Assessing Open Source Large Multimodal Models Capabilities at Detecting Deepfake Images

Santiago Posse
sa970924@ucf.edu
University of Central Florida
Orlando, Florida, USA

ABSTRACT

As the prevalence of deepfake technology increases, the need for readily available, easy to use detection tools is critical. Recently the surge of available large multimodal models (LMMs) has fundamentally changed the way users interact with models. Since they are small, open source LMMs are able to run locally on a wide variety of hardware while keeping safe and private connections in a closed system. This research assesses local LMMs capabilities at detecting real and deepfake images using a variety of prompts. The dataset used consists of 20,000 real and deepfake images split equally, with the deepfake images generated by Stable Diffusion Inpainting. For prompts that require human evaluation, 200 of the images are used and scored on a scale of 1-5. Results reveal open source LMMs are good at detecting real images but struggle to identify deepfake images. However, as prompts became more detailed, many models were able to increase their deepfake detection accuracy. The code and responses can be found at https://github.com/santiagoposse/Assessing-LMMs-Capabilities-at-Detecting-Deepfakes.

1 INTRODUCTION & PROBLEM STATEMENT

The introduction of generative models and their capabilities has caused an influx of deepfake and altered images. This poses many risks, including defaming individuals and spreading misinformation. While generative AI can increase productivity in industries, it is significantly harder to detect deepfakes compared to creating them. Currently, deepfake detection is handled with complex models [4] in combination with multiple attention frameworks [17]. Deepfake detection systems require an extensive technical background for proper utilization.

However, large multimodal models (LMMs) have become increasingly available in both cloud base models like ChatGPT [10] and open source models [8] [5] [16] [14] [1]. LMMs have a wide range of capabilities due to the massive amount of data it is trained on. Unfortunately, this does not entirely transfer to open source models. Open source LMMs, that most system can load, are limited in size due to memory constraints. This leads to most small open source models building upon larger pre-trained models [8]. Most importantly, by combining high quality data in fine-tuning and vision encoders, small open source models can reach GPT-4V level performance at a fraction of the compute necessary [16].

The purpose of this paper is to demonstrate open source LMMs capabilities at assessing deepfake images. All models tested will get a variety of prompts alongside an image. The prompts vary from binary yes and no, to explanatory responses looking for correctness, clarity, and relevance. The hope is that open source LMMs can demonstrate their capabilities at providing accurate responses with the benefits of private, local, and small models. The remainder of the paper is summarized as follows. Section 2 summarizes related work pertaining to deepfake detection and detecting with large language models. Section 3 introduces the methods used in this study. Section 4 highlights the results of the models across all the prompts. Section 5 will delve and discuss the challenges and possible improvements to the study. Lastly, section 6 will summarize the key takeaways of the study and future research that could be done.

2 RELATED WORK

Introduction of Vision Capabilities. Early models like Flamingo [3] introduced the idea of connecting a pre-trained vision encoder to a LLM and demonstrated zero-shot and few-shot capabilities across different visual language task. An image alongside a prompt could be inputted to the model and a response that demonstrates image understanding would be returned. After ChatGPT proposed their own vision capabilities, many open source model including LLaVA [8] are proposed. By introducing visual instruction tuning LLaVA reaches higher benchmark scores across a broader range of task, to achieve this LLaVA uses a vision-language connector which is a MLP layer between the vision encoder and the large language model.

Previous Deepfake Detection. Deepfake detection has risen in significance due to the rapid and constant evolution of generative AI [9] [11]. Several methods have been introduced to leverage different architectures. Many early deep learning developments involved using convolutional neural networks (CNNs) to extract features and use a dense network to classify deepfake images [2]. As deepfake generation became more complex, video detection frameworks were also proposed. Similarly to before, features were extracted using CNNs, and for the sequential process it was handled by an LSTM [6] which subsequently fed the output into a decoder. Existing models during this time treated deepfake detection as a binary problem [17], focusing on feature extraction, which as time passed, became increasingly difficult due to generative tools becoming more advanced.

Using Large Language Models (LLMs) for Detection. A majority of LLMs use the transformer architecture [15] which allows for long-distance dependencies and memory from a large corpus. More recently, LLMs have upgraded too multimodal model capabilities. This allows the model to receive and output multiple types of inputs like images, voice, and video[10]. In light of these advancements, researchers in the forensics field assessed how capable the model was at digital forensics. Their findings demonstrated that the model is competent at completing forensics tasks, but the security aspect

was a issue [12]. Recently, research was conducted detecting deepfake faces with both GPT-4 and Gemini. The models tend to do much better when classifying real images compared to deepfake images[7]. The type of model that is used to generate also plays a significant role showing up to a 7% increase in accuracy[7].

3 METHODOLOGY

3.1 Model Selection

To keep the size of the models similar for consistency, all models are between 7B and 12B parameters. While this limits the possible choices, it removes bias of having a much larger model. Additionally, the size range are popular and accessible to user who have a GPU with 12GB of VRAM or more. Six models were picked for the study, each model varies in their architecture. The models include LLaVA [8] at 7B parameters using the CLIP vision encoder and Vicuna language model. LLaVA-Llama [5] uses the CLIP vision encoder but is fine-tuned to the LLaMA instruct and is 8B parameters. MiniCPM-V [16] at 8B uses the SigLip vision encoder along with Qwen2-7B LLM. BakLLaVA is based on the Mistral LLM augmented with the LLaVA architecture at 7B parameters. LLaMA 3.2-vision [1] is an expansion of LLaMA 3 but includes the CLIP vision encoder, reaching 11B parameters. Finally, Gemma 3 [14] uses the SigLip vision encoder with the gemma language model.

3.2 Dataset Selection and Reduction

The dataset used is part of the DeepFakeFace [13] set. The dataset includes a total of 120,000 images. It is split between 30,000 real images of celebrities inside a file called Wiki, and 90,000 augmented images. The augmented images were generated by different methods, including Stable Diffusion v1.5 [11], Stable Diffusion Inpainting, and Insight. For this paper, 10,000 of the real images were used and the same 10,000 images were taken from Stable Diffusion Inpainting. The reason for selecting Inpainting is due to how the images are augmented. Unlike the other two generative models, Inpainting only augments the face and not the setting, demonstrated in figure 1. This consistency was key for when evaluating the responses. For the human evaluation scoring metric, only 100 images from Inpainting and 100 Wiki images are selected due to having to manually analyze the answers.

3.3 Creating the Multiple Prompts

The prompts ranged from binary yes or no answers to requiring complex answers with reasoning to score higher in the human evaluation metric. The range of prompts allowed for a better analysis on how the models respond to the language in the prompt.

- **Prompt 1**: Is this picture a deepfake, answer with yes or no.
- **Prompt 2**: Has the person in the picture been altered or augmented in any way, answer with yes or no.
- **Prompt 3**: Analyze this image and determine if it is a deepfake. Provide an answer, along with an explanation of your reasoning.
- **Prompt 4**: Analyze the image of a person for signs of deepfake augmentation or manipulation. Look for specific indicators, including: facial inconsistencies, skin texture anomalies, eye and teeth issues, edge artifacts, lighting mismatches, background distortions. Based on your analysis, do you find evidence of deepfake manipulation? If so, list the specific indicators you detected and explain why they suggest deepfake manipulation. If no evidence is found, state that there are no clear signs of deepfake augmentation.

Figure 1: The four images demonstrate the possible augmentation that could occur based on the different generation methods. For the purpose of this study, only Wiki and Inpainting images were used.

Prompt 1 and 2 use differing language to see if the models have a bias towards the word deepfake or the opposite. Prompt 3 and 4 requires the model to both answer if the image is a deepfake or not and provide an explanation as to what parts of the image are augmented. Prompt 4 is hyper specific with exact details on what to possible anomalies could be present and to look for them.

3.4 Evaluation Metrics

Due to wanting an emphasis on seeing how the models react to the different images, for prompt 1 and 2 there are three accuracy results. The Wiki accuracy measures how accurate the models are at predicting real images, and Inpainting accuracy measures deepfake images. Finally, to account for bias between the binary prompts, F1-score will be taken.

Prompt 3 and 4 use a scoring method given to evaluators that is as follows:

- **Score 1**: Hallucination, irrelevant answer.
- **Score 2**: Provided an incorrect answer.
- **Score 3**: Refused to provide an answer or requested more context before answering.
- **Score 4**: Provides the correct answer but is not always coherent or relevant. Additionally, casts some doubt on if they are certain with their response.
- **Score 5**: Provides the correct answer that is coherent and provides correct reasons for selecting yes or no. There is additionally no doubt in the response.

Using the above scoring method, we have a qualitative way of measuring correctness, coherence, and relevance.

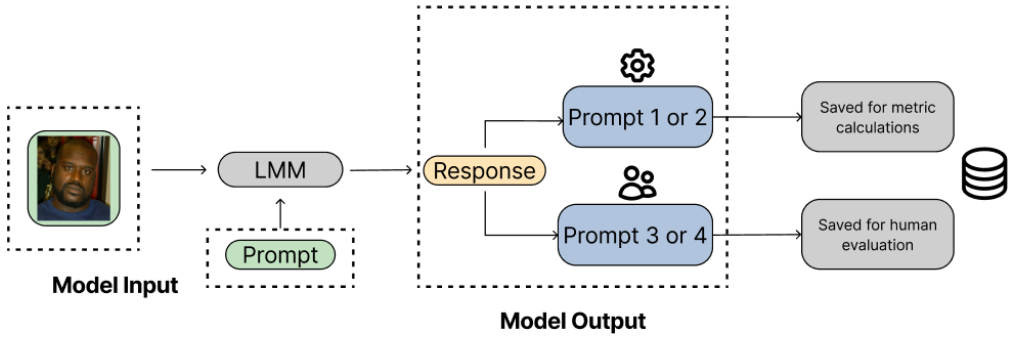


Figure 2: The pipeline for the approach. The pipeline represents how one image is handled, this approach is done for every image and repeated for all models. Depending on the current prompt changes where the data is stored. Prompt 1 and 2 are directly stored in files to be analyzed through evaluation calculations. Prompt 3 and 4 responses are saved to be scored based on the human evaluation metric.

3.5 Approach and Pipeline

The images are inputted into each model in two separate groups, the real Wiki images and the augmented Inpainting images. Along with the image, one of the four prompts are also given, the input is similar to a dialog system where the model is expected to respond given the prompt and context in the image. The model then has one chance at providing an answer for each image. If the model fails to provide an answer for prompt 1 or 2 it is not counted and does not affect the accuracy rating, if the model fails to provide an answer for prompts 3 or 4 then it will receive a rating score of 3. All responses are recorded regardless of answer to their respective file. In total there were 48 response files, 12 files for each prompt, and two files for each model. Separating each model by Wiki and Inpainting responses streamlined the organization and made for a cleaner pipeline demonstrated with figure 2.

4 EVALUATION & RESULTS

4.1 Prompt Results

4.1.1 ***Prompt 1***. This prompt was created to analyze if the models have an understanding of the term "deepfake". The results in table 1 show there is a large discrepancy towards picking the real (Wiki) images. Overall, the models perform very poorly when evaluating the deepfake images. This suggests that when the models analyze the images, they are unable to detect the anomalies present in the deepfake (Inpainting) images. This is likely due to prompt 1 only asking for a binary yes or no response without additional context on what anomalies could be. All models additionally had poor F1 scores ranging from 0.028 to 0.509. Based on these results models like LLaVa-Llama, LLaVA, and BakLLaVA have higher true positive responses compared to the other three models. LLaVA-LLama has the highest F1 score. Although achieving a higher F1 score the model has a below average real image accuracy at 68.95%, suggesting the model is guessing with both real and deepfake images. Between the three higher F1 score models, BakLLaVA had the highest average accuracy with 91.43% real image accuracy and 23.04% deepfake image accuracy.

On the other hand, we have three models that achieved similarly low F1 scores, these models include MiniCPM-V, LLama 3, and Gemma 3. There is a trend between the three models where they have a very high accuracy on real images but poor deepfake image accuracy. This suggests these models only respond with yes, all the images are real. All the models have 99% accuracy or higher on real images but 6% or less on deepfake images. This large imbalance in responses is what leads the models to have a low F1 score. Regardless of which group the models are part of, they all demonstrate poor performance.

Table 1: Results from prompt 1 reported in accuracy (%). 'Wiki' are the real image dataset and 'Inpainting' is the dataset created from Stable Diffusion Inpainting. The best results are highlighted in **bold.** F1-score is reported to offset the bias from the models selecting yes.

Model	**Wiki**	**Inpainting**	**F1 Score**
LLaVA [8]	79.32	23.63	0.329
LLaVA-Llama [5]	68.95	**44.68**	**0.509**
MiniCPM-V [16]	99.35	03.52	0.068
BakLLaVA	91.43	23.04	0.350
LLaMA 3 [1]	99.90	01.43	0.028
Gemma 3 [14]	**99.92**	06.22	0.117

4.1.2 ***Prompt 2***. Prompt 2 was designed to shift the language away from "deepfake" and instead ask whether a person in the image appears "altered or augmented." This prompt aims to assess whether models perform better when using more generalized or indirect phrasing. The results in table 2 demonstrate that this slight change in wording improves overall balance for some models. Compared to Prompt 1, several models saw a significant increase in deepfake image accuracy. Most notably, BakLLaVA showed the best improvement, reaching 56.41% on deepfake images and an overall F1 score of 0.578, the highest among all models. However, this came at the cost of lower real image accuracy, suggesting the model is more aggressive or sensitive to possible manipulations like a blurred background. Other models also benefited from the wording change. LLaVA-Llama improved its deepfake accuracy from 44.68% to 46.22% and increased its F1 score to 0.566, maintaining a relatively balanced performance. Similarly, LLaVA improved from 23.63% to 34.87% on deepfake images, resulting in a better F1 score.

Table 2: Results from prompt 2 reported in accuracy (%) and combined F1-score. Prompt 2 responses sees more models increasing accuracy for deepfake images at the cost of accuracy on real images.

Model	**Wiki**	**Inpainting**	**F1 Score**
LLaVA [8]	74.30	34.87	0.515
LLaVA-Llama [5]	66.87	46.22	0.566
MiniCPM-V [16]	97.69	15.14	0.259
BakLLaVA	61.28	**56.41**	**0.578**
LLaMA 3 [1]	**98.63**	05.72	0.107
Gemma 3 [14]	91.11	21.74	0.333

However, some models like MiniCPM-V, LLaMA 3, and Gemma 3 remained heavily biased toward classifying images as real, with real image accuracy over 90% and deepfake image accuracy under 22%. This shows that despite the improved prompt wording, these models still struggle to recognize deepfake specific cues or tend to default to no as a safe response. Overall, Prompt 2's phrasing seems to encourage more models to engage with the idea of possible image alteration, and in some cases, leads to better balance between real and fake image classification. The trade-off often comes at the cost of reduced confidence in identifying real images, and overall, many models still lack robust deepfake detection accuracy without detailed prompting.

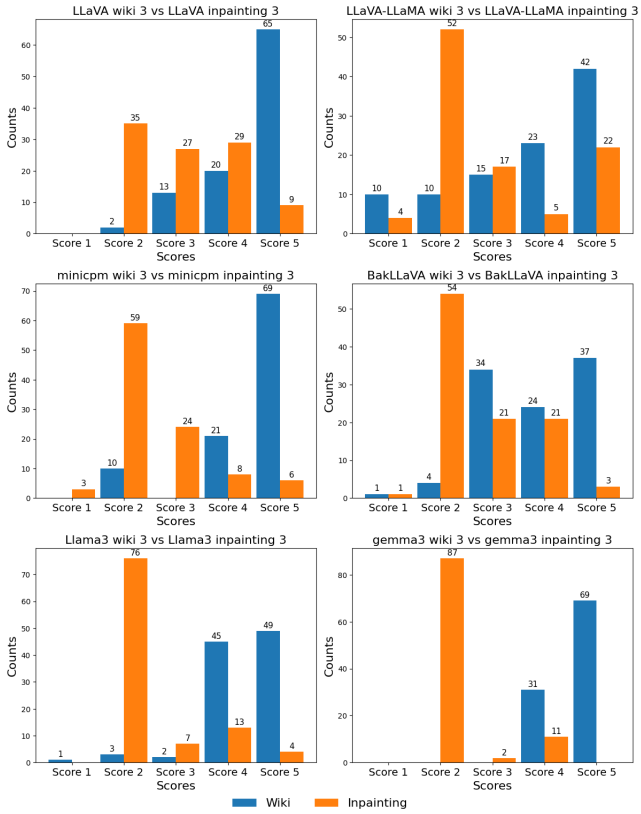
4.1.3 ***Prompt 3.*** Prompt 3 introduces another layer that requires the models to explain and provide examples as to why the image was classified as real or deepfake. This transition from a binary response to a free form explanation allows for a deeper analysis of a model's reasoning and capabilities at detecting deepfakes. Each model responses were scored on a 1 to 5 scale based on the human evaluation scoring method. The distribution of scores is visualized in figure 3, which highlights clear distinction in model performance when comparing real images and deepfake images.

For real images, models generally performed well with a majority of responses falling in scores 4 and 5. MiniCPM-V and Gemma 3 both achieved 69 responses with a score of 5 with LLaVA closely behind at 65. This demonstrates that these models were able to confidently and correctly identify real images while also providing coherent and contextually relevant justifications. Score 4, which reflects correct answers with poor or doubtful explanations were also prevalent. For example, while LLaMA 3 was not very confident and received only 49 responses with a score of 5, it still scored 45 responses with a 4. This demonstrates that even models that are less confident and have lower frequency of responses receiving a 5, these models could still provide reasonably good and correct responses.

Most models received 80 or more combined 4 and 5 score responses except for LLaVA-LLaMA and BakLLava. These two models tend to refuse to respond at a higher rate compared to the other models. Overall, hallucinations and incorrect answer were rare for real images, reinforcing that models are strong in detecting and explaining real and not augmented images.

On the other hand, the deepfake image responses demonstrates a sharp decline in model performance and an inverse distribution compared to real images which can be noticed in figure 3. The

Figure 3: Human Evaluation distribution for prompt 3. Wiki Scores tend to be confident and correct. Inpainting scores are mostly incorrect or the model refuses to respond. In general, the distribution of responses are inverse from real and deepfake images.



frequency of responses scoring a 5 dropped significantly for all models, Gemma 3 received 0, MiniCPM-V only had 5, and BakLLaVA fell from 37 to 3. This suggests that while some models can explain real images well, they struggle when the augmented features are introduced. There is also a decline in responses receiving a score of 4. For example, Llama 3 decreased from 45 to 13 responses scoring 4. In addition, score 3 responses became more common for deepfake images especially in models like MiniCPM-V at 24, LLaVA at 27 and BakLLaVA at 21. This shift demonstrates that many models when given an image with subtle anomalies to facial features hesitate more or request more context to be able to provide a decision.

The most notable aspect of model responses for deepfake images is the substantial amount of incorrect responses. MiniCPM-V had 59, LLaVA-LLaMA had 52, and Gemma 3 had 87. These scores demonstrate worse performance compared to prompt 2 on average when analyzing only if the response is correct or not (scores 4 and 5). A benefit from this style of prompt as that the models are more likely to refuse to respond or ask for more context, this is a better response than being incorrect. Prompt 3 helps reveal that even with a more focused prompt that requires explanation for reasoning, the models struggle at detecting deepfake images.

4.1.4 ***Prompt 4.*** Prompt 4 was designed to further refine the evaluation of deepfake detection by giving models a guideline to potential anomalies building upon prompt 3. By doing so, this prompt aims to examine whether increasing how specific the prompt is improves the model's ability to recognize and explain if an image is a deepfake. As seen in figure 4, responses to prompt 4 maintained similar overall trends that we saw in prompt 3 responses. The more specific prompting in prompt 4 did not result in improvements across the board, but sometimes resulting in worse performance, especially for real images.

For real images, most models maintained a distribution of score 4 and 5 responses, but the bias compared to Prompt 3 is lower. For example, MiniCPM-V dropped from 69 score 5s in prompt 3 to 45 in prompt 4, but this was compensated with a small increase of responses receiving a score of 4. Gemma 3 saw the most significant shift of scores from prompt 3 to 4, with a large increase in score 4 responses from 31 to 82. This pattern from some models suggests that while they continue to produce correct answers, they may have expressed reduced confidence when given the list of possible anomalies. On the other hand, LLaVA-LLaMA increased its number of score 5 responses for real images from 42 to 57, while keeping a stable count of score 4s, 23 in Prompt 3 and 21 in Prompt 4. This indicates that for the other models, like Llama 3, the detailed criteria may have improved clarity or supported more confident outputs.

On deepfake images, model performance remained limited, with most models still heavily weighted towards receiving a score of 2 similarly to prompt 3. Only two models stood out with improvements gained from prompt 4 which are MiniCPM-V and Gemma 3. MiniCPM-V saw an increase in performance within both scores 4 and 5, increasing from 8 to 37 and 6 to 15 respectively from prompt 3. Gemma 3 saw a significant increase in responses being deemed correct with 86 of them given a score of 4, the largest increase across any prompt for any model in deepfake detection.

Overall, the results suggest that although models can see some benefits from detailed prompts, they still lack the underlying capabilities to consistently identify subtle anomalies. The improvements observed in some models, particularly for Gemma 3, suggest that prompt engineering can enhance reasoning, but it requires more specialized prompts for each individual model.

4.2 Response Evaluation

4.2.1 ***Prompt 1 and 2.*** As explored in section 4.1 prompt 1 and 2 explored how language affects classification for the deepfake images. An additional important topic to small LLM responses is if they follow direction given the prompt. For prompt 1 and 2 the prompt specifically asked for only "yes" or "no" responses, if the model responded with anything else the response would be deemed as invalid. As seen in table 3 there is not a significant different between prompt 1 and 2. This is important because it means the language different of "deepfake" and "altered or augmented" does not pose an increase in invalid responses.

4.2.2 ***Prompt 3 and 4.*** By providing a guideline as to what the potential anomalies are, there was a noticeable structure difference in how some of the models responded. For example, MiniCPM-V responded in a more structured format where each potential anomaly was considered and either rejected or expanded upon in a step by step fashion. The structured response approach by the model saw an increase in deepfake detection compared to prompt 3 responses. This is in contrast to BakLLaVA which responses differed in structure and in general do not provide reasoning for each possible anomaly. This is reflected in the scores where this model saw a decrease in performance. The difference in quality and structure of responses can be seen in figure 5

Figure 4: Human Evaluation scoring for prompt 4. Similar trends appear like in prompt 3 with some outliers gaining significant improvements. Increasing how specific a particular prompt is written does not lead to general improvements.

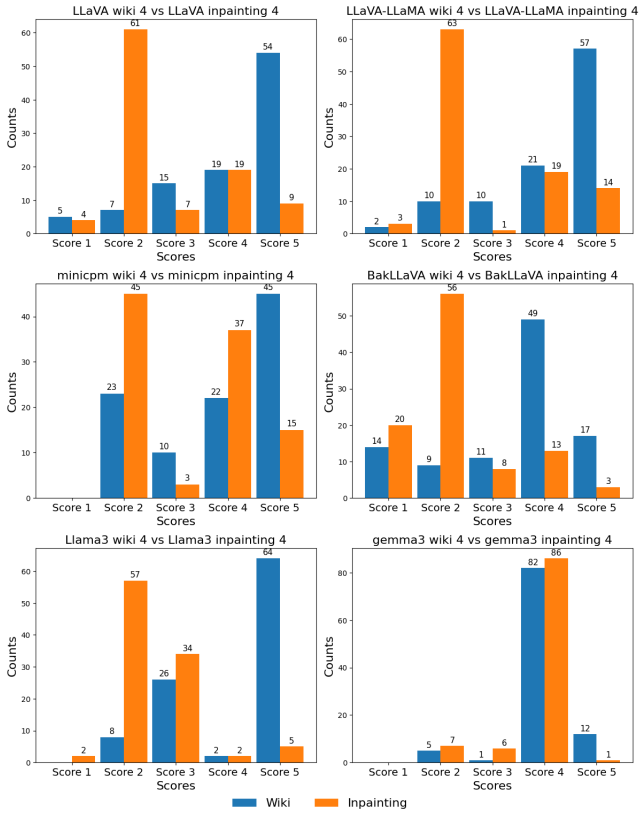
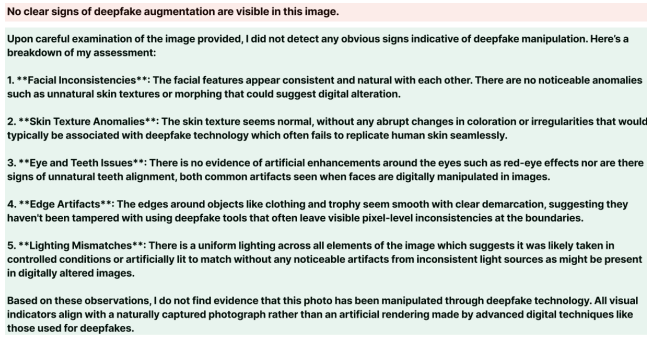


Table 3: Count for valid responses in both prompt 1 and 2. Each entry can have a max of 20000 total responses. Based on the results small LLMs still follow directions and respond correctly.

Model	**Prompt 1**	**Prompt 2**
LLaVA [8]	17964	19996
LLaVA-Llama [5]	20000	20000
MiniCPM-V [16]	20000	18183
BakLLaVA	19303	19987
LLaMA 3 [1]	17104	17093
Gemma 3 [14]	20000	20000

Figure 5: Example responses of structure in prompt 4. The red prompt is from BakLLaVA response. Green is from MiniCPM-V. Both responses are based on the same image.



5 DISCUSSION & CHALLENGES

This study provides key insights into the current capabilities and shortcomings of small open source LMMs for deepfake detection. Across all prompts, models demonstrated significantly higher accuracy and confidence when classifying real images. This suggests that LMMs have a general understanding of natural images, this can be observed in how accurate all the models were at describing the images, especially in prompt 3 and 4. This also demonstrated how deepfake detection is prompt sensitive. Performance on identifying deepfakes improved as prompt complexity and specificity increased. While there is not a significant difference in detection between prompt 3 and 4, these explanatory prompts saw increased performance compared to binary prompts. Additionally, language matters, the choice of wording influenced model behavior. For example, prompt 2, which replaces the term "deepfake" with "altered or augmented", led to a more balanced classification distribution for some models. This implies a sensitivity to prompt wording can leverage better reasoning or performance from models.

There were several challenges and limitations that emerged during this study. The use of only using Stable Diffusion Inpainting presented more challenges to the models due to only modifying facial features in a subtle way. This led to it also being difficult for human evaluators to make sure the model is correctly or incorrectly explaining the anomalies. Additionally, there is a level of bias to the evaluation of the scoring of prompt 3 and 4 because it was competed by one person. With hardware limitations the 30,000 images could not be processed, which could potentially provide more data and lead to a different distribution in the model responses.

There are multiple areas for potential improvement that can be applied to this specific study. Firstly, reducing the number of images from one method of augmentation and potentially adding other methods like Stable Diffusion v1.5, this could provide more insight as to how models respond to different techniques of manipulation. Furthermore, the human evaluation could have been much shorter, having evaluators score over a thousand prompts could lead to fatigue in evaluation. Significantly reducing the number of prompts to grade and additionally having more evaluators of different backgrounds could lead to a broader distribution that has less bias.

6 CONCLUDING REMARKS

This work presents a focused evaluation of the capabilities of small open source large multimodal models to detect deepfakes using progressively complex prompts. By leveraging a dataset with subtle facial anomalies and integrating both quantitative and qualitative analyses, it is uncovered that prompt design, model architecture, and task framing can influence model performance.
This study contributes:

- A novel, prompt driven evaluation system to classify real vs deepfake images.
- The use of a large image set featuring subtle face augmentation.
- A system that provides a quantitative and qualitative evaluation approach that combines accuracy and response quality across different types of prompts.

The findings highlight that while small LMMs can detect real images, their deepfake detection capabilities are dependent on prompt structure and types of anomalies present. For future work, there will be a larger focus on reducing bias by expanding the amount of human evaluators. Introduce additional evaluation techniques to test how model architecture can affect performance and detection. Additionally, expanding the diversity of the dataset to account for different techniques of augmentation seen across different generative methods.

REFERENCES

[1] 2024. The Llama 3 Herd of Models. arXiv:2407.21783 [cs.AI] https://arxiv.org/abs/2407.21783
[2] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. 2018. MesoNet: a Compact Facial Video Forgery Detection Network. In *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*. IEEE, 1–7. https://doi.org/10.1109/wifs.2018.8630761
[3] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikoł aj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. 2022. Flamingo: a Visual Language Model for Few-Shot Learning. In *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), Vol. 35. Curran Associates, Inc., 23716–23736. https://proceedings.neurips.cc/paper_files/paper/2022/file/960a172bc7fbf0177ccccbb411a7d800-Paper-Conference.pdf
[4] François Chollet. 2017. Xception: Deep Learning with Depthwise Separable Convolutions. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1800–1807. https://doi.org/10.1109/CVPR.2017.195
[5] XTuner Contributors. 2023. XTuner: A Toolkit for Efficiently Fine-tuning LLM. https://github.com/InternLM/xtuner.
[6] David Güera and Edward J. Delp. 2018. Deepfake Video Detection Using Recurrent Neural Networks. In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. 1–6. https://doi.org/10.1109/AVSS.2018.8639163
[7] Shan Jia, Reilin Lyu, Kangran Zhao, Yize Chen, Zhiyuan Yan, Yan Ju, Chuanbo Hu, Xin Li, Baoyuan Wu, and Siwei Lyu. 2024. Can ChatGPT Detect DeepFakes? A Study of Using Multimodal Large Language Models for Media Forensics. arXiv:2403.14077 [cs.AI] https://arxiv.org/abs/2403.14077
[8] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved Baselines with Visual Instruction Tuning. arXiv:2310.03744 [cs.CV] https://arxiv.org/abs/2310.03744
[9] Yixin Liu, Kai Zhang, Yuan Li, Zhiling Yan, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, Jianfeng Gao, Lifang He, and Lichao Sun. 2024. Sora: A Review on Background, Technology, Limitations, and Opportunities of Large Vision Models. arXiv:2402.17177 [cs.CV] https://arxiv.org/abs/2402.17177

[10] OpenAI. 2023. GPT-4V(ision) System Card. https://api.semanticscholar.org/CorpusID:263218031

[11] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. arXiv:2307.01952 [cs.CV] https://arxiv.org/abs/2307.01952

[12] Mark Scanlon, Frank Breitinger, Christopher Hargreaves, Jan-Niclas Hilgert, and John Sheppard. 2023. ChatGPT for digital forensic investigation: The good, the bad, and the unknown. *Forensic Science International: Digital Investigation* 46 (2023), 301609. https://doi.org/10.1016/j.fsidi.2023.301609

[13] Haixu Song, Shiyu Huang, Yinpeng Dong, and Wei-Wei Tu. 2023. Robustness and Generalizability of Deepfake Detection: A Study with Diffusion Models. arXiv:2309.02218 [cs.CV] https://arxiv.org/abs/2309.02218

[14] Gemma Team. 2025. Gemma 3 Technical Report. arXiv:2503.19786 [cs.CL] https://arxiv.org/abs/2503.19786

[15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention Is All You Need. arXiv:1706.03762 [cs.CL] https://arxiv.org/abs/1706.03762

[16] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, Qianyu Chen, Huarong Zhou, Zhensheng Zou, Haoye Zhang, Shengding Hu, Zhi Zheng, Jie Zhou, Jie Cai, Xu Han, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2024. MiniCPM-V: A GPT-4V Level MLLM on Your Phone. arXiv:2408.01800 [cs.CV] https://arxiv.org/abs/2408.01800

[17] Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, and Nenghai Yu. 2021. Multi-attentional Deepfake Detection. arXiv:2103.02406 [cs.CV] https://arxiv.org/abs/2103.02406