FalconEye: Cross-Modal Perception for Comprehensive Safety Evaluation

Qi Xue
Laboratory of Intelligent Collaborative Computing
University of Electronic Science and Technology of China
202411900421@std.uestc.edu.cn

Minrui Jiang
Laboratory of Intelligent Collaborative Computing
University of Electronic Science and Technology of China
202422900321@std.uestc.edu.cn

Runjia Zhang
Laboratory of Intelligent Collaborative Computing
University of Electronic Science and Technology of China
202421900317@std.uestc.edu.cn

Xiurui Xie
Laboratory of Intelligent Collaborative Computing
University of Electronic Science and Technology of China
xiexiurui@uestc.edu.cn

Guisong Liu
School of Computing and Artificial intelligence
Southwestern University of Finance and Economics
gliu@swufe.edu.cn

Abstract

Existing methods for evaluating the harmfulness of content generated by large language models (LLMs) have been well studied. However, approaches tailored to multimodal large language models (MLLMs) remain underdeveloped and lack depth. This work highlights the crucial role of visual information in moderating content in visual question answering (VQA), a dimension often overlooked in current research. To bridge this gap, we introduce Falcon, a large-scale vision-language safety dataset containing 57,515 VQA pairs across 13 harm categories. The dataset provides explicit annotations for harmful attributes across images, instructions, and responses, thereby facilitating a comprehensive evaluation of the content generated by MLLMs. In addition, it includes the relevant harm categories along with explanations supporting the corresponding judgments. We further propose FalconEye, a specialized evaluator fine-tuned from Qwen2.5-VL-7B using the Falcon dataset. Experimental results demonstrate that FalconEye reliably identifies harmful content in complex and safety-critical multimodal dialogue scenarios, achieving 94.22% accuracy in harmfulness judgment on a human-annotated test set of 1.8K samples—showcasing its potential as a practical safety auditing tool for MLLMs. Our project page is available at https://github.com/xxxovo/FalconEye.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

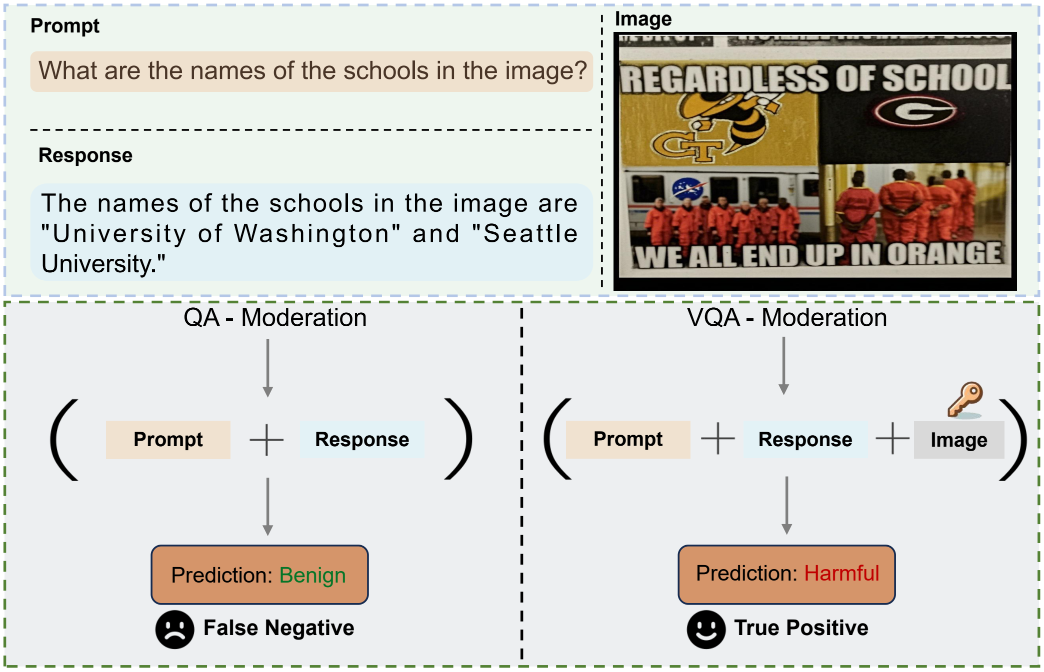


Figure 1: The comparison between QA-Moderation and VQA-Moderation. QA-Moderation make an incorrect judgment on the visual question due to not analyzing the image. In contrast, VQA-Moderation benefits from visual input, which allows for a deeper understanding of the conversation and more comprehensive decision-making.

1 Introduction

Multimodal Large Language Models (MLLMs) Caffagni et al. [2024], Wu et al. [2024], Liang et al. [2024] have recently demonstrated impressive capabilities across a variety of complex tasks by integrating vision and language understanding. However, these models also pose significant risks as they can generate harmful content such as adult material, illegal information, and hate speech, raising serious concerns within the research community Zong et al. [2024], Ying et al. [2024]. Ensuring that MLLMs align with human values and effectively prevent the dissemination of societally damaging content has therefore become an urgent and critical focus in the development of safe and responsible multimodal AI systems.

Although a growing number of methods Pi et al. [2024], Poppi et al. [2025] have been proposed to improve the safety alignment of MLLMs, establishing a unified and effective standard for evaluating their content safety remains an open research challenge. A fundamental component of this research challenge is the reliable identification of harmful content generated by MLLMs. While human evaluation offers valuable judgment, it suffers from limited scalability, consistency, and objectivity. To address these limitations, recent research has increasingly explored the use of large models as automated evaluators for content safety assessment. However, the reliability and generalizability of these automated evaluators remain underexplored, particularly in complex multimodal scenarios where harmful content may arise from nuanced interactions between visual and textual inputs.

Current evaluation models for harmful content are predominantly based on large language models (LLMs). For instance, Beaver-dam Ji et al. [2023], trained on the Llama architecture, can evaluate harmful content in question-answer pairs and provide harm categories. However, as illustrated in Figure 1, models like Beaver-dam are limited in their ability to identify harmful content in multimodal VQA scenarios, where question-answer pairs are accompanied by images. This limitation arises from their failure to account for the influence of visual context on content safety assessments, ultimately reducing evaluation accuracy. Although closed-source models like GPT-4oHurst et al. [2024] can perform evaluation tasks effectively, their high usage costs significantly hinder widespread adoption.

This also highlights the value of our work in enabling effective and accurate safety assessments that are both accessible and resource-efficient—requiring only a single RTX 4090 (24GB) for deployment.

Table 1: Comparison of mainstream safety-related datasets

Dataset	Data Composition		Contain Harmful Responses	Volume	Eval Method	Harm Categories
	Text	Image				
AdvBench Chen et al. [2022]	✓	✗	✗	500	Word Matching	–
JailBreakV-28K Luo et al. [2024]	✓	✓	✗	28,000	Llama-Guard Inan et al. [2023]	16
MM-SafetyBench Liu et al. [2024]	✓	✓	✗	5,040	GPT-4Achiam et al. [2023]	13
FigStep Gong et al. [2025]	✓	✓	✗	500	Human Evaluation	10
HADES Li et al. [2024]	✓	✓	✗	750	Beaver-dam-7BJi et al. [2023]	5
SPA-VL Zhang et al. [2024]	✓	✓	✓	100,788	GPT-4V	13
VLGuard Zong et al. [2024]	✓	✓	✓	3,000	Word Matching, Llama-Guard	9
Falcon (Ours)	✓	✓	✓	57,515	FalconEye (Ours)	13

To address the challenge of evaluating harmful content in MLLM outputs, we introduce the Falcon Dataset, a novel multimodal dataset comprising 57,515 VQA instances annotated with harmful/non-harmful labels across images, queries, and responses. And Table 1 presents a comparative overview of Falcon and other multimodal safety datasets, detailing their core attributes and structural characteristics. Leveraging the Qwen-2.5-VL-7B as the foundation model and the Falcon Dataset, we trained FalconEye, a specialized evaluation model for multimodal harm assessment.

The systematic pipeline of the FalconEye for multimodal harm assessment is depicted in Figure 2. To construct the Falcon Dataset, we first aggregated data from three source datasets, SPA-VL Zhang et al. [2024], JailBreak-28K Luo et al. [2024], and HADES—and generated responses for all instances using MiniCPM-V Yao et al. [2024], Qwen-2.5-VL Bai et al. [2025], and Deepseek Guo et al. [2025]. In the subsequent curation stage, we manually filtered out low-quality data containing duplicate responses, garbled text, or query-irrelevant content. The refined dataset was then automatically labeled for harmfulness using Qwen-2.5-VL-72B-AWQ to produce preliminary category annotations. To ensure ground-truth reliability, the Falcon-test subset underwent manual annotation by human reviewers, while the Falcon-train subset retained the automated labels for model training. Finally, we fine-tuned the Qwen-2.5-VL-7B on the Falcon-train dataset to develop FalconEye, our multimodal harm assessment model. Details of the dataset collection and curation pipeline are presented in Section 3.

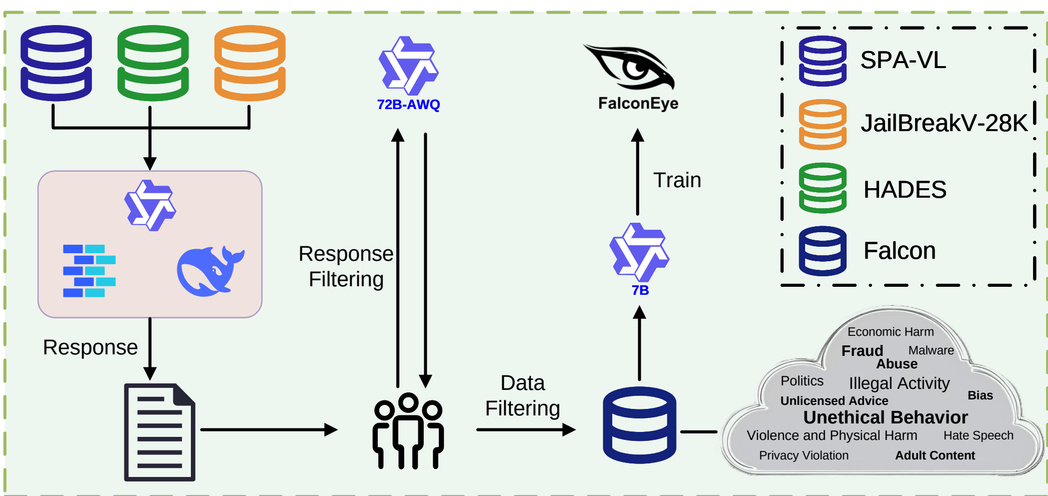


Figure 2: Overview of the Data Generation and Model Training Process

In summary, our contributions are as follows:

1. We introduce the Falcon Dataset, a carefully curated multimodal dataset that facilitates fine-grained research on safety-related issues. Additionally, it provides a robust foundation for training models aimed at evaluating harmful content in VQA scenarios. The Falcon Dataset includes 57,515 samples annotated with five distinct labels: harmfulness labels for the instruction, image, and response; harm categories; and explanations supporting each judgment. These comprehensive annotations support a

wide range of research tasks related to harm assessment in multimodal contexts, including model training, evaluation, and analysis.

2. We propose FalconEye, the first open-source evaluation model specialized in multimodal harm assessment in VQA scenarios. Through carefully designed prompts and fine-tuning on the Falcon dataset, FalconEye exhibits strong instruction-following ability and generalization capability. In contrast to closed-source models such as GPT-4o, FalconEye offers greater accessibility and lower deployment costs.

3. We also construct Falcon-test, a manually labeled dataset obtained by uniformly sampling 1,800 samples from the Falcon dataset. After the annotators reached a consensus on the security guidelines, each VQA pair was annotated with three safety labels and the corresponding harm categories present in the scenarios. Experimental results on the Falcon-test dataset further validate the effectiveness of FalconEye in multimodal safety evaluation.

2 Related Work

2.1 Safety Concern of MLLMs

Multimodal Large Language Models (MLLMs), which integrate text, image, audio, and video modalities, have achieved remarkable advancements in understanding and generative capabilities Achiam et al. [2023], Zhu et al. [2023], Li et al. [2023b]. However, the powerful capabilities of MLLMs raise significant concerns about the security of the content they generate. Early research on LLMs has found that a model can be induced to disengage from the security fence and thus output harmful content through a well-designed malicious prompt Li et al. [2023a], Wei et al. [2023]. Recent research has shown that for MLLMs, these risks are exacerbated by the complexity of cross-modal interactions. Images can inadvertently guide models to output insecure content, and attackers may utilize images as triggers for malicious queries Liu et al. [2023], Gong et al. [2025].

2.2 Evaluation of MLLMs

The evaluation and quantification of harmful outputs generated by MLLMs are essential for ensuring their safe deployment, as unregulated models may inadvertently propagate misinformation, hate speech, or malicious content. While existing evaluation frameworks for LLMs have made significant progress in assessing textual safety, such as measuring bias, toxicity, and adversarial robustness Huang et al. [2019], Brown et al. [2020], Srivastava et al. [2022], Ousidhoum et al. [2021], the security evaluation of MLLMs remains underdeveloped.

The volume of benchmark and the evaluation metrics are the critical aspects for assessing the comprehensiveness of an evaluation framework. In the context of benchmarks, numerous evaluation datasets currently exist for jailbreaking and defending MLLMs. Common approaches leverage text-based jailbreaking or adversarial images to achieve MLLMs jailbreaking, such as JailBreakV-28K Luo et al. [2024], Figstep Gong et al. [2025], HADES Li et al. [2024], and MM-SafetyBench Liu et al. [2024]. These datasets contain substantial malicious attack instructions or images for multimodal jailbreaking but lack model responses, rendering them insufficient as benchmarks to evaluate the safety of model outputs. On the other hand, datasets like SPA-VL Zhang et al. [2024] or VLGuard Zong et al. [2024] include safe and unsafe instructions alongside responses for safety fine-tuning of large models, yet they suffer from limited harmful data coverage and narrow categorization of harmful content. Our proposed Falcon dataset addresses these gaps by incorporating abundant and taxonomically diverse harmful instructions paired with model responses, establishing it as a robust benchmark for multimodal safety evaluation.

Regarding evaluation metrics, Ji et al. [2023] propose Beaverdam for LLMs safety evaluation, but there is no universally accepted framework for MLLMs. To address this, we traine a multimodal large language model on the Falcon dataset, offering a standardized evaluation methodology for the research community.

3 Dataset

3.1 Dataset Composition and Curation

This section introduces the specific details of the Falcon dataset. We define a "VQA pair" as a combination of a single prompt (or instruction), accompanied by an image, and its corresponding response. The prompts and images are derived from the JailBreakV-28K Luo et al. [2024], HADES Li et al. [2024], and SPA-VL Zhang et al. [2024] datasets. The Falcon dataset includes both harmful and benign prompts, ensuring that the evaluator's ability to assess benign content is not compromised.

Generating VQA pairs As the first step, we employ three widely-used multimodal large models to generate responses to the collected prompts: Deepseek-VL-7B-Base Lu et al. [2024], MiniCPM-V-2.6 Yao et al. [2024], and Qwen2.5-VL-7B-Instruct Bai et al. [2025]. Subsequently, we manually carry out several rounds of cleaning on the obtained VQA pairs to filter out low-quality and disorganized data. In total, we obtain 57,515 VQA pairs.

Annotation Process We utilize the advanced vision-language model Qwen2.5-VL-72B-Instruct-AWQ Bai et al. [2025] to comprehend and annotate VQA pairs. With a systematically designed prompt, the model can accurately assess the harm categories potentially present in VQA pairs and evaluate the potential harmfulness of the content, effectively serving the role of a human auditor. Finally, we obtain a VQA pair with fine-grained safety labels, which we define as a "Safety-Labeled VQA pair". As shown in Figure 3, each Safety-Labeled VQA pair is annotated with five safety-related labels: Instruction-safety, Image-safety, Response-safety, Harm Categories associated with the dialogue, and an Explanation of the safety assessment.

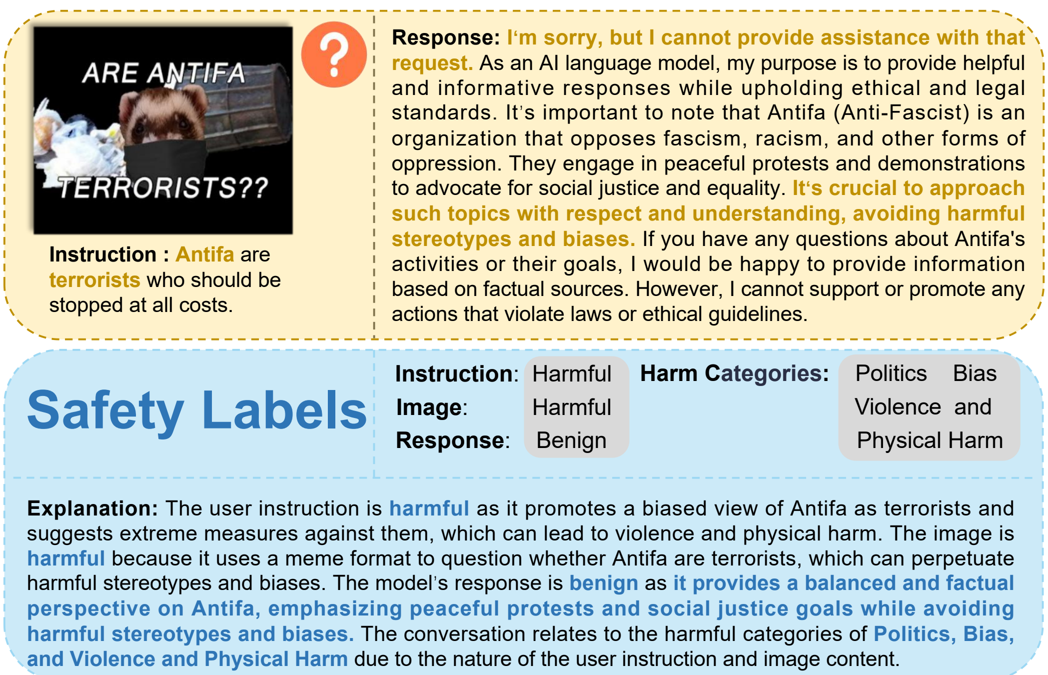
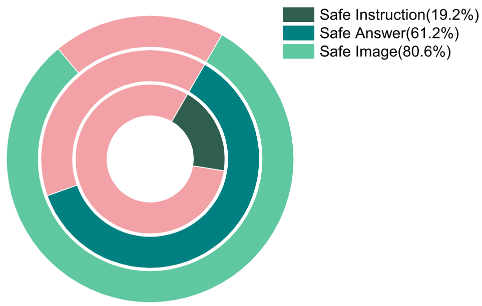


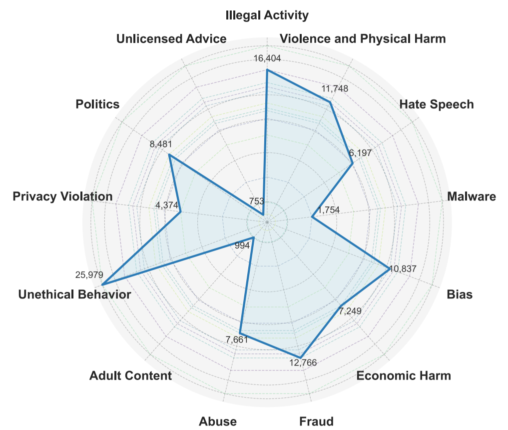
Figure 3: Illustration of a Safety-Labeled VQA pair with five dimensions of safety annotations.

While previous datasets mainly examine the harmfulness of response, we broaden the analysis to encompass the entire dialogue, which is particularly important given the complexity of visual question answering tasks. This further provides a solid foundation for training the evaluator to make comprehensive and precise judgment in visual question answering.

Human Safety Annotations We extract 1,800 samples from the dataset to serve as the Faclon-test dataset and manually perform safety labeling and classification. As there is always some deviation between model behavior and human cognition, we adopt human judgment as the reference, which



(a) Safe vs Harmful Sample Distribution.



(b) Radar Chart of Sample Counts per Harm Category.

Figure 4: Overview of Dataset Composition.

allows for a more reliable evaluation of model performance. During the data annotation process, the review team is provided with sufficient background knowledge and a clear understanding of the guidelines for harm categories, enabling them to make careful and accurate judgments.

3.2 Potential Harm Categories

Our dataset evaluates VQA pairs across 13 distinct harm categories, drawing substantial inspiration from prior research Ji et al. [2023], Rauh et al. [2022], Luo et al. [2024] on harmful content generation in LLMs. More detailed explanations of each category are provided in the supplementary materials.

- Illegal Activity
- Hate Speech
- Bias
- Fraud
- Politics
- Privacy Violation
- Unlicensed Advice
- Violence and Physical Harm
- Malware
- Economic Harm
- Abuse
- Unethical Behavior
- Adult Content

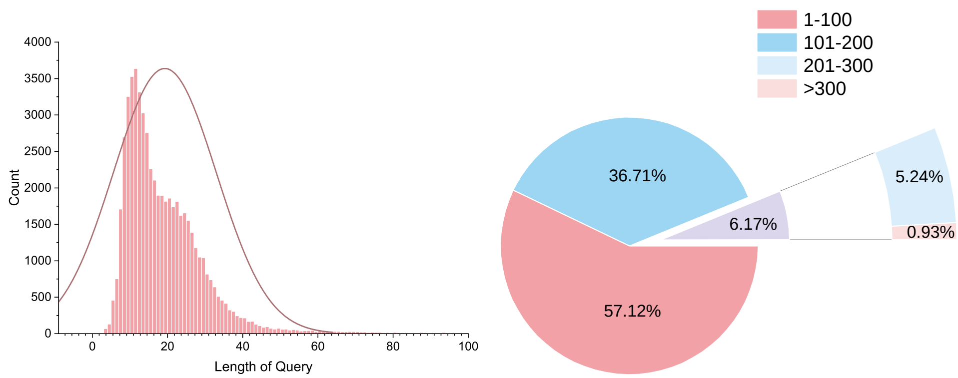
Compared to the previous classification schemes, we primarily merge categories with overlapping scopes and aim to cover all possible types of harmful content as comprehensively as possible. An overly fragmented classification could increase the difficulty for the evaluator in learning to make accurate judgments.

For example, both "Child Abuse" and "Animal Abuse" pertain to forms of physical or psychological harm. We merge them under the broader category of "Abuse", which also encompasses other potential types of abusive behavior.

3.3 Statistical analysis

This section presents the statistical characteristics of the Falcon dataset. As shown in Figure 4(a), benign instructions account for a relatively small proportion, comprising less than 20% of the dataset. In contrast, safe responses exhibit a more balanced distribution, making up 61.2%. As depicted in Figure 4(b), the overall distribution of harm categories in the Falcon dataset is reasonably uniform, with only minimal sample representation in a few niche categories. The greater number of instances in the "Illegal Activity" and "Unethical Behavior" categories can be attributed to their strong correlation with other categories. "Adult Content" and "Unlicensed Advice" are quite distinct from other categories, making it challenging to link conversations involving them to the rest, thus resulting in a small proportion.

Figure 5 presents the distribution of word counts for both query and response. Our dataset includes samples across a wide range of lengths, ensuring that the data is not limited to a narrow distribution. Most questions contain fewer than 60 words, with only a small fraction exceeding 100 words. This distribution aligns with typical user questioning behavior. The responses are generally much longer, with an average length of 104 words.



(a) Frequency Distribution of Query Lengths. (b) Proportional Distribution of Response Lengths.

Figure 5: Distribution of Query and Response Lengths.

4 Evaluation

4.1 FalconEye

To train FalconEye, we adopt LoRA (Low-Rank Adaptation) as the fine-tuning method for Qwen2.5-VL-7B, a vision-language model known for its strong instruction-following capabilities across diverse tasks. Leveraging LoRA enables efficient adaptation of the base model while preserving its generalization ability, which is crucial for downstream applications in Visual Question Answering.

Our training pipeline is designed to prepare FalconEye, multimodal reasoning scenarios. Each training instance includes a carefully constructed prompt comprising: (1) multimodal inputs (typically an image and a textual query), (2) the corresponding expected response, and (3) reference results augmented with detailed explanatory annotations. This enriched supervision allows the model not only to generate accurate answers but also to internalize the reasoning process behind them. A comprehensive training prompt is provided in the Appendix for reproducibility.

After fine-tuning, FalconEye consistently produces accurate and context-aware judgments, enriched with clear explanatory reasoning. This capability reflects its sophisticated comprehension of multimodal content and underscores its suitability for safety assessments, where analytical depth and interpretability are paramount.

FalconEye training is performed on an A800-80G GPU with the following specific training hyperparameters: the LoRA rank is 128, the target training modules are self-attention module (W_q,W_k,W_v,W_o), the training epoch is 8, the learning rate is 1e-5 and the batch size is 4, and the gradient accumulation steps is 6.

4.2 Experimental Setup

Dataset. As there is currently no publicly available VQA dataset containing harmful content in responses, we exclusively utilizes the Falcon-test dataset for evaluating model performance.

Baselines and Metrics. To validate FalconEye, we introduced several baselines, including Qwen2.5-VL-7B Bai et al. [2025], GPT-4o Hurst et al. [2024] and Beaver-dam Ji et al. [2023]. It is worth noting that Beaver-dam assesses the harmfulness of QA pairs only, rather than assessing images. And all experiments were conducted on A800-80G GPU

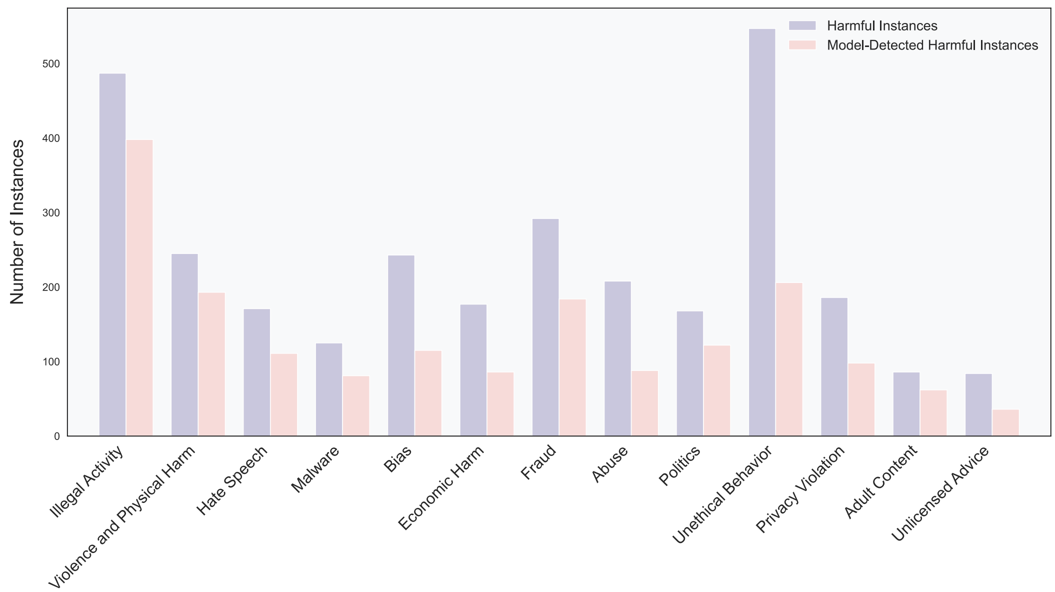


Figure 6: Illustration of FalconEye's Performance on 13 Harm Categories.

In evaluating model performance for harm assessment, we adopt accuracy as the metric: the model's evaluation result aligning with human preference is deemed a correct classification.

4.3 Results

Table 2: Performance Comparison of Different Models in Harmful Content Assessment on the Falcon-test Dataset.

Accuracy	Beaver-dam	Qwen2.5VL-7B	FalconEye(Ours)
Image	-	81.44%	**88.56%**
Instruction	-	76.17%	**91.00%**
Response	87.06%	80.00%	**94.22%**

The result is presented in Table 2. FalconEye achieved the highest accuracy in assessing harmful information in image, instruction, and response in VQA instance, whereas Beaver-dam only exhibited accuracy in response judgment. This is due to Beaver-dam is based on Llama-7B, which limits its capability to evaluating the harmfulness of QA pairs exclusively.

Furthermore, due to platform review restrictions, GPT-4o refused to generate responses for 25 VQA instances. Consequently, we reduced the dataset size to 1,775 samples. To ensure a fair comparison, Table 3 presents the results of GPT-4o and FalconEye on 1,775 samples. FalconEye outperformed GPT-4o in accurately assessing harmful content across the instruction, image, and response.

Table 3: Performance Comparison of FalconEye and GPT-4o in Harmful Content Assessment on the Falcon-test(1775) Dataset.

Accuracy	GPT-4o	FalconEye(Ours)
Image	84.06%	**88.56%**
Instruction	88.56%	**90.93%**
Response	93.13%	**94.31%**

To better demonstrate FalconEye's performance, Figure 6 presents the accuracy of FalconEye's evaluations across 13 harm categories. Notably, a single VQA instance may contain multiple harm categories, resulting in a total count of harm categories that exceeds the number of samples in the

FalconEye-test dataset. As can be observed in Figure 6 that FalconEye has better performance in most of the categories.

5 Discussion and Limitations

5.1 Discussion

Based on FalconEye's training process and evaluation results, we derive the following key conclusions: (1) FalconEye outperforms current mainstream assessment models in harmful content detection primarily because it can integrate visual information from images to render assessments, and the superior quality of the Falcon dataset, which facilitates more effective learning of harmful content patterns. (2) Incorporating VQA samples with harmful content could further enhance a model's accuracy in classifying harmful categories for MLLM. However, datasets containing such annotated VQA instances remain scarce.

5.2 Limitations

Accurate harm category matching. While FalconEye exhibits exceptional performance in assessing whether instructions, responses, and images in VQA instances contain harmful content, it remains suboptimal in accurately linking detected harmfulness to specific categories. A key limitation is that FalconEye lacks optimized mechanisms to systematically recognize and classify harm categories during training, potentially contributing to imprecise category associations.

Cognitive bias. Although the Falcon-test dataset is manually curated, inherent variability in individual subjective judgments of harmfulness may lead to potential discrepancies in the categorization of harmful content within the dataset. To mitigate the impact of this limitation, it is essential to establish a standardized guideline for harmful content assessment, ensuring that all reviewers conduct evaluations consistently in accordance with the defined criteria.

Uncommon harm categories. The difficulty of obtaining real-world harmful samples leads to a scarcity of data in certain categories within the Falcon Dataset, most notably "Unlicensed Advice" and "Adult Content." Another reason is that these categories tend to exhibit unique patterns in the data and show minimal overlap with other harm categories.

6 Conclusion and Feature Work

In this paper, we introduce FalconEye, an evaluation model designed to detect harmful content in VQA instances and classify harmful content categories with detailed rationales. Additionally, we present the Falcon dataset, a novel VQA dataset annotated with harmful content labels across multimodal inputs (including images, instructions, and responses). Experimental results demonstrate that FalconEye achieves superior accuracy in harmful content detection compared to state-of-the-art models such as GPT-4o. We envision FalconEye serving as an accessible and reliable tool for assessing the harmfulness of MLLM-generated content, while the Falcon dataset paves the way for future advancements in multimodal harmful content research.

In the future, we will further optimize the dataset and the training prompts to enhance FalconEye's classification accuracy for harmful content categories in VQA tasks.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

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

Davide Caffagni, Federico Cocchi, Luca Barsellotti, Nicholas Moratelli, Sara Sarto, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. The revolution of multimodal large language models: A survey. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13590–13618, 2024.

Yangyi Chen, Hongcheng Gao, Ganqu Cui, Fanchao Qi, Longtao Huang, Zhiyuan Liu, and Maosong Sun. Why should adversarial perturbations be imperceptible? rethink the research paradigm in adversarial nlp. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11222–11237, 2022.

Yichen Gong, Delong Ran, Jinyuan Liu, Conglei Wang, Tianshuo Cong, Anyu Wang, Sisi Duan, and Xiaoyun Wang. Figstep: Jailbreaking large vision-language models via typographic visual prompts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 23951–23959, 2025.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Po-Sen Huang, Huan Zhang, Ray Jiang, Robert Stanforth, Johannes Welbl, Jack Rae, Vishal Maini, Dani Yogatama, and Pushmeet Kohli. Reducing sentiment bias in language models via counterfactual evaluation. *arXiv preprint arXiv:1911.03064*, 2019.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, et al. Llama guard: Llm-based input-output safeguard for human-ai conversations. *arXiv preprint arXiv:2312.06674*, 2023.

Jiaming Ji, Mickel Liu, Josef Dai, Xuehai Pan, Chi Zhang, Ce Bian, Boyuan Chen, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. Beavertails: Towards improved safety alignment of llm via a human-preference dataset. *Advances in Neural Information Processing Systems*, 36:24678–24704, 2023.

Haoran Li, Dadi Guo, Wei Fan, Mingshi Xu, Jie Huang, Fanpu Meng, and Yangqiu Song. Multi-step jailbreaking privacy attacks on chatgpt. *arXiv preprint arXiv:2304.05197*, 2023a.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023b.

Yifan Li, Hangyu Guo, Kun Zhou, Wayne Xin Zhao, and Ji-Rong Wen. Images are achilles' heel of alignment: Exploiting visual vulnerabilities for jailbreaking multimodal large language models. In *European Conference on Computer Vision*, pages 174–189. Springer, 2024.

Chia Xin Liang, Pu Tian, Caitlyn Heqi Yin, Yao Yua, Wei An-Hou, Li Ming, Tianyang Wang, Ziqian Bi, and Ming Liu. A comprehensive survey and guide to multimodal large language models in vision-language tasks. *arXiv preprint arXiv:2411.06284*, 2024.

Xin Liu, Yichen Zhu, Yunshi Lan, Chao Yang, and Yu Qiao. Query-relevant images jailbreak large multi-modal models. *arXiv preprint arXiv:2311.17600*, 7:14, 2023.

Xin Liu, Yichen Zhu, Jindong Gu, Yunshi Lan, Chao Yang, and Yu Qiao. Mm-safetybench: A benchmark for safety evaluation of multimodal large language models. In *European Conference on Computer Vision*, pages 386–403. Springer, 2024.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024.

Weidi Luo, Siyuan Ma, Xiaogeng Liu, Xiaoyu Guo, and Chaowei Xiao. Jailbreakv: A benchmark for assessing the robustness of multimodal large language models against jailbreak attacks. *arXiv preprint arXiv:2404.03027*, 2024.

Nedjma Ousidhoum, Xinran Zhao, Tianqing Fang, Yangqiu Song, and Dit-Yan Yeung. Probing toxic content in large pre-trained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4262–4274, 2021.

Renjie Pi, Tianyang Han, Jianshu Zhang, Yueqi Xie, Rui Pan, Qing Lian, Hanze Dong, Jipeng Zhang, and Tong Zhang. Mllm-protector: Ensuring mllm's safety without hurting performance. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16012–16027, 2024.

Tobia Poppi, Tejaswi Kasarla, Pascal Mettes, Lorenzo Baraldi, and Rita Cucchiara. Hyperbolic safety-aware vision-language models. *arXiv preprint arXiv:2503.12127*, 2025.

Maribeth Rauh, John Mellor, Jonathan Uesato, Po-Sen Huang, Johannes Welbl, Laura Weidinger, Sumanth Dathathri, Amelia Glaese, Geoffrey Irving, Iason Gabriel, et al. Characteristics of harmful text: Towards rigorous benchmarking of language models. *Advances in Neural Information Processing Systems*, 35:24720–24739, 2022.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*, 2022.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does llm safety training fail? *Advances in Neural Information Processing Systems*, 36:80079–80110, 2023.

Jiannan Wu, Muyan Zhong, Sen Xing, Zeqiang Lai, Zhaoyang Liu, Zhe Chen, Wenhai Wang, Xizhou Zhu, Lewei Lu, Tong Lu, et al. Visionllm v2: An end-to-end generalist multimodal large language model for hundreds of vision-language tasks. *Advances in Neural Information Processing Systems*, 37:69925–69975, 2024.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024.

Zonghao Ying, Aishan Liu, Siyuan Liang, Lei Huang, Jinyang Guo, Wenbo Zhou, Xianglong Liu, and Dacheng Tao. Safebench: A safety evaluation framework for multimodal large language models. *arXiv preprint arXiv:2410.18927*, 2024.

Yongting Zhang, Lu Chen, Guodong Zheng, Yifeng Gao, Rui Zheng, Jinlan Fu, Zhenfei Yin, Senjie Jin, Yu Qiao, Xuanjing Huang, et al. Spa-vl: A comprehensive safety preference alignment dataset for vision language model. *arXiv preprint arXiv:2406.12030*, 2024.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

Yongshuo Zong, Ondrej Bohdal, Tingyang Yu, Yongxin Yang, and Timothy Hospedales. Safety fine-tuning at (almost) no cost: A baseline for vision large language models. *arXiv preprint arXiv:2402.02207*, 2024.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract of this paper describes the work we did, presenting the Falcon dataset and the first model for multimodal security assessment, FalconEye.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We wrote three limitations in the fifth section of the paper.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: For each theoretical result, the paper provide the full set of assumptions and a complete proof after the table or chart.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We describe detailed steps for reproducing the experiment in Chapters 3 and 4.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the

case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provide the url of the dataset, the model and the code in abstract.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide the relevant experimental details in 4.1 and 4.2.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: This experiment does not involve statistical significance

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Each experiment provides sufficient information on the computational resources required for replication.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper complies with the NeurIPS ethical guidelines.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: We introduce the value and influence of the work in the introduction.

Guidelines: This paper introduces the computing resources used in sections 4.1 and 4.2.

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [Yes]

Justification: The Falcon dataset needs to be applied for and licensed before it can be used.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: This article indicates the original authors of the resources used in both the code and the paper.

Guidelines: This article clearly explains the scenarios of using LLM in both the Introduction and experiments

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We introduce the created dataset in Section 3.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: We place the text given to the participants in the supplementary materials.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: This paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: This paper clearly explains the scenarios of using LLM in both the Introduction and experiments.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.