

DEPARTMENT: AFFECTIVE COMPUTING
AND SENTIMENT ANALYSIS

Are Foundation Models the Next-Generation Social Media Content Moderators?

Mohammad Nadeem, Laeeba Javed, *Aligarh Muslim University, Aligarh, India*

Shahab Saquib Sohail, *VIT Bhopal University, Sehore, India*

Erik Cambria, *Nanyang Technological University, Singapore*

Amir Hussain, *Edinburgh Napier University, UK*

Abstract—Recent progress in AI tools and systems has been significant, especially in their reasoning and efficiency. Notable examples include Generative AI-based large language models (LLMs) like GPT-3.5, GPT-4, Gemini, among others. These foundation models are versatile and effective in handling tasks including code generation, poem writing, itinerary planning etc. based on text input. Moreover, they can manage multi-modal inputs also including text, images, and audio. In our work, we evaluate the effectiveness of fine-tuned deep-learning models compared to general-purpose LLMs in moderating image-based content. We used deep learning models such as CNN, ResNet50, and VGG-16, trained them for violence detection on an image dataset, and tested them on a separate dataset. The same test dataset was also evaluated using LLaVa and GPT-4, two LLMs that can process images. Results demonstrate that VGG-16 model had the highest accuracy at 0.94, while LLaVa had the lowest at 0.66. GPT-4 showed superiority over Llava with an accuracy value of 0.9242. LLaVa recorded the highest precision of all models. Interestingly, GPT-4 showed better results compared to deep learning models when trained on smaller datasets which highlights the usefulness of foundational LLMs for applications having limited training data.

Artificial intelligence (AI) has advanced significantly in recent years 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, Galactica, LLaMA from FAIR, Alpaca from Stanford, BloombergGPT, Chinchilla from DeepMind, and Palm. LLMs can process various types of data including video, images and text which makes them significant language processors [1]. 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. As investigated in [2], 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, medicine, healthcare, sustainability, education, and decision making. In addition, LLMs are being used to evaluate their performance on prediction tasks. For example, Patrinos et al. [3] 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. ChatGPT is widely used for tasks requiring contextual comprehension and excels at catching subtleties in the context of language.

Advancements in LLMs have improved their capacity to handle many forms of input data, such as text, picture, audio, and video. 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 foundation models 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. [4]. 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 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.

Deep learning, known for its hierarchical representation learning, has been instrumental in managing complicated patterns inside enormous datasets that has led us to advances in image analysis, recognition, and comprehension. 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 to facilitate automatic identification of violent behaviors using body language and nonverbal. However, incorporating large language models brings an interesting perspective since they have intrinsic language comprehension skills which might augment the contextual understanding of images [5]. 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 fine-tuned models comprehensively with general-purpose LLMs.

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 foundation models; 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; finally, Section 6 presents the concluding remarks and future directions.

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. Therefore, to maintain the integrity of posts and adhere to legal standards, content moderation has attracted researchers and policymakers greatly. A number of works in this direction have 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 has been all hands on deck for human moderators, who've been the go-to folks for dealing with harmful content. In [6], authors have argued that for effectively handling social media content for adequate moderation, human moderation is better as humans can better deal with the requirements of contextual understanding and its challenges. 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.

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 powerful tools, particularly adept at image and video classification tasks valuable for content moderation. These models are trained on massive datasets labeled as containing harmful or safe content. Once trained, they can then be used to automatically identify and flag potentially harmful content for further review by human moderators. Authors have achieved [7] state-of-the-art performance in detecting inappropriate content on platforms like YouTube using EfficientNet-BiLSTM architecture. Additionally, Moustafa [8] has used convolutional neural networks for tasks like pornography detection. These works set a benchmark to compare 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. Additionally, these models may struggle to generalize to new types of content not included in their training data. 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 [2], translating languages [9], and writing different kinds of creative content. Advances in LLMs have made them more versatile as they can handle not just text but also multi-modal input.

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. They also show promise in rule-based community moderation with noteworthy accuracy and precision. Moreover, LLMs can improve the interaction between users and platforms, aiding in clearer communication [10]. 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.

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. LLMs also perform well in analyzing social media sentiment, though ethical issues must be carefully considered.

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 [11] 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 concerns to be dealt with [12]. 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.

MODELS EMPLOYED

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

Fine-tuned deep learning models

Out of the fine-tuned models, ResNet50 and VGG16 were pre-trained models. We also considered a simple CNN baseline model.

ResNet50 model ResNet50 is a popular pre-trained network that was developed in 2016 to overcome the problematic behavior of deep networks when gradients for the earlier layers become very small (the problem of vanishing gradients) [13]. In our research, we utilized the ResNet-50 model, which was pre-trained using the ImageNet dataset. We applied transfer learning to leverage the ResNet-50 model's capability for feature extraction which it had acquired from training on a

diverse large dataset. The pre-trained models primarily work as feature extractors, and their final fully connected layers are replaced with new layers. First, a global average pooling layer was added to the network to reduce the spatial dimensions and allow for a denser representation of the features. Next, a dense layer with one neuron and 'sigmoid' activation function was included for the final classification task. The model employed the Adam optimizer and a binary cross-entropy loss function for training. We trained the model with 100 epochs and batch size of 32.

VGG16 The VGG16 architecture, developed in 2014, is a well-known deep convolutional neural network design [14]. 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. The 'sigmoid' activation function was used in the last layer with one neuron, preceded by a flattened layer. For training, we used specific hyperparameter settings including the 'Adam' optimizer, a binary cross-entropy loss function, a batch size of 32, and an epoch value of 100.

CNN Convolutional neural networks (CNNs) are widely used for image classification due to their effective structure, which includes convolutional layer followed by pooling layer. At the end of the network, flattened and fully connected layers are used. We used a basic CNN model which was composed of a total of nine layers. First convolutional layer took images of dimension 128x128x3 as input and processed them with the help of 32. It is followed by a second layer with 64 filters and a third layer with 128 filters. All convolutional layers use a 3x3 filter size and a ReLU activation function, and are each followed by a Max Pooling layer with a 2x2 pool size. Subsequently, the extracted features were transformed into a one-dimensional vector using a flattening layer. Lastly, two dense layers were added into the network. The first one had 128 neurons with ReLU activation, and the final layer contained one neuron with a sigmoid activation function. The architectural setup of CNN model is outlined in Table 1.

Foundation Models

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

TABLE 1. Architectural setup of CNN model

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		

LLava LLava (Large Language and Vision Assistant) is an advanced multi-modal model published in December 2023 [15]. It combines visual processing features with a large language model (LLM) that results in model capable of handling textual and visual input. It combines the visual encoding ability of CLIP with language processing ability of Vicuna's. The model has been fine-tuned with the help of a proprietary dataset. LLava has showcased decent performance on complex datasets in multi-modal conversations when presented with novel visuals or instructions.

GPT-4 Generative Pre-trained Transformer 4, often known as GPT-4, is the most recent model in OpenAI's GPT family comprised of foundation models [16]. It is a flexible multi-modal model capable of handling both textual and visual inputs. GPT-4 produces a wide range of outputs, including both images and text. Though internal architecture of GPT-4 is unknown, one may expect improvements in areas such as model size, training data, training processes, and fine-tuning approaches. In comparison to previous versions, notable improvements are observed while using GPT-4 across a wide range of benchmarks and applications.

PROPOSED METHODOLOGY

This section presents the overall approach used to analyze general-purpose LLMs and specific deep-learning models. It includes several steps, as detailed below.

Dataset

In the current study, we have used real-life violence situations dataset. 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

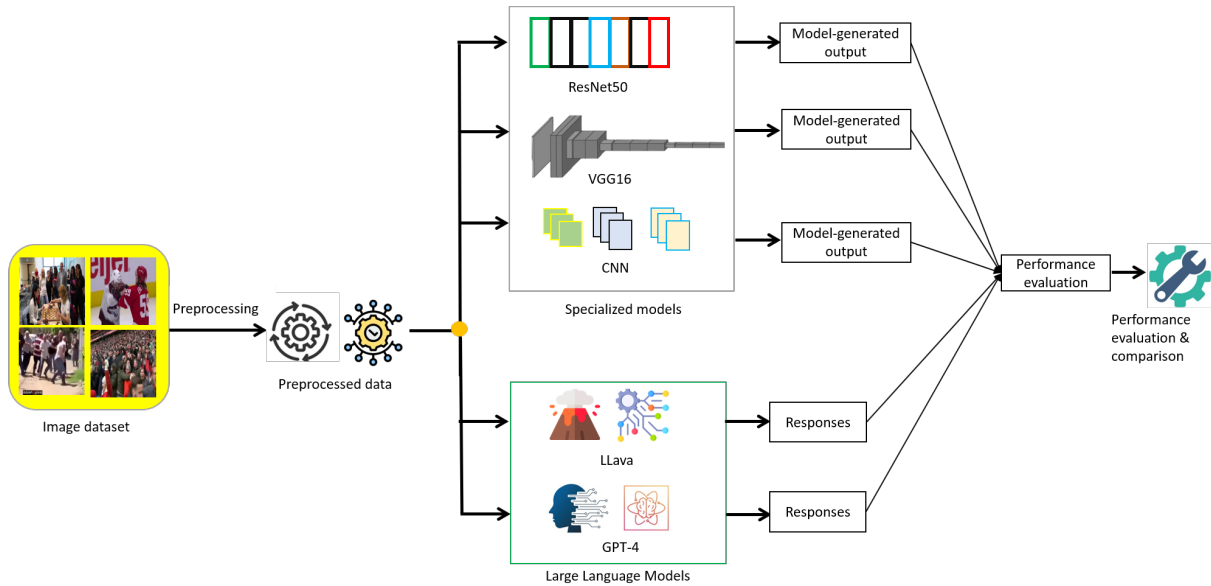


FIGURE 1. Overall flow diagram of the adopted methodology

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 [17]. 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.

Initial processing of images

From the overall set of 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'). We used the same test dataset of 554 images to test all LLMs and deep-learning models. Moreover, the images had different resolutions and were brought to the same pixel resolution (128x128x3) before being passed to models.

Response generation from LLMs

The process of inputting images into Large Language Models, obtaining their outputs, and then analyzing the results involves several phases as discussed next.

Query preparation A direct query was devised for LLMs to instruct them to choose the most appropriate category for the provided image. Initially, we included Gemini LLM (developed by Google) in our experiments. However, it did not generate an appropriate answer as it could not process the images of people yet. 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 found that large language models (LLMs) sometimes provided additional information, like the reasoning for choosing a specific category. To keep responses consistent, we tried to restrict the LLMs with their answers to the category name only.

Response collection Once the query was finalized as described earlier, we supplied LLMs with the images one by one. To address potential memory retention concerns with particular LLMs when confronted with similar images, a new window was opened after processing 10 images inside one window. The specified text prompt was presented to both LLMs for each image, and the LLM's resultant categorization was carefully documented.



FIGURE 2. Samples taken from the dataset for both categories

Response cleaning The LLM outputs require post-processing due to possible differences in answer patterns for each query. 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 answers of LLMs. The complete mechanism to generate responses from LLMs is outlined in Algorithm 1.

Algorithm 1 Response collection process

- 1: **for** each image m in test set **do**
 - 2: **for** each LLM L **do**
 - 3: Supply m to L .
 - 4: Ask L through prompt to identify class of m .
 - 5: Store the output of L .
 - 6: **end for**
 - 7: **end for**
 - 8: Post-process the output, if necessary.
-

Using Algorithm 1, we successfully determined categories of all images in the test set for Llava and GPT-

4. The outputs were recorded and then analyzed with the outcomes from fine-tuned deep-learning models for comparison.

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.

Figure 1 provides a comprehensive understanding of the adopted methodology. The performance of each model was measured by comparing its responses to the true categories. For each model, a confusion matrix was created to check if the predictions corresponded with the actual labels. Furthermore, we calculated four key evaluation metrics: recall, precision, F1 score and

Algorithm 2 LLMs and deep learning models comparison

- 1: Create train (T_r), validation (V), and test (T_S) sets by splitting overall violence detection dataset D .
- 2: **for** each general-purpose LLM (G_l) **do**
- 3: Use Algorithm 1 to record Output O_l .
- 4: Compute evaluation metrics E_l from O_l .
- 5: **end for**
- 6: **for** each fine-tuned deep model (F_d) **do**
- 7: Pre-process T_r , V and T_S , if necessary.
- 8: Design the architecture F_d .
- 9: Supply T_r and V to train F_d .
- 10: Run trained F_d on T_S to calculate evaluation metrics E_d .
- 11: **end for**
- 12: Compare E_d and E_l .

accuracy as they are standard measures for evaluating classification problems. We could not compute the AUC score because the LLMs provided specific category names rather than probability scores.

FINDINGS

This section outlines the results obtained from fine-tuned deep learning models and foundational LLMs.

Deep learning models vs LLMs

Table 2 shows the results obtained from both the deep learning models and LLMs and Figure 3 contains the corresponding confusion matrices.

The results showed notable differences among the models' performance. VGG16 model surpassed the rest of the models in terms of accuracy (0.9422), while Llava had the lowest accuracy at 0.6657. Among the large language models, GPT-4 had the best accuracy value of 0.9242. ResNet50 outperformed both the CNN and Llava models in accuracy. Although Llava LLM ranked lowest in accuracy, it achieved the highest precision rate of 0.9818 which was better than GPT-4 and all the deep learning models. It demonstrates LLava's capability to minimize false positives. On the other hand, the CNN model had the lowest precision with a value of 0.8996. Interestingly, both LLMs' precision scores were significantly higher than those of fine-tuned models.

The recall (sensitivity) metric evaluates model's efficiency in identifying the true positive instances. ResNet50 led in recall with a value of 0.9692, followed closely by VGG16 at 0.9589. Conversely, Llava had the lowest recall at 0.3699 which shows a higher likelihood of missing true positives. For the F1-score, VGG16

performed best with a value of 0.9459. Between LLMs, GPT-4 surpassed Llava with a F1-score of 0.9242.

It is 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 surpassed the baseline deep-learning model but did not reach the accuracy levels of more specialized VGG16 and ResNet50 models. However, with future improvements and more comprehensive training, future LLMs could potentially perform comparable to specialized deep learning models.

Reduced dataset

Deep learning models rely heavily on training data volume, affecting their functionality, generalizability, and flexibility. Larger and more diverse datasets allow the model to recognize and analyze complex patterns and make them capable of performing well on unseen data. Furthermore, the scalability of deep learning models is strongly correlated with data size. To optimize the model's huge parameter space, a large dataset is required as inadequate data may result in insufficient information for the model to calibrate them effectively. On the other hand, LLMs have become popular as they provide meaningful replies across several applications without the need for special fine-tuning or targeted training [18]. They are pre-trained on large, diversified datasets covering a wide range of images, topics and languages. The versatility of LLMs also makes them extremely adaptable.

Therefore, it is plausible to analyze if specialized deep learning models continue to outperform in our scenario, even with less training data. For this analysis, we conducted two sets of experiments. In the first scenario, we reduced the original dataset to 50%, and in the second, to 25%. The dataset at 50% included 5,254 images, with 4,728 for training and 526 for validation. When the dataset was reduced to 25%, the train set contained 2,364 instances and validation set 263. We used the original test set of 554 images for both cases. The same methodology as described in Algorithm 2 was followed. Table 3 contains the corresponding results obtained from reduced datasets.

As anticipated, the performance of deep learning models declines which underscores their dependence on the size of the dataset (see Table 3). Furthermore, the drop in performance is significant across all evaluation metrics, as shown in Figure 4. For example, the accuracy of the VGG16 model fell from 0.9422 to 0.9241 and then to 0.9097 in the two scenarios. A similar pattern is seen with other deep-learning models. The

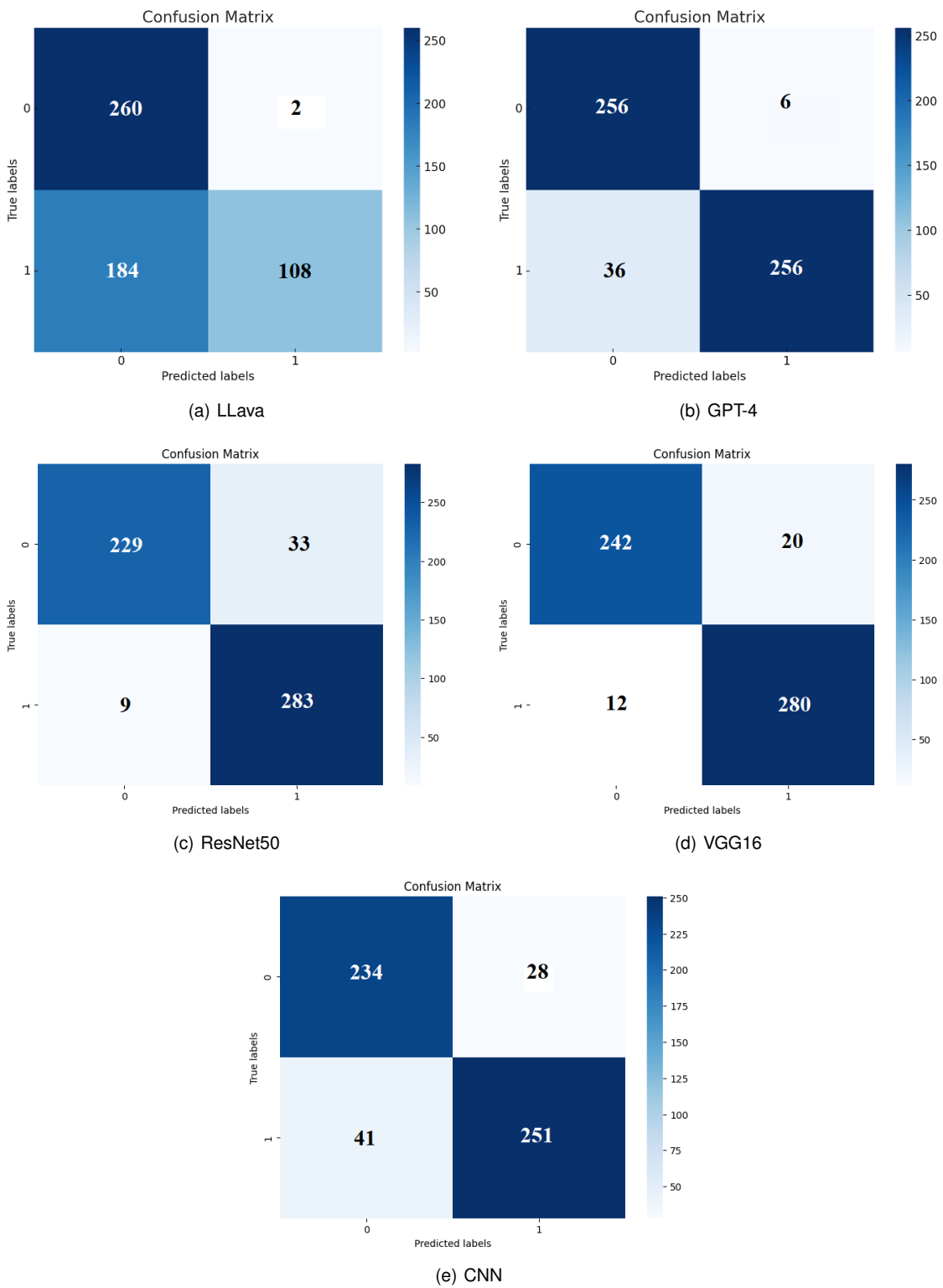


FIGURE 3. Confusion matrix of all LLMs and deep learning models

TABLE 2. Deep learning models vs LLMs on various evaluation measures

Class	Model	Evaluation measures			
		Accuracy	Precision	Recall	F1 Score
General purpose LLM	LLava	0.6657	0.9818	0.3699	0.5373
	GPT-4	0.9242	0.9771	0.8767	0.9242
Fine-tuned DLM	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

TABLE 3. Results obtained from deep learning models with smaller training set

Dataset Portion	Model	Evaluation measures			
		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

need for extensive training highlights the advantage of LLMs over deep learning models, particularly when available training data is limited.

Explainability

In recent times, researchers have been concentrating on explainable artificial intelligence (XAI) [19]. The interpretability and explainability of models' decisions have become crucial in comprehending the reasoning process, especially in sensitive areas like 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 [20].

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 Llava and GPT-4 and observed that they use distinct ways of reasoning. 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.

Limitations

The proposed study has the following shortcomings. First, we used a small test set of 554 images, which limits the generalizability of the results. It does not adequately cover the scenarios indicative of violent and non-violent events. 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

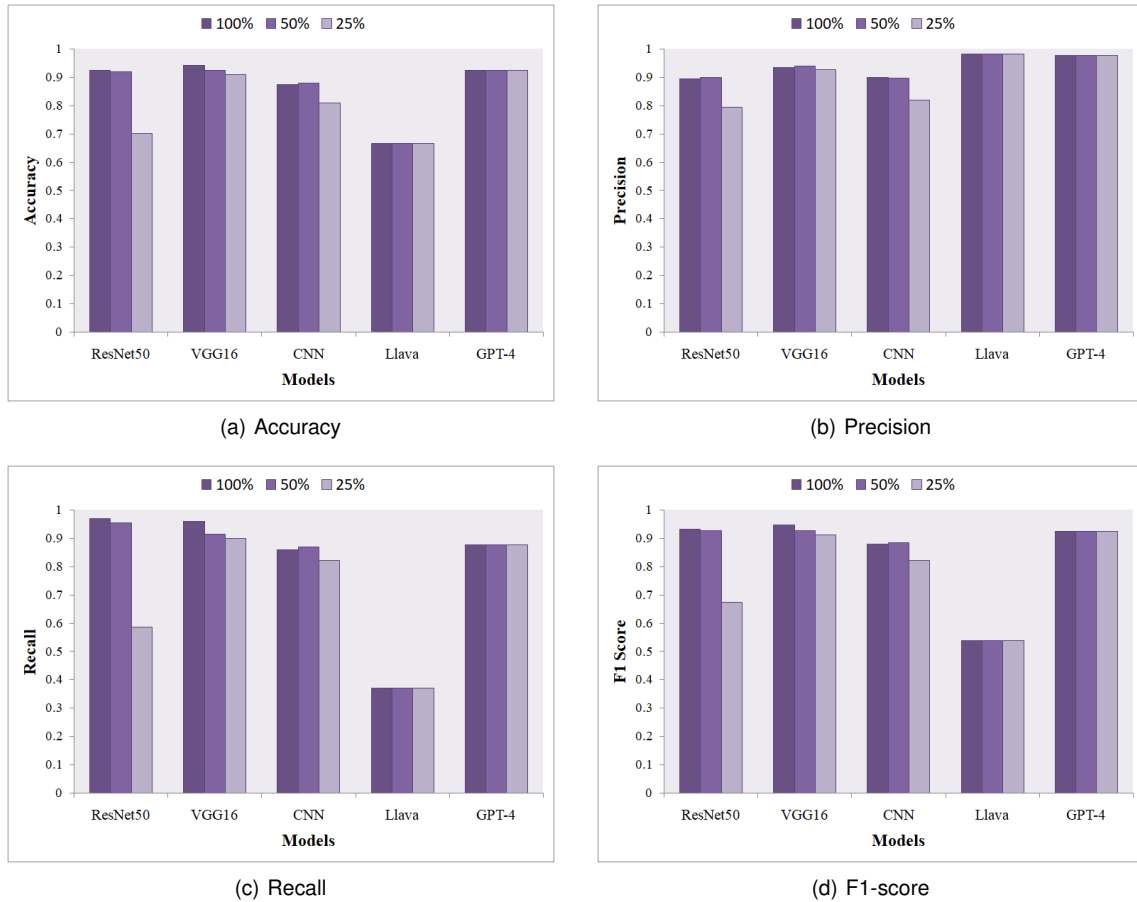


FIGURE 4. Change in performance for deep learning models and LLMs

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.

CONCLUSION

The current study presents an approach for comparing the efficacy of fine-tuned deep-learning models to foundation models 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. 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 utility of suggested technique with more sophisticated multi-modal tasks. 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. P. Dhruv and S. Naskar, "Image classification using convolutional neural network (CNN) and recurrent neural network (RNN): A review," *Machine Learning*

- and Information Processing: Proceedings of ICMLIP 2019, 2020, pp. 367–381.
2. S. S. Sohail et al., “Decoding ChatGPT: a taxonomy of existing research, current challenges, and possible future directions,” *Journal of King Saud University-Computer and Information Sciences*, 2023, p. 101675.
 3. G. P. Patrinos et al., “Using ChatGPT to predict the future of personalized medicine,” *The Pharmacogenomics Journal*, vol. 23, no. 6, 2023, pp. 178–184.
 4. 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, 2022, p. 129.
 5. T. Brown et al., “Language models are few-shot learners,” *Advances in neural information processing systems*, vol. 33, 2020, pp. 1877–1901.
 6. T. Gillespie, “Content moderation, AI, and the question of scale,” *Big Data & Society*, vol. 7, no. 2, 2020, p. 2053951720943234.
 7. K. Yousaf and T. Nawaz, “A deep learning-based approach for inappropriate content detection and classification of youtube videos,” *IEEE Access*, vol. 10, 2022, pp. 16283–16298.
 8. M. Moustafa, “Applying deep learning to classify pornographic images and videos,” *arXiv preprint arXiv:1511.08899*, 2015.
 9. M. Anas et al., “Can Generative AI Models Extract Deeper Sentiments as Compared to Traditional Deep Learning Algorithms?” *IEEE Intelligent Systems*, vol. 39, no. 2, 2024, pp. 5–10.
 10. M. Franco, O. Gaggi, and C. E. Palazzi, “Analyzing the use of large language models for content moderation with chatgpt examples,” *Proceedings of the 3rd International Workshop on Open Challenges in Online Social Networks*, 2023, pp. 1–8.
 11. D. Kumar, Y. AbuHashem, and Z. Durumeric, “Watch your language: large language models and content moderation,” *arXiv preprint arXiv:2309.14517*, 2023.
 12. Y. Du et al., “Enhancing job recommendation through llm-based generative adversarial networks,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, 2024, pp. 8363–8371.
 13. K. He et al., “Deep residual learning for image recognition,” *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
 14. K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.
 15. M. Nadeem et al., “Vision-Enabled Large Language and Deep Learning Models for Image-Based Emotion Recognition,” *Cognitive Computation*, 2024, pp. 1–14.
 16. OpenAI, “GPT-4 technical report,” 2023, [arXiv:2303.08774](https://arxiv.org/abs/2303.08774).
 17. M. M. Soliman et al., “Violence recognition from videos using deep learning techniques,” *2019 Ninth International Conference on Intelligent Computing and Information Systems (ICICIS)*, 2019, pp. 80–85.
 18. 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, 2023, pp. 15–23.
 19. Q. M. Areeb et al., “Filter bubbles in recommender systems: Fact or fallacy—A systematic review,” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 13, no. 6, 2023, p. e1512.
 20. E. Cambria et al., “SenticNet 8: Fusing Emotion AI and Commonsense AI for Interpretable, Trustworthy, and Explainable Affective Computing,” *International Conference on Human-Computer Interaction (HCI)*, 2024.