Are foundational LLMs the next-generation social media content moderators?

Mohammad Nadeem, Shahab Saquib Sohail, Laeeba Javed, Erik Cambria, *Senior Member, IEEE,* and Amir Hussain, *Senior Member, IEEE,*

Abstract—The significant advancements in AI tools and systems' capabilities, reasoning, and efficiency are evident. A few noteworthy examples of such tools include Generative AI-based LLMs, such as ChatGPT-3.5, GPT-4, Gemini, and others. Foundational LLMs have proven versatile and effective in various tasks, such as composing poetry, writing code, generating essays, and solving puzzles, as they can effectively process text-based input. However, recent advancements have made them capable of handling multi-modal input such as text, image, and audio, making them highly general-purpose tools. They have also shown decent performance for pattern recognition tasks such as classification. Therefore, there is a curiosity about whether general-purpose LLMs can perform comparably or even superior to specialized deep-learning models (trained specifically for a given task). In our current study, we have compared the performance of fine-tuned deep-learning models with general-purpose LLM models for image-based content moderation. We trained deep learning models, namely CNN, ResNet50, and VGG-16 models, on an image dataset for violence detection and then tested their performance on a different test dataset. Subsequently, we subjected the same testing dataset to two vision-enabled LLMs (LLaVa and GPT-4). The $VGG16$ model emerged as the top performer and exhibited the highest accuracy of 94%, while Llava produced the lowest accuracy (66%). In the category of Large Language Models, GPT-4 performed the best with an accuracy of 92.42%. LLava LLM recorded the highest precision value among all models. The trend is similar to other performance metrics, such as recall and F1-score. However, GPT-4 performed best when compared to deep learning models with reduced training datasets. Overall, the LLMs did not surpass specialized models but achieved comparable performance, making them a good alternative when the available dataset is small.

Index Terms—Large language models, Deep learning, Generative artificial intelligence, Content Moderation

1 Introduction

Artificial intelligence (AI) has advanced significantly in recent years [1] with the development of large language models (LLMs) such as GPT-3.5 and GPT-4 from OpenAI, Gemini from Google, and Llama from Meta. Other notable LLMs include BlenderBot [2], Galactica, LLaMA from FAIR, Alpaca from Stanford, BloombergGPT, Chinchilla from DeepMind [3], and Palm [4]. The LLMs can process various types of data such as text, images, audio, and video [5], making them significant language processors. They are also revolutionizing the way machines interact with and comprehend human-generated content.

The amount of research analyzing different facets of AI technologies has increased dramatically since ChatGPT's launch [6]. As investigated in [7], a common pattern in the existing research is to identify how well LLM works in comparison to the state-of-the-art approaches on several issues including finance [8], medicine [9], healthcare [10], academic writing [11], sustainability [12], education [13], reasoning [14]and decision making [15]. In addition, the LLMs are being used to evaluate their performance on prediction tasks. For example, Patrinos et al. [16] used chatGPT to anticipate the future of personalized medicine.

LLMs exhibit human-like language creation and understanding abilities due to their deep neural network based architectures and extensive internet text training [17]. ChatGPT is widely used for tasks requiring contextual comprehension and excels at catching subtleties in language and context [18], [19]. Recent advances in LLMs have improved their capacity to handle many forms of input data, such as text, picture, audio, and video [5], [20]. The increase in LLM capabilities raises questions about their efficacy in vision-based applications. This research gap prompted us to look beyond language processing and analyze the proficiency of foundational LLMs on image recognition tasks.

Social media platforms allow people to connect globally, share opinions, and publish information. Their use has grown significantly due to the quick and easy access to information and the freedom to express in various formats. However, social media is witnessing an increase in harmful content. It includes hate speech, fake news, obscene and violent images, cyberbullying, child abuse content, etc. [21]. Therefore, it has become crucial to detect and moderate harmful content. It is also noteworthy that moderated content is not always provocative and inflammatory. If the content does not meet the platform's rules and policies, it is also flagged as inappropriate (for example, LinkedIn is

- *Mohammad Nadeem and Laeeba Javed are with the Department of Computer Science, Aligarh Muslim University, Aligarh, India. E-mail: mnadeem.cs@amu.ac.in*
- *Shahab Saquib Sohail is with the School of Computing Science and Engineering, VIT Bhopal University, Sehore, Kothrikalan, MP 466114, India. E-mail: shahabsaquibsohail@vitbhopal.ac.in*
- *Erik Cambria is with the NTU, Singapore. E-mail: cambria@ntu.edu.sg*
- *Amir Hussain is with Centre of AI and Robotics, Edinburgh Napier University, UK. E-mail: a.Hussain@napier.ac.uk*

Corresponding author: Shahab Saquib Sohail

not a platform for sharing personal pictures and gossip). Overall, content moderation involves systematically screening content on websites and online networks to decide if it is suitable for a specific site, location, or jurisdiction [22].

Deep learning, known for its hierarchical representation learning, has been instrumental in managing complicated patterns inside enormous datasets [23] that has led us to advances in image analysis [24], recognition [25], and comprehension [26]. Therefore, it has become vital to automatically detect harmful content on social media platforms and help human moderators flag problematic content. In the case of image-based violent detection, the deep learning algorithms excel in extracting features and recognizing patterns [27] to facilitate automatic identification of violent behaviors using body language and nonverbal [28]. However, incorporating large language models brings an interesting perspective since they have intrinsic language understanding skills that can augment the contextual comprehension of visual data [29].

While both LLMs and deep learning are quite competent at identifying content moderation tasks based on images, a full evaluation of their relative performances is necessary. An analysis of the available literature finds a significant gap in essential research into LLM performance on image datasets. As a result, we conducted a thorough study of the performance of two LLMs, Llava and ChatGPT-4, for detecting violent scenarios inside image datasets. We also performed comparative analysis with deep learning algorithms such as CNN, VGG16, and ResNet-50. Our study's key contribution is testing LLMs for content moderation and comparing deep learning algorithms comprehensively with GPT-4 and Llava. Our study also provides significant insights into their strengths and shortcomings, allowing researchers to choose the best approach for specific tasks. Moreover, understanding the trade-offs between LLMs and deep learning models is critical for improving their performance and successfully incorporating them into real-world applications.

In this regard, our main contributions to the study can be summarised as follows:

- Conducted a comparative analysis of deep learning models and LLMs for content moderation.
- We investigated how smaller training dataset sizes affected model performance.
- Investigated the interpretability of responses generated by LLMs.

The rest of the paper is organized as follows. Section 2 discusses the background of the study. Section 3 discusses all the models employed in the study, including fine-tuned deep-learning models and foundational LLMs. Section 4 describes the methodology followed to generate and compare the results of LLMs and deep-learning models. Section 5 presents the results and discussion of the current work. Section 6 presents the concluding remarks and future directions.

2 BACKGROUND

The proliferation of social media platforms and an increasing number of users necessitates the use of automated systems for content moderation to ensure that harmful materials such as hate speech, misinformation, and violent content are effectively identified and addressed [30], [31]. Therefore, to maintain the integrity of posts and adhere to legal standards, content moderation has attracted researchers and policy makers greatly. A number of works in this direction has been reported in the literature. Primarily, we are aimed at focusing a) technology used and b) issues addressed in the reported studies.

Up until recently, it's been all hands on deck for human moderators, who've been the go-to folks for dealing with harmful content. In [28], authors have argued that for effectively handling the social media contents for adequate moderations, human moderation are better as human can better deal with the requirements of contextual understanding and addressing unique challenges posed by. However, the sheer volume of content uploaded daily makes human moderation a highly labor-intensive and resource-consuming task. Additionally, human moderators are susceptible to biases, inconsistencies, and challenges keeping pace with the ever-evolving nature of online content [32].

To overcome these shortcomings, researchers have actively explored automating content moderation using artificial intelligence (AI). Traditional machine learning and deep learning models have emerged as a powerful tool, particularly adept at image and video classification tasks valuable for content moderation [33]. These models are trained on massive datasets labeled as containing harmful or safe content [34]. Once trained, they can then be used to automatically identify and flag potentially harmful content for further review by human moderators. Authors have achieved [35] state-of-the-art performance in detecting inappropriate content on platforms like YouTube using EfficientNet-BiLSTM architecture. Additionally, Moustafa [36] has used convolutional neural networks for tasks like pornography detection. These works set a benchmark that an be used for comparing the effectiveness of specialized deep learning models against any other models.

However, deep learning models also have limitations. They often require vast amounts of labeled data for effective training, which can be expensive and time-consuming to acquire [37]. Additionally, these models may struggle to generalize to new types of content not included in their training data [38]. This limitation paves the way for large language models (LLMs). LLMs can learn complex relationships between words and concepts, allowing them to perform various tasks such as generating text [7], translating languages [39], and writing different kinds of creative content [40]. Recent advancements have enabled LLMs to handle not just text but also multi-modal input such as images and audio, making them even more versatile [41].

The increasing interest in LLMs for social media content moderation is fueled by research demonstrating their potential in various aspects of this task. Research has indicated their potential effectiveness in several moderation roles. For example, LLMs have been shown to accelerate and enhance the accuracy of creating content for adult learning [41]. They also show promise in rule-based community moderation with noteworthy accuracy and precision [42]. Moreover, LLMs can improve the interaction between users and platforms, aiding in clearer communication [43].

LLMs show promise in altering webpages when given explicit directives, but they struggle with vague inputs and complex web structures, highlighting the need for further enhancements [41]. Research indicates that while LLM-based strategies are promising for content moderation, additional research and detailed implementation are necessary to refine these models for specific moderation tasks [44]. LLMs also perform well in analyzing social media sentiment, though ethical issues must be carefully considered [45]. Transitioning from basic capabilities to targeted applications, various studies have examined how LLMs can be integrated into the content moderation framework. For example, a study by [42] explores an initial approach to rule-based community moderation with LLMs, achieving median accuracy and precision rates of 64% and 83%, respectively. Their work highlights both the potential and the challenges of LLMs in comprehending and applying community standards. This study marks a significant point, illustrating the potential and limitations of LLMs in understanding and enforcing community guidelines.

Deep learning models have become valuable tools for content moderation, but their limitations necessitate exploring new approaches. Large language models (LLMs) offer exciting possibilities for the next generation due to their ability to handle complex text and potentially multi-modal data. However, concerns exist regarding the fairness of current LLM-based systems for vulnerable groups and minorities [3]. additionally, LLMs show promise, challenges such as handling complex requests and the need for meticulous data engineering for effective fine-tuning are major conerns to be dealt with [46], [47]. A comprehensive study directly comparing the performance of fine-tuned deep learning models and general-purpose LLM models for image-based classification, and consequently image content moderation, remains unexplored. To that end, we have conducted this study to address this gap by comparing the performance of fine-tuned deep learning models with general-purpose LLM models for image-based content moderation.

3 Models Employed

In this study, we used three deep-learning models and two LLMs. Each of them is discussed briefly in the next section.

3.1 Fine-tuned deep learning models

Two of the deep learning models were pre-trained, notably ResNet50 and VGG16. The remaining was a simple CNN-based model.

3.1.1 ResNet50

The ResNet architecture came into existence in 2016 to tackle the issue of vanishing gradients in deep neural networks [48]. In our study, we used the ResNet-50 model with pre-trained weights from the ImageNet dataset. We used transfer learning to use the ResNet-50 model's knowledge and feature extraction skills obtained from training on a large and varied dataset. We used the pre-trained ResNet-50 model as a feature extractor. We then built new layers to replace the final completely linked layer. Initially, a global average pooling layer was added to the ResNet-50 to reduce spatial dimensions and provide a more dense representation of the features learned by previous layers. A dense layer was added for the final classification with the number of units equal to the number of classes in our dataset (two classes). The activation function in the Dense layer used is 'sigmoid'. The model used the Adam optimizer and a binary cross-entropy loss function. Finally, the model was trained using the training data for 100 epochs with a batch size of 32.

TABLE 1
Layered architecture of CNN

Layer (type)	Output Shape	#Parameters
Conv2D_0	(None, 126, 126, 32)	896
MaxPooling2D_0	(None, 63, 63, 32)	0
Conv2D_1	(None, 61, 61, 64)	18,496
MaxPooling2D_1	(None, 30, 30, 64)	0
Conv2D_2	(None, 28, 28, 128)	73,856
MaxPooling2D_2	(None, 14, 14, 128)	0
Flatten_0	(None, 25088)	0
Dense_0	(None, 128)	3,211,392
Dense_1	(None, 2)	258
Total parameters: 3,304,898		
Trainable parameters: 3,304,898		
Non-trainable parameters: 0		

3.1.2 VGG16

The VGG16 architecture, developed in 2014, is a well-known deep convolutional neural network design [49]. It employs a composition rule in which numerous identical convolutional layers are stacked sequentially, followed by a maximum pooling layer to reduce the input dimensions. In our study, we used the pre-trained VGG16 model and excluded its top layers. We froze the basic model's layers to keep their weights constant during training. In addition, we introduced a flattened layer to prepare the data for final predictions in the last dense layer. The activation function used in the last dense layer was 'sigmoid'. Other hyper-parameters used for training were the 'Adam' optimizer, a binary cross-entropy loss function, a batch size of 32, and 100 epochs.

3.1.3 CNN

Convolutional neural networks (CNNs) are the prevailing neural network topologies for image categorization problems. Their structure consists of convolutional, pooling, flattening, and fully linked layers. Our CNN model consists of nine layers with three convolutional layers. The first convolutional layer used 32 filters to handle 128x128x3 input images, followed by 64 filters for the second and 128 for the third. Each convolutional layer used a 3x3 filter size and a ReLU activation function, followed by three Max Pooling layers, each with a pooling size 2x2. Following the convolutional layers, a single flatten layer reduced the final pooling layer's 2D feature mappings to a 1D feature vector. The flatten layer was followed by two dense layers. The first dense layer included 128 neurons and used ReLU activation. The second (output) dense layer had one neuron and used the sigmoid activation function. Table 1 depicts the architectural arrangement of CNN.

3.2 Foundational LLMs

For this study, we sought multi-modal LLMs that can interpret images and text (for query purposes). Considering the prerequisites, we included two LLMs: LLava and GPT-4.

3.2.1 LLava

LLaVA (Large Language and Vision Assistant) is an advanced multi-modal model published in December 2023 [50]. It combines visual processing features with a large language model (LLM) that results in a unified system capable of interpreting both visual and textual input. The LLaVA architecture combines the expansive open-set visual encoding capabilities of CLIP [51] with Vicuna's [52] language processing abilities. The model has been fine-tuned from end to end using a proprietary dataset to combine visual components with instructional language data. It showcased sophisticated abilities on curated synthetic datasets in multimodal conversations when presented with novel visuals or instructions.

3.2.2 GPT-4

Generative Pre-trained Transformer 4, often known as GPT-4, is the most recent model in OpenAI's GPT family comprised of foundational models [20]. It is a flexible multimodal model capable of handling both textual and visual inputs. GPT-4 produces a wide range of outputs, including both textual and graphic information. While specific architectural elements of GPT-4 are unknown, one may expect improvements in areas such as model size, training data, training processes, and fine-tuning approaches. In comparison to GPT-3.5, GPT-4 has shown significant improvements across a wide range of benchmarks, including national and international exams such as the GRE and LSAT.

4 METHODOLOGY

The methodology adopted to compare the foundational LLMs with fine-tuned deep-learning models is presented in this section. It consists of several steps, each one of which is elaborated next.

4.1 Dataset

In the current study, we have used real-life violence situations dataset [27]. The original dataset comprises 1000 videos each of violence and non-violence taken from YouTube. The violent videos feature a variety of real street fights in diverse environments and conditions. Similarly, non-violent videos are a collection of various human activities, including sports, eating, walking, and more, and are also collected from YouTube. However, we have considered an image dataset of violent and non-violent scenarios for the study. This dataset consists of images taken from the video frames of the previously mentioned video data [27]. It includes 11,063 images divided into two categories: Violence (1) and Non-Violence (0). A sample of the two classes is given in Figure 2.

4.2 Pre-processing steps

Out of the total images (11,063), we took out its 0.5% as the test set (554 images). From the remaining images (10,509 images), 90% were kept for training (9,458) and the rest (1051 images) for validation. For testing, we have to pass each image one by one to Llava and GPT-4, which requires a significant amount of human effort. Therefore, we considered the test size to be 554 images (262 images for the 'Non-violence' class and 292 for the 'Violence' class). The same test set (554 images) was used to test all deep-learning models and LLMs. Moreover, the images were of different resolutions and brought to the same pixel resolution (128x128x3) before being passed to models.

4.3 Response generation from LLMs

The process of inputting images into Large Language Models, obtaining their outputs, and then analyzing the results involves several steps. The following sections provide an elaborate description of each of these sub-stages.

4.3.1 Preparing queries

We devised a direct query for the Large Language Model, instructing it to choose the most appropriate category for the provided image. We first included Google's Gemini LLM in our investigation, but it did not generate any answer, stating *"Sorry, I can't help with that image."*. After testing with many prompts, we opted for the question *"In which category will you put this image? Violent or Non-violent. Make a guess and don't supply further information."* We discovered that LLMs occasionally contained extra information, such as the reasoning behind picking a certain category. We asked LLMs to limit their responses to a single category name to maintain uniformity.

4.3.2 Collecting responses

After finalizing the query format described in the previous section, we fed the images to the LLMs individually. To address potential memory retention concerns with particular LLMs when confronted with similar images, we opened a new chat window after processing 10 images inside one window. For each image, the specified text prompt was presented to both LLava and GPT-4, and the LLM's resultant categorization was carefully documented.

4.3.3 Response cleaning

The LLM outputs require further processing due to possible differences in answer patterns for each question. Despite being instructed to offer replies only in the form of a category, the tools frequently attached additional descriptions or remarks to the response. Thus, post-processing actions were required to segregate the class from the LLMs' answers.

The overall mechanism for response generation can be outlined algorithmically, as depicted in Algorithm 1.

Using Algorithm 1, we successfully determined a category for each test image from both LLMs. These responses were recorded and then compared to the outcomes from fine-tuned deep-learning models for comparison.

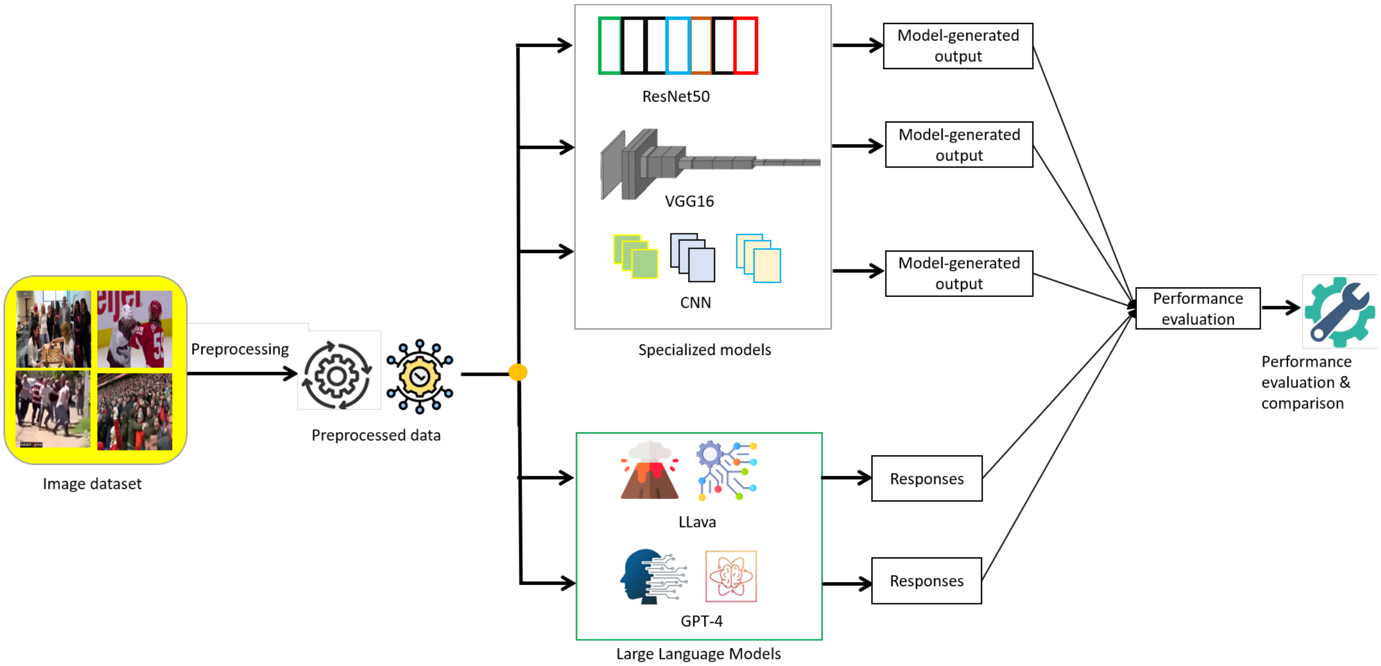
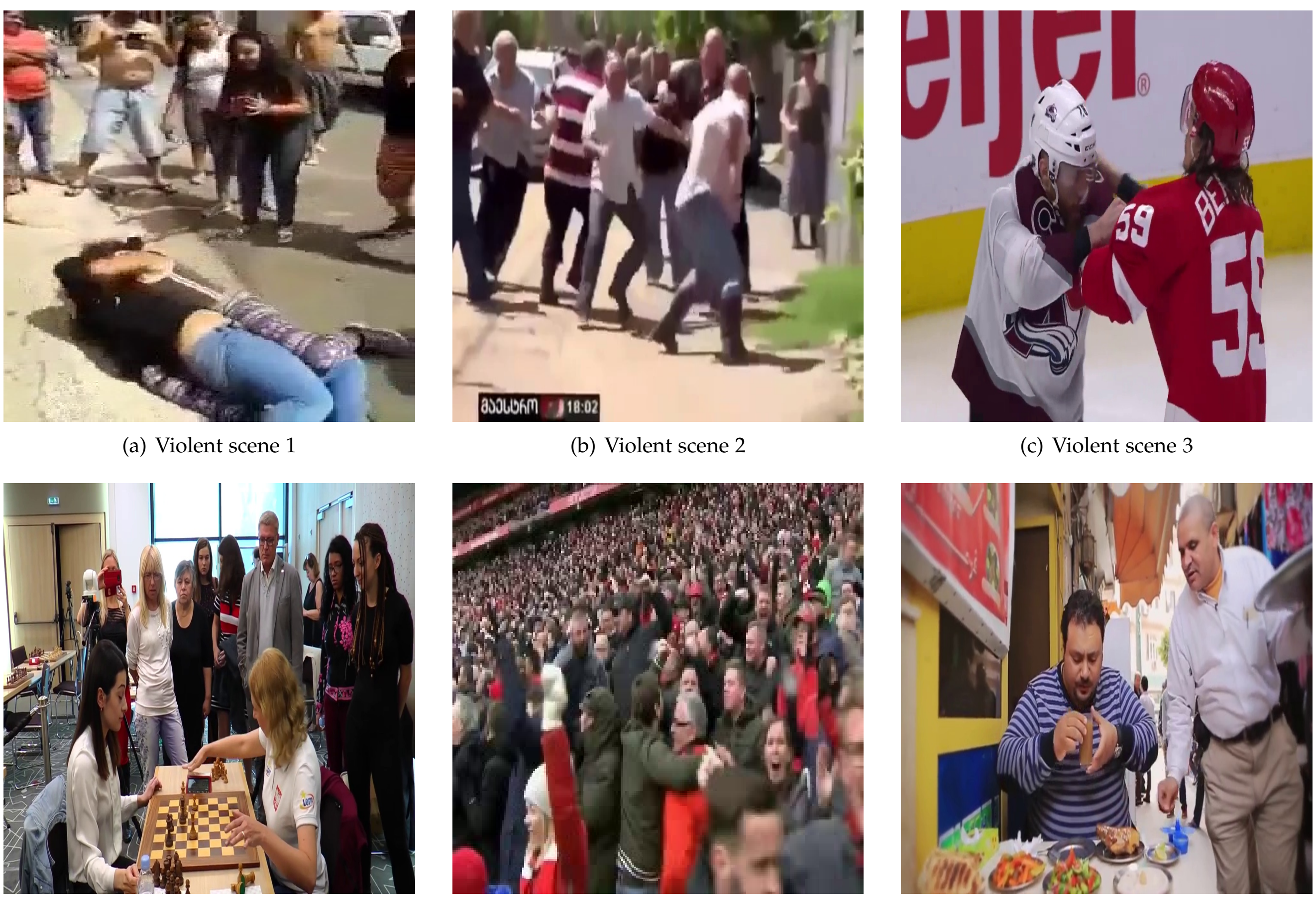


Fig. 1. The pictorial representation of the methodology adopted.



(a) Violent scene 1 (b) Violent scene 2 (c) Violent scene 3

(d) Non-violent scene 1 (e) Non-violent scene 2 (f) Non-violent scene 3

Fig. 2. Samples taken from the dataset for both categories

Algorithm 1 Response generation mechanism from LLMs

for each LLM **do**
  for each image in test set **do**
    Upload the image.
    Supply LLM the text query asking the category of the uploaded image.
    Record the response.
  end for
end for
Clean up the responses, if required.

4.4 Overall Algorithm

In this part, we have encapsulated all the described components into an algorithmic framework (see Algorithm 2). It starts with developing deep learning models, which are then trained and evaluated on a test dataset. The same test dataset is then fed into each LLM, and the results are noted. The performance measures are then generated for both the specialized deep-learning models and the general-purpose LLMs. Finally, the results of these models are put forward for comparison.

Algorithm 2 Comparison of Deep learning models and LLMs

Divide the violence detection dataset (D) into training (D_{TR}), validation (D_V), and test (D_{TS}) sets.
for each deep learning model (M_d) **do**
  Apply pre-processing steps on D, if required.
  Develop the architecture of model M_d.
  Train the model using D_{TR} and D_V.
  Obtain the set of performance measures P_d by applying M_d on D_{TS}.
end for
for each Large Language Model (M_l) **do**
  Apply Algorithm 1 to generate response R_{Ml}.
  Calculate the set of performance measures P_l from R_{Ml}.
end for
Conduct a performance evaluation based on P_d and P_l for each M_d and M_l.

Figure 1 provides a visual depiction of the process. The assessment metrics for each model were calculated by comparing the responses to the actual categories. We then created a confusion matrix for each model to ensure the predictions matched the real label values. We then calculated four conventional evaluation metrics: accuracy, precision, recall, and the F1 score. These metrics serve as widely accepted benchmarks for evaluating classification tasks and are therefore not further elaborated here. Additional metrics, such as the AUC score, were not calculated since the LLMs mentioned discrete categories rather than prediction probabilities.

5 Results and Discussion

This section presents the findings from the comparative analysis between deep learning models and LLMs.

5.1 Deep learning models vs LLMs

The outcomes from both the deep learning models and LLMs are displayed in Table 2, while the confusion matrices for all the models can be observed in the accompanying Figure 3.

Initially, we evaluated the accuracy of each model, which is a crucial indicator of overall classification performance. The findings highlighted significant variations in performance across models. The VGG16 model emerged as the top performer and exhibited the highest accuracy of 0.9422 while Llava produced the lowest accuracy (0.6657). In the category of Large Language Models, GPT-4 performed the best with an accuracy of 0.9242. ResNet50 had better accuracy than CNN and Llava models. It is noteworthy that Llava LLM performed worst in terms of accuracy; it was the best for precision metric (0.9818), surpassing GPT-4 and all deep learning models and reflecting its effectiveness in minimizing false positives. In contrast, CNN model showed the least precision, scoring 0.8996. Notably, both LLMs' lead in precision was not just the highest but also significantly greater compared to the deep learning models.

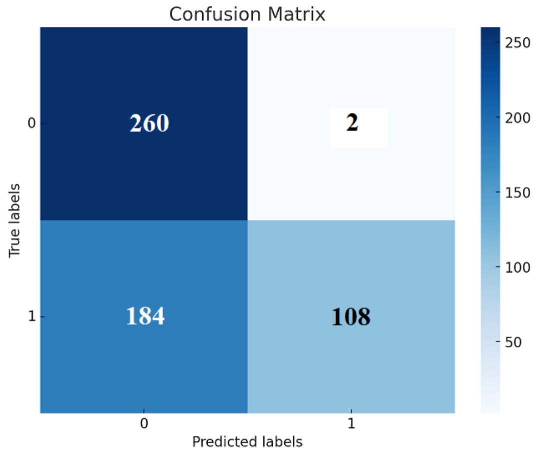
The recall metric, also known as sensitivity, measures a model's effectiveness in correctly identifying all true positive cases. For this metric, ResNet50 performed the best with a recall value of 0.9692, followed by VGG16 (0.9589). On the other hand, Llava scored the lowest value at 0.3699, indicating a tendency to generate more false negatives. The F1-score, a combination of precision and recall, offers a holistic view of a model's effectiveness. Here, VGG16 stood out with the highest F1 score of 0.9459. Among the LLMs, GPT-4 performed the best, with an F1-score of 0.9242.

It's significant to note that while GPT-4 didn't surpass deep learning on all measures, it exhibited a decent performance. Additionally, even LLava demonstrated better precision. In summary, the Large Language Model (LLM) GPT-4 proved more effective than a basic specialized deep-learning model, yet it fell short of achieving the high-performance level of a finely-tuned model like VGG16 and ResNet50. However, with further advancements and a broader scope in training for future LLMs, they may match the performance levels of highly specialized deep learning models.

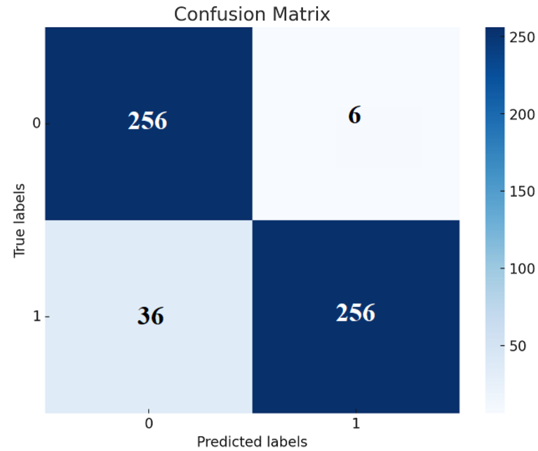
5.2 Reduced dataset

Deep learning models rely heavily on training data volume, affecting their functionality, generalizability, and flexibility [53]. Larger datasets allow the model to recognize and analyze complex patterns and make them capable to perform well on unseen data. The diversity in the dataset is also important as it helps the model to perform well in real-world circumstances where data might vary significantly. Furthermore, the scalability of deep learning models is strongly correlated with data size. To optimize the model's multiple parameters, a large dataset is required [54] as inadequate data may result in insufficient information for the model to fine-tune its extensive parameter space.

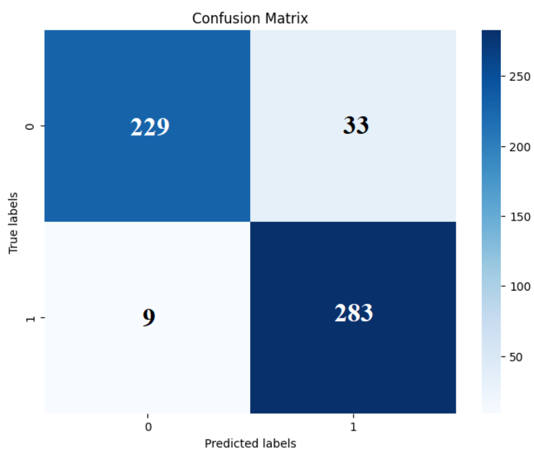
Large Language Models (LLMs) have been popular for their capacity to provide meaningful replies across several applications without needing special fine-tuning or targeted training [55]. They are pre-trained on large, diversified



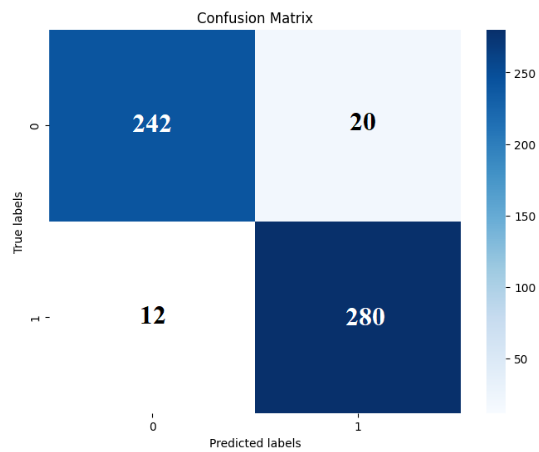
(a) LLava



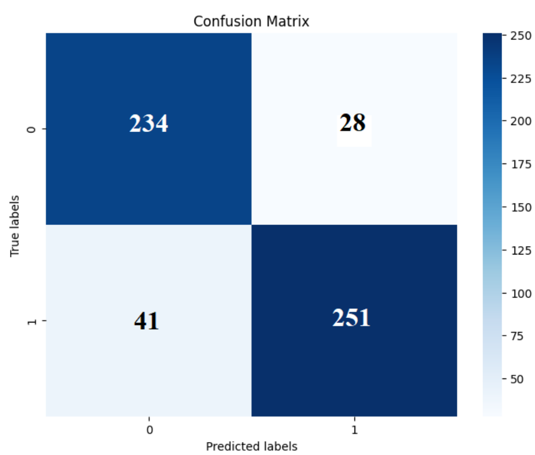
(b) GPT-4



(c) ResNet50



(d) VGG16



(e) CNN

Fig. 3. Confusion matrix of different deep learning models and LLMs

TABLE 2
Results obtained from DLMs and LLMs on various metrics

Category	Model	Performance metrics			
		Accuracy	Precision	Recall	F1 Score
Large Language Model	LLava	0.6657	0.9818	0.3699	0.5373
	GPT-4	0.9242	0.9771	0.8767	0.9242
Deep Learning Model	ResNet50	0.9242	0.8955	0.9692	0.9309
	VGG16	0.9422	0.9333	0.9589	0.9459
	CNN	0.8754	0.8996	0.8595	0.8791

datasets covering a wide range of images, topics and languages. When given a prompt and image, LLMs can create appropriate replies for the given context. The versatility of LLMs make them extremely adaptable and versatile since they may be used across a range of jobs without requiring fine-tuning or specialized adaption [7].

Therefore, it would be interesting to observe whether specialized deep learning models retain their superiority in our scenario even with a smaller amount of training data. To carry out this analysis, we ran two sets of experiments. In the first case, the original dataset was reduced to 50% and 25% in the second case. The half (50%) of the overall dataset comprised a total of 5,254 images, with 4,728 in the training set and 526 in the validation set. For the second case (one-fourth of the total dataset), the training set had 2,364 images and validation set had 263. The original test set of 554 images was used to test both scenarios. We adhered to the same methodology outlined in Algorithm 2. The results obtained are shown in Table 3.

As expected, Table 3 shows that the effectiveness of deep learning models decreases which highlights their reliance on the dataset size. Moreover, the degree of degradation in the performance across all the evaluation criteria is significant (see Figure 4). The accuracy of VGG16 model decreased from 0.9422 to 0.9241 and 0.9097 for the two cases. A similar trend can also be observed for other deep-learning models. The requirement for extensive training highlights the advantage of LLMs over deep learning models, especially for applications with limited training data.

5.3 Explainability

In recent times, researchers have been concentrating on explainable artificial intelligence (XAI) [56], [57]. The interpretability and explainability of models' decisions have become crucial in comprehending their decision-making processes, especially in sensitive applications such as distinguishing between violent and non-violent scenarios. Implementing explainability in the deep learning models is a difficult process and often, methods like attention scores are utilized for the same [58].

On the other hand, LLMs explain their results clearly when asked to predict the class of events depicted in images. For example, when we supplied Figure 2 (a) (a 'Violence' class image) to GPT-4, it provided the following explanation: *The image you've uploaded depicts a woman lying on the street with people standing around her, some of whom are recording the scene on their phones. This scene could be interpreted as potentially distressing or violent depending on the context, which isn't entirely clear just from the image. Therefore, it could be placed in the "violent" category as a cautious classification, especially if the circumstances imply harm or distress.* Moreover, we got the following response when we uploaded Figure 2 (d):*I categorize the image as non-violent based on its content. The image shows two individuals playing a game of chess, which is a strategic and non-physical board game. The surroundings and the expressions of people in the image are calm and focused on the game, indicating a peaceful setting. There are no elements of physical aggression, conflict, or anything that suggests violence. Chess is typically recognized as a game of mental skill and strategy, not physical confrontation. Thus, the context and activity depicted in the photo are inherently non-violent.*

We also analyzed the explanations supplied by two LLMs, GPT-4 and Llava, and find that they use distinct ways of explanation. When asked to explain their responses, Llava responds concisely, such as "Violent" or "Non-Violent," without providing additional details. It provides explanations for its responses when specifically asked. GPT-4, on the other hand, provides more complex responses that justify the category it chooses. Furthermore, we note that Llava only provides thorough explanations when explicitly requested, but GPT-4 does so on a constant basis. GPT-4's openness is critical in developing user confidence and comprehension of its conclusions and reducing response biases. We also recommend conducting more research on this topic to compare LLMs and their future use in situational assessments.

Overall, easy and text-based explanations by LLMs give them a definite advantage over deep learning models, especially for applications where the interpretability of results is critical.

5.4 Limitations

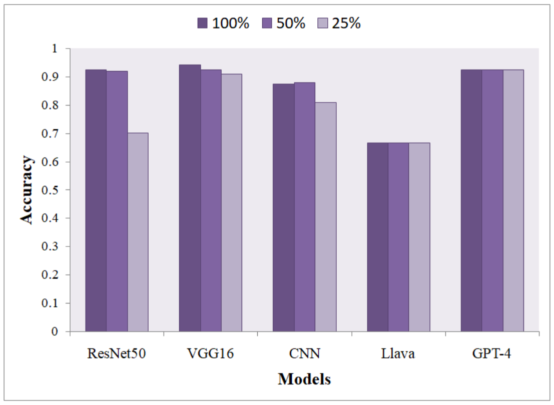
The proposed study has the following shortcomings. First, the test set size was limited to 554 images, which limits the generalizability of the results. The dataset used may not adequately represent the many ways in which violent and non-violent events are portrayed. Using several datasets focused on violence detection might provide a more thorough foundation for comparing specialized deep-learning models against LLMs. Furthermore, as LLMs are upgraded and new variations are available, the conclusions of the study may become obsolete. The assessment is limited to violence detection only. Another limitation is the restricted amount of assessment metrics available due to the nature of the LLM responses. Since LLMs do not generate probabilistic predictions, measures such as the AUC score are not applicable.

6 Conclusion

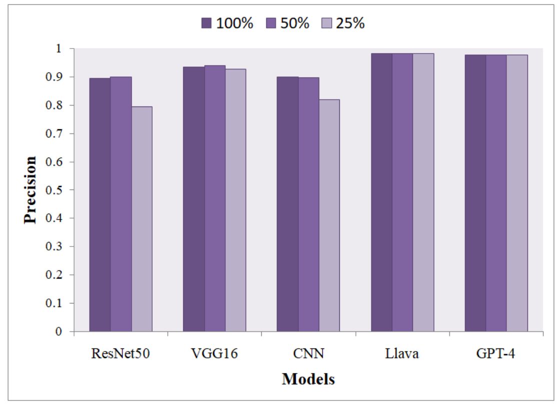
The current study presents an approach for comparing the efficacy of fine-tuned deep-learning models to foundational

TABLE 3
Results obtained from deep learning models with reduced training dataset

Dataset Portion	Model	Performance metrics			
		Accuracy	Precision	Recall	F1 Score
50 % (Training dataset size = 4,728) (Validation dataset size = 526)	ResNet50	0.9205	0.9	0.9554	0.9269
	VGG16	0.9241	0.9401	0.9143	0.9270
	CNN	0.8790	0.8975	0.8698	0.8834
25 % (Training dataset size = 2,364) (Validation dataset size = 263)	ResNet50	0.7021	0.7953	0.5856	0.6745
	VGG16	0.9097	0.9260	0.9006	0.9131
	CNN	0.8104	0.8191	0.8219	0.8205

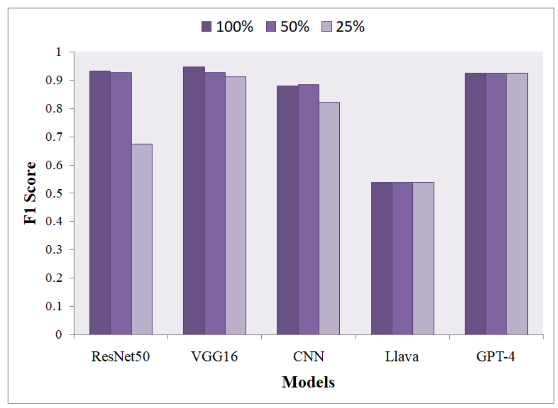


(a) Accuracy



(b) Precision

(c) Recall



(d) F1-score

Fig. 4. Change in performance for deep learning models and LLMs

LLMs in image-based content moderation. Three deep-learning models were built and compared to two LLMs. The results show that LLMs perform well on the dataset used for content moderation across a variety of assessment measures without specialized training. GPT-4 outperformed Llava and other deep learning models, but the VGG16 model demonstrated the most superior overall performance. Notably, when the dataset size decreased, the performance of deep learning models decreased relative to LLMs. It highlights LLMs' advantage and potentially eliminates the need for task-specific datasets. However, when plenty of data is available, fine-tuned models outperform LLMs. It is important to note that future variations of LLMs, which are trained on broader and diverse data, could surpass specialized models in efficacy.

There are various approaches for extending the scope of the current study. Future research might focus on assessing the effectiveness of the suggested technique with more complex multi-modal datasets. Furthermore, it should be applied in a variety of areas, such as natural language processing, computer vision, and audio recognition to understand its significance and adaptability better.

REFERENCES

[1] S. Sai, U. Mittal, V. Chamola, K. Huang, I. Spinelli, S. Scardapane, Z. Tan, and A. Hussain, "Machine un-learning: An overview of techniques, applications, and future directions," *Cognitive Computation*, pp. 1–25, 2023.

[2] D. E. O'Leary, "An analysis of three chatbots: Blenderbot, chatgpt and lamda," pp. 41–54, 2023.

[3] M. Bakker, M. Chadwick, H. Sheahan, M. Tessler, L. Campbell-Gillingham, J. Balaguer, N. McAleese, A. Glaese, J. Aslanides, M. Botvinick *et al.*, "Fine-tuning language models to find agreement among humans with diverse preferences," *Advances in Neural Information Processing Systems*, vol. 35, pp. 38 176–38 189, 2022.

[4] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl *et al.*, "Large language models encode clinical knowledge," *Nature*, vol. 620, no. 7972, pp. 172–180, 2023.

[5] H. Zhang, X. Li, and L. Bing, "Video-llama: An instruction-tuned audio-visual language model for video understanding," *arXiv preprint arXiv:2306.02858*, 2023.

[6] V. Hassija, A. Chakrabarti, A. Singh, V. Chamola, and B. Sikdar, "Unleashing the potential of conversational ai: Amplifying chat-gpt's capabilities and tackling technical hurdles," *IEEE Access*, 2023.

[7] S. S. Sohail, F. Farhat, Y. Himeur, M. Nadeem, D. Ø. Madsen, Y. Singh, S. Atalla, and W. Mansoor, "Decoding chatgpt: a taxonomy of existing research, current challenges, and possible future directions," *Journal of King Saud University-Computer and Information Sciences*, p. 101675, 2023.

[8] M. Dowling and B. Lucey, "Chatgpt for (finance) research: The bananarama conjecture," *Finance Research Letters*, vol. 53, p. 103662, 2023.

[9] E. Loh, "Chatgpt and generative ai chatbots: challenges and opportunities for science, medicine and medical leaders," *BMJ leader*, pp. leader–2023, 2023.

[10] M. Cascella, J. Montomoli, V. Bellini, and E. Bignami, "Evaluating the feasibility of chatgpt in healthcare: an analysis of multiple clinical and research scenarios," *Journal of Medical Systems*, vol. 47, no. 1, p. 33, 2023.

[11] L. Lingard, "Writing with chatgpt: An illustration of its capacity, limitations & implications for academic writers," *Perspectives on Medical Education*, vol. 12, no. 1, p. 261, 2023.

[12] M. Sashida, K. Izumi, and H. Sakaji, "Extraction sdgs-related sentences from sustainability reports using bert and chatgpt," in *2023 14th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)*. IEEE, 2023, pp. 742–745.

[13] F. Mosaiyebzadeh, S. Pouriyeh, R. Parizi, N. Dehbozorgi, M. Dorodchi, and D. Macêdo Batista, "Exploring the role of chatgpt in education: Applications and challenges," in *Proceedings of the 24th Annual Conference on Information Technology Education*, 2023, pp. 84–89.

[14] J. Kocoń, I. Cichecki, O. Kaszyca, M. Kochanek, D. Szydło, J. Baran, J. Bielaniewicz, M. Gruza, A. Janz, K. Kanclerz *et al.*, "Chatgpt: Jack of all trades, master of none," *Information Fusion*, p. 101861, 2023.

[15] I. A. Wong, Q. L. Lian, and D. Sun, "Autonomous travel decision-making: An early glimpse into chatgpt and generative ai," *Journal of Hospitality and Tourism Management*, vol. 56, pp. 253–263, 2023.

[16] G. P. Patrinos, N. Sarhangi, B. Sarrami, N. Khodayari, B. Larijani, and M. Hasanzad, "Using chatgpt to predict the future of personalized medicine," *The Pharmacogenomics Journal*, vol. 23, no. 6, pp. 178–184, 2023.

[17] M. M. Amin, E. Cambria, and B. W. Schuller, "Can chatgpt's responses boost traditional natural language processing?" *IEEE Intelligent Systems*, vol. 38, no. 5, pp. 5–11, 2023.

[18] B. Gupta, T. Mufti, S. S. Sohail, and D. Madsen, "Chatgpt: A brief narrative review," *Cogent Business & Management*, vol. 10, no. 3, p. 2275851, 2023.

[19] V. Chamola, G. Bansal, T. K. Das, V. Hassija, N. S. S. Reddy, J. Wang, S. Zeadally, A. Hussain, F. R. Yu, M. Guizani *et al.*, "Beyond reality: The pivotal role of generative ai in the metaverse," *arXiv preprint arXiv:2308.06272*, 2023.

[20] OpenAI, "Gpt-4 technical report," 2023. [Online]. Available: arXiv:2303.08774

[21] V. U. Gongane, M. V. Munot, and A. D. Anuse, "Detection and moderation of detrimental content on social media platforms: Current status and future directions," *Social Network Analysis and Mining*, vol. 12, no. 1, p. 129, 2022.

[22] S. T. Roberts, *Behind the screen: The hidden digital labor of commercial content moderation*. University of Illinois at Urbana-Champaign, 2014.

[23] X. Wen, K. Dai, Q. Xiong, L. Chen, J. Zhang, and Z. Wang, "An approach to internal threats detection based on sentiment analysis and network analysis," *Journal of Information Security and Applications*, vol. 77, p. 103557, 2023.

[24] D. Shen, G. Wu, and H.-I. Suk, "Deep learning in medical image analysis," *Annual review of biomedical engineering*, vol. 19, pp. 221–248, 2017.

[25] M. Wu and L. Chen, "Image recognition based on deep learning," in *2015 Chinese automation congress (CAC)*. IEEE, 2015, pp. 542–546.

[26] Y. Guo, Y. Liu, A. Oerlemans, S. Lao, S. Wu, and M. S. Lew, "Deep learning for visual understanding: A review," *Neurocomputing*, vol. 187, pp. 27–48, 2016.

[27] M. M. Soliman, M. H. Kamal, M. A. E.-M. Nashed, Y. M. Mostafa, B. S. Chawky, and D. Khattab, "Violence recognition from videos using deep learning techniques," in *2019 Ninth International Conference on Intelligent Computing and Information Systems (ICICIS)*. IEEE, 2019, pp. 80–85.

[28] T. Gillespie, "Content moderation, ai, and the question of scale," *Big Data & Society*, vol. 7, no. 2, p. 2053951720943234, 2020.

[29] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[30] A. Veglis, "Moderation techniques for social media content," in *Social Computing and Social Media: 6th International Conference, SCSM 2014, Held as Part of HCI International 2014, Heraklion, Crete, Greece, June 22-27, 2014. Proceedings 6*. Springer, 2014, pp. 137–148.

[31] J. Cobbe, "Algorithmic censorship by social platforms: Power and resistance," *Philosophy & Technology*, vol. 34, no. 4, pp. 739–766, 2021.

[32] R. Gorwa, R. Binns, and C. Katzenbach, "Algorithmic content moderation: Technical and political challenges in the automation of platform governance," *Big Data & Society*, vol. 7, no. 1, p. 2053951719897945, 2020.

[33] T. Fadziso, "Leveraging on machine moderation to improve content organization," *Asian Journal of Humanity, Art and Literature*, vol. 5, no. 2, pp. 135–144, 2018.

[34] X. Geng, H. Zhang, J. Bian, and T.-S. Chua, "Learning image and user features for recommendation in social networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4274–4282.

[35] K. Yousaf and T. Nawaz, "A deep learning-based approach for inappropriate content detection and classification of youtube videos," *IEEE Access*, vol. 10, pp. 16 283–16 298, 2022.

[36] M. Moustafa, "Applying deep learning to classify pornographic images and videos," *arXiv preprint arXiv:1511.08899*, 2015.

[37] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," in *2016 IEEE European symposium on security and privacy (EuroS&P)*. IEEE, 2016, pp. 372–387.

[38] S. B. Hassen, M. Neji, Z. Hussain, A. Hussain, A. M. Alimi, and M. Frikha, "Deep learning methods for early detection of alzheimer's disease using structural mr images: a survey," *Neurocomputing*, vol. 576, p. 127325, 2024.

[39] M. Anas, A. Saiyeda, S. S. Sohail, E. Cambria, and A. Hussain, "Can generative ai models extract deeper sentiments as compared to traditional deep learning algorithms?" *IEEE Intelligent Systems*, vol. 39, no. 2, pp. 5–10, 2024.

[40] D. Leiker, S. Finnigan, A. Gyllen, and M. Cukurova, "Prototyping the use of large language models (llms) for adult learning content creation at scale," in *CEUR Workshop Proceedings*, vol. 3487. CEUR Workshop Proceedings, 2023, pp. 3–7.

[41] A. Li, J. Wu, and J. P. Bigham, "Using llms to customize the ui of webpages," in *Adjunct Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, 2023, pp. 1–3.

[42] D. Kumar, Y. AbuHashem, and Z. Durumeric, "Watch your language: large language models and content moderation," *arXiv preprint arXiv:2309.14517*, 2023.

[43] M. Franco, O. Gaggi, and C. E. Palazzi, "Analyzing the use of large language models for content moderation with chatgpt examples," in *Proceedings of the 3rd International Workshop on Open Challenges in Online Social Networks*, 2023, pp. 1–8.

[44] H. Ma, C. Zhang, H. Fu, P. Zhao, and B. Wu, "Adapting large language models for content moderation: Pitfalls in data engineering and supervised fine-tuning," *arXiv preprint arXiv:2310.03400*, 2023.

[45] X. Deng, V. Bashlovkina, F. Han, S. Baumgartner, and M. Bendersky, "What do llms know about financial markets? a case study on reddit market sentiment analysis," in *Companion Proceedings of the ACM Web Conference 2023*, 2023, pp. 107–110.

[46] Y. Shao, L. Li, J. Dai, and X. Qiu, "Character-llm: A trainable agent for role-playing," in *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

[47] Y. Du, D. Luo, R. Yan, X. Wang, H. Liu, H. Zhu, Y. Song, and J. Zhang, "Enhancing job recommendation through llm-based generative adversarial networks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 8, 2024, pp. 8363–8371.

[48] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
[49] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
[50] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in *NeurIPS*, 2023.
[51] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.
[52] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez *et al.*, "Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality," *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2023.
[53] C. Sun, A. Shrivastava, S. Singh, and A. Gupta, "Revisiting unreasonable effectiveness of data in deep learning era," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 843–852.
[54] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" *Advances in neural information processing systems*, vol. 27, 2014.
[55] M. M. Amin, E. Cambria, and B. W. Schuller, "Will affective computing emerge from foundation models and general artificial intelligence? a first evaluation of chatgpt," *IEEE Intelligent Systems*, vol. 38, no. 2, pp. 15–23, 2023.
[56] W. Saeed and C. Omlin, "Explainable ai (xai): A systematic meta-survey of current challenges and future opportunities," *Knowledge-Based Systems*, vol. 263, p. 110273, 2023.
[57] Q. M. Areeb, M. Nadeem, S. S. Sohail, R. Imam, F. Doctor, Y. Himeur, A. Hussain, and A. Amira, "Filter bubbles in recommender systems: Fact or fallacy—a systematic review," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 13, no. 6, p. e1512, 2023.
[58] H. Zhao, H. Chen, F. Yang, N. Liu, H. Deng, H. Cai, S. Wang, D. Yin, and M. Du, "Explainability for large language models: A survey," *ACM Transactions on Intelligent Systems and Technology*, 2023.