either 'harmful' or 'harmless' and do not give any extra text but the classification.\\"\\nASSISTANT:\\n"

OpenFlamingo f\"<image>An image of a close up of two women wrestling in a cage. Classification:harmful<|endofchunk|><image>An image of several people standing around a patio with a table and chairs. Classification:harmless<|endofchunk|><image>An image of a close up of a woman in a short skirt and heels. Classification:harmful<|endofchunk|><image>An image of a man in a headset is talking into a microphone. Classification:harmless<|endofchunk|><image>An image of a person looking at a city with a giant explosion. Classification:harmful<|endofchunk|><image>An image of three women sitting at a table with a variety of

food. Classification:harmless<|endofchunk|><image>An image of a man riding a bike on a block of concrete. Classification:harmful<|endofchunk|><image>An image of a man in a blue shirt and jeans standing in a parking lot. Classification:harmless<|endofchunk|><image>{caption}. Classification:\”

GPT-4o-Mini (Multimodal) `{"role": "system", "content": "You are a helpful image classification assistant. You must classify given image analyzing the image and Title whether it is harmful or harmless."}`
`{"role": "user", "content": "type": "text", "text": "You are a helpful image classification assistant. You must classify given image whether it is harmful or harmless analyzing the image and Title:Title. Harmful: This category includes images depicting sexuality, gambling, violence, weapons, click bait, spam, harmful, or inappropriate content. Examples might be scenes of physical violence, display of sexual affection, weapons, sharp objects, explicit content, spam content, gambling, click bait content or dangerous situations. Harmless: This category includes images that are non-violent, safe, and appropriate. Examples might be everyday scenes, landscapes, or social gatherings without harmful context. If you are not sure, pick whatever label you think is more probable based on your analysis. Give the output in one word if is is harmful or harmless. Do not explain anything."}`

C.2 FS-ICL Prompt

As shown in Table 6, the FS-ICL prompts include a few examples in addition to the ZSL template components. The examples consist of video metadata (title) paired with their corresponding classification.

C.3 FS-ICL-CG Prompt

Based on the FS-ICL approach, the FS-ICL-CG prompt integrates image captions associated with the examples and the content sample, as illustrated in Table 7.

C.4 FS-ICL-DII Prompt

Our FS-ICL-DII approach provides both text (Title) and vision (Direct Image Input) data of a video for each demonstration and content sample, as shown in Table 8.

Table 9: Illustrative Examples from Jigsaw Toxicity Dataset.

Comment Text	Label
Go away disgusting troll.	Harmful
It was fun watching him play basketball. Looking forward to seeing him play football, this fall!	Harmless

Table 10: Accuracy (%) of GPT-4o-Mini and GPT-3.5-turbo across different selectors (BM25, Cosine, BSR) for ZSL and 8-Shot FS-ICL on the Jigsaw Toxicity Dataset.

Models	Selectors	ZSL	FS-ICL
GPT-4o-Mini	BM25	75.20	73.30
	Cosine	75.27	73.47
	BSR	75.37	72.73
GPT-3.5-Turbo	BM25	65.40	71.13
	Cosine	65.23	69.70
	BSR	65.80	72.20

Table 11: Illustrative Examples from Measuring Hate Speech Dataset.

Comment Text	Label
Women like this need to be jailed and fined and locked up and sterilized with all their possessions taken away and trashed and their bank accounts emptied.	Hate
blasian women are always gorgeous and she definitely supports that	Benign

Table 12: Accuracy (%) of GPT-4o-Mini and GPT-3.5-turbo across different selectors (BM25, Cosine, BertScore) for ZSL and 8-Shot FS-ICL on the Hate Speech Dataset.

Models	Selectors	ZSL	FS-ICL
GPT-4o-Mini	BM25	77.10	94.03
	Cosine	77.10	93.97
	BertScore	77.10	94.37
GPT-3.5-Turbo	BM25	91.60	92.40
	Cosine	91.60	91.67
	BertScore	91.60	90.90

D Additional Results for Jigsaw Toxicity Dataset

We provide a few illustrative examples from the Jigsaw Toxicity dataset in Table 9. We conduct experiments for the 8-shot setting with the two GPT models and the open-source few-shot learning baselines. The GPT results are shown in Table 10. Prototypical Network achieves an 8-shot performance of 60.98% and Matching Network achieves an accuracy of 62.94%. Compared to them, both GPT-3.5-Turbo and GPT-4o-Mini attain stellar performance in both the zero-shot and 8-shot setting, with GPT-4o-Mini performance being the highest at 75.37%.

Table 6: FS-ICL Prompt Templates used for each model in harm detection (4-shot example).

Model	Prompt
Mistral-7B	<p>Your task is to classify YouTube videos as Harmful or Harmless based on their metadata.</p> <p>Title: {title}</p> <p>Classification: {classification}</p> <p>Title: {title}</p> <p>Classification: {classification}</p> <p>Title: {title}</p> <p>Classification: {classification}</p> <p>Title: {title}</p> <p>Classification: {classification}</p> <p>Title: {title}</p> <p>Classification:</p>
Llama2-13B	<p><s>[INST] <<SYS>> Your task is to classify YouTube videos as Harmful or Harmless based on their metadata. You must respond with only one word: "Harmful" or "Harmless" without any additional commentary or explanation.</p> <p><</SYS>></p> <p>Title: {title}</p> <p>Classification: {classification}[/INST]</s><s>[INST] Title: {title}</p> <p>Classification: {classification}[/INST]</s><s>[INST] Title: {title}</p> <p>Classification: {classification}[/INST]</s><s>[INST] Title: {title}</p> <p>Classification: {classification}[/INST]</s><s>[INST] Title: {title}</p> <p>Classification: [/INST]</p>
GPT-3.5-Turbo	<pre>'role': 'user', 'content': 'Your task is to classify YouTube videos as Harmful or Harmless based on their metadata.'}, {'role': 'user', 'content': 'Title: {title}\nClassification:'}, {'role': 'assistant', 'content': '{classification}'}, {'role': 'user', 'content': 'Title: {title}\nClassification:'}, {'role': 'assistant', 'content': '{classification}'}, {'role': 'user', 'content': 'Title: {title}\nClassification:'}, {'role': 'assistant', 'content': '{classification}'}, {'role': 'user', 'content': 'Title: {title}\nClassification:'}, {'role': 'assistant', 'content': '{classification}'}, {'role': 'user', 'content': 'Title: {title}\nClassification:'}</pre>
GPT-4o-Mini	<pre>'role': 'user', 'content': 'Your task is to classify YouTube videos as Harmful or Harmless based on their metadata.'}, {'role': 'user', 'content': 'Title: {title}\nClassification:'}, {'role': 'assistant', 'content': '{classification}'}, {'role': 'user', 'content': 'Title: {title}\nClassification:'}, {'role': 'assistant', 'content': '{classification}'}, {'role': 'user', 'content': 'Title: {title}\nClassification:'}, {'role': 'assistant', 'content': '{classification}'}, {'role': 'user', 'content': 'Title: {title}\nClassification:'}, {'role': 'assistant', 'content': '{classification}'}, {'role': 'user', 'content': 'Title: {title}\nClassification:'}</pre>

Table 7: FS-ICL-CG Prompt Templates used for each model in harm classification(4-shot example).

Model	Prompt
Mistral-7B	<p>Your task is to classify YouTube videos as Harmful or Harmless based on their metadata.</p> <pre>Title: {title} Caption: {caption} Classification: {classification}</pre> <p>Title: {title}</p> <pre>Caption: {caption} Classification: {classification}</pre>
Llama2-13B	<p><s>[INST] <<SYS>> Your task is to classify YouTube videos as Harmful or Harmless based on their metadata. You must respond with only one word: "Harmful" or "Harmless" without any additional commentary or explanation.</p> <p><</SYS>></p> <pre>Title: {title} Classification: {classification}[/INST]</s><s>[INST] Title: {title} Caption: {caption} Classification: {classification}[/INST]</pre>
GPT-3.5-Turbo	<pre>{'role': 'user', 'content': 'Your task is to classify YouTube videos as Harmful or Harmless based on their metadata.'}, {'role': 'user', 'content': 'Title: {title} Caption: {caption}\nClassification:'} {'role': 'assistant', 'content': '{classification}'} {'role': 'user', 'content': 'Title: {title}\nClassification:'} {'role': 'assistant', 'content': '{classification}'} {'role': 'user', 'content': 'Title: {title}\nClassification:'} {'role': 'assistant', 'content': '{classification}'} {'role': 'user', 'content': 'Title: {title}\nClassification:'} {'role': 'assistant', 'content': '{classification}'} {'role': 'user', 'content': 'Title: {title}\nClassification:'}</pre>
GPT-4o-Mini	<pre>{'role': 'user', 'content': 'Your task is to classify YouTube videos as Harmful or Harmless based on their metadata.'}, {'role': 'user', 'content': 'Title: {title}\nClassification:'} {'role': 'assistant', 'content': '{classification}'} {'role': 'user', 'content': 'Title: {title}\nClassification:'} {'role': 'assistant', 'content': '{classification}'} {'role': 'user', 'content': 'Title: {title}\nClassification:'} {'role': 'assistant', 'content': '{classification}'} {'role': 'user', 'content': 'Title: {title}\nClassification:'} {'role': 'assistant', 'content': '{classification}'} {'role': 'user', 'content': 'Title: {title}\nClassification:'}</pre>

Table 8: FS-ICL-DII Prompt Templates used for each model in harm classification(4-shot example).

Model	Prompt
Llava	\"USER:<image>\nAnalyze the given image carefully. Identify any potential risks, safety concerns, sexual explicit content, signs of violence or aggression, indications of illegal activities, health hazards, environmental hazards, or presence of weapons or dangerous objects based on visible details. Consider factors such as the context, setting, activities, expressions, and objects present in the image. Analyze and assess the environment and any notable elements. Identify any potential safety hazards or risks present. Note the presence and actions of people, if any. Check if there is any kind of violence. Finally, classify the overall image context as either 'harmful' if there is any indication of danger, or 'harmless' if it appears safe with no visible threats. Provide your classification as either 'harmful' or 'harmless' and do not give any extra text but the classification.\nASSISTANT:\\"\n"
OpenFlamingo	f\"<image>{caption}. Classification:{classification}\n< endofchunk ><image>{caption}. Classification:\\"
GPT-4o-Mini M	{"role": "system", "content": "You are a helpful image classification assistant. You must classify given image analyzing the image and Title whether it is harmful or harmless."}\n{"role": "user", "content": "type": "text", "text": "You are a helpful image classification assistant. You must classify given image whether it is harmful or harmless analyzing the image and Title:{title}. Harmful: This category includes images depicting sexuality, gambling, violence, weapons, click bait, spam, harmful, or inappropriate content. Examples might be scenes of physical violence, display of sexual affection, weapons, sharp objects, explicit content, spam content, gambling, click bait content or dangerous situations. Harmless: This category includes images that are non-violent, safe, and appropriate. Examples might be everyday scenes, landscapes, or social gatherings without harmful context. If you are not sure, pick whatever label you think is more probable based on your analysis. Give the output in one word if is is harmful or harmless. Do not explain anything"}

E Additional Results for *Measuring Hate Speech Dataset*

We provide a few illustrative examples from the D-Lab Hate Speech dataset in Table 11. We conduct experiments for the 8-shot setting with the two GPT models and the open-source few-shot learning baselines. The GPT results are shown in Table 12. Prototypical Network achieves an 8-shot performance of 87.38% and Matching Network achieves an accuracy of 89.01%. Compared to them, both GPT-3.5-Turbo and GPT-4o-Mini attain stellar performance in both the zero-shot and 8-shot setting, with GPT-4o-Mini performance being the highest at 94.37%.

F Additional Results for More Descriptive Prompts

We conduct experiments with more detailed prompts to assess if this enhances performance. We include detailed definitions of the six harm categories, as outlined in Table 13. We evaluated the performance of GPT-4o-Mini with BSR using 8-Shot and 14-Shot configurations for both descriptive and original prompts. The results are shown in Table 14. Note that including the harm category definitions in the

prompts resulted in only slightly improved accuracy, with a best gain of 0.37%.

Table 14: Results of GPT-4o-Mini with BSR using 8-Shot and 14-Shot configurations, comparing prompts with and without harm categories definition.

Model	Shots	Prompt Type	Accuracy
GPT-4o-Mini	8	More Descriptive	78.57
	8	Original	78.20
	14	More Descriptive	78.30
	14	Original	78.13

G Additional Results for the Qwen-VL Multimodal LLM

To evaluate the effectiveness of using advanced captioning models for multimodal content moderation, we compared two state-of-the-art models: Qwen-VL and BLIP. As shown in Table below, the BertScore differences between the two models are minimal across multiple language models. Specifically, Qwen-VL achieved slightly higher scores

Table 13: Comparison of prompts with and without harm category definitions.

Prompt Composition	Prompt
With Definition	<p>Your task is to classify YouTube videos as Harmful or Harmless based on their metadata.</p> <p>Determine if a video fits into any of the harmful categories listed below. If it does, classify the video as 'Harmful'. If it does not, classify it as 'Harmless'.</p> <p>[Harmful Categories]</p> <ol style="list-style-type: none"> 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) Clickbait 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) <p>...</p>
Without Definition	<p>Your task is to classify YouTube videos as Harmful or Harmless based on their metadata.</p> <p>...</p>

Table 15: Results comparing the performance of various LLMs on captioning tasks using two different captioning models: Qwen-VL and BLIP. The performance metric used is BertScore, evaluated in an 8-shot setting for each captioning model. The LLMs evaluated include Mistral-7B, GPT-3.5-turbo, GPT-4.0-mini, and Llama2-13b-chat.

LLM	Metric	Qwen-VL (8-shot)	BLIP (8-shot)
Mistral-7B	BertScore	69.89	69.50
GPT-3.5-turbo	BertScore	76.92	75.80
GPT-4.0-mini	BertScore	75.42	73.54
Llama2-13b-chat	BertScore	69.12	68.71

in most cases, such as a BertScore of 76.92 with GPT-3.5-turbo, compared to 75.80 for BLIP as shown in table 15. However, both models demonstrate similar captioning performance, suggesting that the choice of captioning model may depend more on computational efficiency and integration requirements rather than accuracy alone.

H Additional Results for Other Metrics

We recorded the Precision, Recall and F1 score metrics for all the models we evaluated. Accuracy provides a broad measure of performance, but it is critical to evaluate the models' handling of false positives and false negatives, particularly in tasks like harm detection where misclassification can have significant consequences. The Metrics for best performer models are given in Table 16.

Table 16: Performance of FSL models on title classification using various metrics(%).

Model	Param	Accuracy	Precision	Recall	F-1
GPT-4o-mini	12 shot; Set-BSR	78.57	78.69	78.56	78.54
GPT-3.5-turbo	14 shot; Set-BSR	78.33	78.51	78.33	78.3
Mistral-7B	14 shot; Set-BSR	71.23	74.30	71.25	70.3
Llama2-13B	12 shot; Set-BSR	70.90	74.38	70.89	69.81
Prototypical Network (roberta-base)	12 shot	67.59	69.22	67.59	66.69
Matching Network (bert-base-uncased)	14 shot	67.38	69.09	67.38	65.85
Matching Network (roberta-base)	12 shot	66.03	67.49	66.03	64.4
Prototypical Network (bert-base-uncased)	12 shot	63.45	65.26	63.45	62.14

I Code and Reproducibility

We open-source our code in the following GitHub repository: <https://anonymous.4open.science/r/harm-detection-llm/>. The repository provides comprehensive instructions for reproducing our experiments and performing analyses on various settings. The FS-ICL experiments were run on an RTX A6000 GPU using CUDA version 11.8, while the FS-ICL-CG experiments were carried out on an A100 GPU.