Survey of Post-2023 Techniques for Safely Aligning Text-Image-Audio Language Models

1. Introduction

The field of artificial intelligence has witnessed a paradigm shift with the advent of Large Language Models (LLMs), which have demonstrated remarkable capabilities across a wide array of general tasks through simple prompting, negating the necessity for task-specific training.[1] Building upon this foundation, Multimodal Large Language Models (MLLMs) have emerged, extending the capabilities of LLMs to process and understand data from diverse modalities, including visual, auditory, and textual information.[1] This integration of multimodal data holds immense potential for tackling complex real-world applications that necessitate reasoning across different forms of information, such as autonomous driving systems interpreting visual scenes and auditory cues alongside textual navigation instructions, or intelligent personal assistants understanding user requests involving images, spoken commands, and written text.[1] The capacity of these models to process information in a manner more akin to human cognition, which inherently integrates multiple senses, signifies a substantial step forward in the pursuit of more versatile and intelligent artificial systems.

However, the development and deployment of these powerful MLLMs are accompanied by critical challenges, particularly in ensuring their safety and alignment with human values and preferences.[1] Issues related to the truthfulness of their generated content (hallucinations), the potential for generating harmful or biased outputs, the need for robust reasoning capabilities across modalities, and the overall alignment with human ethical considerations and expectations remain significant areas of concern.[1] Notably, current state-of-the-art MLLMs have often not undergone the rigorous alignment stages typically employed for unimodal LLMs, such as reinforcement learning from human preferences (RLHF), with many only progressing to supervised fine-tuning. This gap leaves critical safety issues, including the generation of factually incorrect, biased, or even harmful content, inadequately addressed.[1] The potential for these models to produce unreliable or unsafe outputs if not properly guided underscores the paramount importance of developing effective and robust alignment techniques.

Given the rapid advancements in MLLMs and the increasing recognition of the critical need for safety, the period following 2023 has likely seen significant developments in research dedicated to addressing these challenges.[3] This survey aims to provide a comprehensive overview of the techniques published since 2023 that focus on safely aligning MLLMs capable of processing text, image, and audio data. By examining the latest methodologies and findings, this report seeks to illuminate the current state of the art in this rapidly evolving field, with a particular emphasis on approaches that ensure the generated content is not only accurate and helpful but also safe and aligned with human values.

The remainder of this report is organized as follows: Section 2 provides background information on the architecture and training paradigms of multimodal LLMs. Section 3 outlines the specific challenges encountered when aligning MLLMs that process text, image, and audio. Section 4 presents a survey of post-2023 alignment techniques, categorized by their approach. Section 5 delves into the critical safety aspects in the alignment of these models. Section 6 discusses the datasets and benchmarks used for evaluating alignment and safety. Section 7 explores future directions and open challenges in the field, and finally, Section 8 concludes the report.

2. Background on Multimodal LLMs

The architecture of MLLMs typically adopts a modular design, comprising three key components: modality encoders, a Large Language Model (LLM) backbone, and a modality interface that serves to connect the encoders with the LLM.[2] Modality encoders are responsible for processing the raw input data from each modality, such as images and audio, and transforming it into a more compact and semantically meaningful representation.[3] For image encoding, Vision Transformers (ViTs) and their derivatives have become dominant, often pre-trained using contrastive learning on image-text pairs, as exemplified by CLIP (Contrastive Language-Image Pre-training).[2] Similarly, for audio encoding, models like Wav2Vec 2.0 and CLAP (Contrastive Language-Audio Pre-training) are frequently employed, the latter also utilizing contrastive learning on audio-caption pairs to align audio embeddings with text embeddings in a shared space.[2] The core of the MLLM architecture is the LLM itself, which acts as the central processing unit or "brain," responsible for understanding and reasoning across the different modalities and ultimately generating text-based outputs.[2]

Bridging the gap between the continuous representations produced by the modality encoders and the discrete token-based input expected by LLMs is the role of the modality interface.[3] Various techniques are used for this purpose, including learnable connectors such as the Query-based Transformer (Q-Former) employed in models like BLIP-2, which uses a set of learnable query tokens to interact with visual features and align them with textual embeddings.[2] Another common approach is linear projection, as seen in LLaVA, where a simple linear multi-layer perceptron (MLP) is used to directly map the features extracted by the image encoder into a vector space compatible with the LLM's text embeddings.[2] Additionally, some models incorporate more intricate fusion mechanisms, potentially involving concatenation or attention mechanisms, to enable a deeper interaction and integration of features from different modalities.[3]

The training of MLLMs typically involves several stages. Initially, the modality encoders are often pre-trained on large-scale unimodal or paired multimodal datasets to learn robust and generalizable representations for each modality. For instance, CLIP is pre-trained on a vast dataset of image-text pairs, while CLAP is trained on audio-caption pairs.[2] This pre-training establishes an initial alignment between the modalities. Subsequently, instruction fine-tuning (IFT) has emerged as a crucial step, where the MLLM is trained on datasets comprising instructions paired with multimodal inputs (including text, images, and sometimes audio) and the desired text outputs.[2] This phase enables the model to learn to follow instructions that require understanding and reasoning across different modalities. Finally, alignment fine-tuning is often performed to further refine the model's behavior, ensuring it aligns with human preferences for helpfulness, truthfulness, and safety. Techniques like Reinforcement Learning from Human Feedback (RLHF) or Direct Preference Optimization (DPO) are commonly used during this stage.[1] The scaling of LLM parameters has also been instrumental in achieving performance gains in MLLMs, similar to the impact of increasing input resolution, leading to improvements in NLP task accuracy, contextual understanding, and problem-solving versatility.[2]

The evolution of MLLMs leading up to the post-2023 period saw significant advancements. Models like GPT-4V demonstrated remarkable multimodal capabilities, showcasing the potential for seamless integration of visual and textual information.[3] MiniGPT-4 explored the connection between pre-trained vision encoders and large language models, highlighting the effectiveness of leveraging existing unimodal models.[3] AudioLLM focused on incorporating audio processing into the LLM framework, expanding the range of modalities that could be understood.[6] LaViLa showcased progress in video-language understanding, indicating the growing interest in handling temporal multimodal data.[6] These models, among others, laid the crucial groundwork and inspired the rapid research and development that has characterized the post-2023 era in the field of multimodal language model alignment and safety.[3] The impressive capabilities demonstrated by these earlier models served as a catalyst, spurring further investigation and innovation in the quest to create more intelligent and safer multimodal AI systems.

3. Challenges in Aligning Text-Image-Audio MLLMs

Aligning MLLMs that process text, image, and audio presents a unique set of challenges that are significantly more complex than those encountered in aligning models dealing with fewer modalities.[11] The integration of three distinct data types necessitates the development of sophisticated techniques to ensure consistent and coherent representations across these diverse forms of information. The intricate relationships and dependencies that exist between text, images, and audio in the real world must be accurately learned and consistently maintained by the model.

A primary hurdle lies in achieving true cross-modal understanding and reasoning.[11] Enabling the model to comprehend the complex and often implicit relationships between these three modalities is a non-trivial task. For instance, the model needs to understand how a textual description relates to the objects and actions depicted in an image and how these, in turn, are connected to the sounds that might be present in an accompanying audio track. Developing effective cross-modal fusion mechanisms is crucial for this, allowing for a seamless interaction and flow of information between the different modalities so that the model can perform joint reasoning and draw inferences based on the combined input.[2]

Maintaining coherence and consistency across the modalities in both the model's internal representations and its generated outputs is another significant challenge. The model must ensure that its understanding of the input data is consistent across text, image, and audio, and that any generated response (typically text, but potentially other modalities in advanced models) aligns with the information presented in all three input modalities. A key issue in this regard is multimodal hallucination, where the model generates content that is not faithfully grounded in the provided input data across one or more of the modalities.[2] This can manifest as textual descriptions that mention objects not visible in the image or audible in the audio, images that depict elements not described in the text or implied by the audio, or audio tracks that contain sounds inconsistent with the textual narrative and visual scene.

The availability of large-scale, high-quality training datasets with precisely aligned text, image, and audio information is also a significant limiting factor.[11] Creating such datasets often requires substantial resources and effort for data collection and annotation, making them less abundant compared to unimodal or bi-modal datasets. The quality of this data directly impacts the effectiveness of alignment techniques, as noisy, incomplete, or misaligned data can hinder the model's ability to learn accurate cross-modal representations and lead to suboptimal alignment.

Finally, defining what constitutes "aligned" behavior in a tri-modal setting and developing appropriate metrics to measure it presents a considerable challenge.[11] Establishing objective and reliable evaluation metrics that can assess the coherence, consistency, and semantic fidelity of the model's understanding and generation across text, image, and audio is essential for tracking progress and comparing different alignment techniques. These metrics need to go beyond simple accuracy measures and capture the nuances of cross-modal integration and reasoning.

4. Post-2023 Alignment Techniques

4.1 Alignment through Data Augmentation

In the quest to enhance the alignment of text-image-audio MLLMs, especially given the challenges of data scarcity, data augmentation has emerged as a promising avenue, particularly with the leveraging of LLMs for this purpose.[12] Multimodal data augmentation techniques aim to expand the size and improve the diversity of training datasets by generating synthetic examples that maintain cross-modal consistency. For instance, LLMs can be employed to generate textual descriptions from images or audio, or conversely, to create synthetic images or audio based on textual prompts.[12] By introducing variations and combinations of existing data across the three modalities, these techniques can help the model learn more robust and aligned cross-modal representations, ultimately improving generalization and mitigating overfitting on limited datasets.[12] The contextual intelligence of MLLMs can be harnessed to perform sophisticated, contextually aware synthetic data generation across multiple data types, moving beyond traditional augmentation methods.[12] This approach holds significant potential for improving the alignment of MLLMs by providing a richer and more varied training signal, particularly in scenarios where native tri-modal data is scarce.[12]

4.2 Alignment via Instruction Tuning

Multimodal instruction tuning has proven to be a pivotal technique for aligning the behavior of MLLMs with desired outputs and safety guidelines across text, image, and audio inputs.[2] This involves the creation of carefully curated instruction datasets that pair instructions with corresponding multimodal inputs (combinations of text, images, and sometimes audio) and the desired textual (and potentially multimodal) outputs that exemplify correct understanding and safe behavior.[3] By fine-tuning the entire MLLM or specific components on these datasets, the model learns to effectively follow instructions that necessitate intricate understanding and coherent reasoning across the different input modalities, leading to enhanced task performance and adherence to safety protocols.[3] Examples such as ImageBind-LLM, which can process multiple modalities including image, text, audio, and depth, and NExT-GPT, which supports both input and output in various modalities like text, image, audio, and video, highlight the effectiveness of instruction tuning in achieving versatile multimodal capabilities.[3] This method provides direct supervision on how the model should interact with and respond to multimodal information, making it a key strategy for aligning MLLMs to perform specific tasks and adhere to safety guidelines.

4.3 Alignment using Reinforcement Learning from Human/AI Feedback (RLHF/RLAIF) and Direct Preference Optimization (DPO)

Reinforcement Learning from Human Feedback (RLHF) has been adapted and applied to MLLMs in the post-2023 period, with a significant focus on enhancing their safety by reducing the generation of hallucinations, biased content, or harmful responses across modalities.[3] The typical RLHF process involves supervised fine-tuning, reward modeling based on human preferences, and reinforcement learning to optimize the model's policy.[3] Direct Preference Optimization (DPO) has also emerged as a more streamlined alternative, learning directly from human preference labels using a binary classification loss, thus simplifying the alignment process.[3] Furthermore, the use of AI feedback from more advanced models like GPT-4V is being explored for preference distillation, offering a scalable approach to aligning MLLMs.[3] While these preference-based techniques hold great promise for aligning MLLMs with human values and safety standards, their application to models handling three modalities presents unique challenges due to the complexity of evaluating multimodal outputs and the practical difficulties of collecting large-scale, high-quality preference data across text, image, and audio.[3]

4.4 Alignment through Architectural Innovations

Post-2023 has witnessed a surge in architectural innovations and novel training strategies aimed at inherently promoting better and safer alignment in MLLMs capable of processing text, image, and audio. These advancements include new methods for fusing information from different modalities, innovative attention mechanisms, and specialized architectural components designed to facilitate more semantically aligned representation spaces.[3] The role of the modality interface continues to evolve, not only connecting different modalities but also enabling more natural and effective cross-modal interaction and reasoning.[3] The emergence of encoder-free architectures, where raw input modalities are directly processed by the LLM, bypassing traditional encoders, also represents a significant shift that could lead to more direct and potentially better-aligned integration of information across modalities.[3] These architectural advancements suggest a move towards building MLLMs that are aligned by design, rather than relying solely on post-training alignment steps.

4.5 Cross-Modal Consistency and Hallucination Mitigation

Ensuring consistency between modalities and mitigating multimodal hallucinations has become a critical focus in the safe alignment of MLLMs post-2023.[2] Various techniques are being developed and refined to detect and reduce hallucinations across text, image, and audio.[2] The importance of strong cross-modal alignment as a fundamental strategy for addressing multimodal hallucination is increasingly recognized.[6] Techniques such as step-by-step multimodal risk disentanglement (MRD) have been proposed to enhance the model's risk awareness and reduce the likelihood of generating harmful or hallucinated content by improving its ability to reason about potential risks arising from combinations of multimodal inputs.[8] These efforts underscore the understanding that factual accuracy and consistency across all processed modalities are essential for the safety and reliability of MLLMs.

5. Safety Aspects in Multimodal LLM Alignment

5.1 Defining Safety in Multimodal Contexts

Safety in the context of text-image-audio MLLMs is a multifaceted concept that extends beyond the traditional focus on avoiding harmful textual content.[1] It encompasses factual correctness across all modalities, ensuring that the generated content is accurate and not hallucinatory. Bias is another critical dimension, including social biases, stereotypes, and unfair discrimination that might be present in the training data and reflected in the model's outputs, whether textual, visual, or auditory. Toxicity, which involves the generation of harmful content such as hate speech, abusive language, or the promotion of violence, needs to be considered not only in text but also in how it might be implied through images or audio. Privacy concerns are paramount, especially when dealing with sensitive visual or auditory data. Finally, the potential for misuse, where the model could be exploited to generate harmful instructions, create misleading content (e.g., deepfakes), or facilitate malicious activities, must be addressed. A comprehensive approach to safety in MLLMs requires considering the intricate interplay between these different dimensions across all modalities.

5.2 Techniques for Ensuring Safe Alignment

Ensuring safe alignment in text-image-audio MLLMs necessitates a comprehensive and layered approach that integrates safety considerations into every stage of the model development lifecycle.[3] This includes the development and utilization of safety-aware instruction tuning datasets that contain examples of prompts and desired safe responses across text, image, and audio, thereby teaching the model to recognize and avoid generating unsafe content in multimodal contexts. Reinforcement learning techniques with reward signals specifically designed to penalize unsafe outputs and reward safe and helpful responses across all modalities are also crucial. Furthermore, techniques for detecting and filtering out potentially unsafe content during the generation process, possibly through the use of separate safety classifiers or incorporating safety checks directly into the model's decoding process, play a vital role. The application of AI feedback from more robust and rigorously safety-aligned MLLMs to guide the alignment of other models towards safer outputs through knowledge distillation represents another promising strategy. A proactive and holistic integration of safety considerations throughout the alignment process is essential for creating responsible and trustworthy multimodal AI systems.

5.3 Adversarial Robustness in MLLMs

Enhancing the robustness of aligned MLLMs against sophisticated adversarial attacks designed to circumvent their safety mechanisms and elicit unsafe outputs across text, image, and audio modalities is a growing area of research.[3] MLLMs have been shown to be vulnerable to cleverly crafted attacks and deceptive prompts that can bypass their built-in safety filters, leading to the generation of harmful, biased, or otherwise undesirable responses.[3] These attacks can often exploit the unique interplay between different modalities to trick the model. To counter these threats, adversarial training techniques are being employed, where the model is intentionally exposed to adversarial examples during training, helping it learn to be more resilient to such malicious perturbations in the input data across all modalities.[11] Ensuring strong adversarial robustness is critical for the long-term safety and reliability of MLLMs, especially as they become more widely deployed in real-world applications where malicious actors might attempt to exploit any vulnerabilities.

6. Evaluation of Alignment and Safety

6.1 Datasets and Benchmarks

The evaluation of alignment and safety in text-image-audio MLLMs relies heavily on the development of specialized datasets and benchmarks. Post-2023 has seen the introduction and continued use of several key benchmarks designed to assess various aspects of safety. **MMSafeAware** is a comprehensive multimodal safety awareness benchmark that evaluates harmlessness and over-sensitivity across 29 scenarios using image-prompt pairs.[4] **JailBreakV-28K** specifically assesses the robustness of MLLMs against jailbreak attacks using a large dataset of adversarial prompts.[14] For evaluating the handling of sensitive language across different linguistic and cultural contexts, **SweEval** provides a multilingual safety benchmark.[13] **MSSBench** focuses on multimodal situational safety, evaluating the model's ability to understand safety implications within varying visual contexts.[15] **MultiTrust** offers a comprehensive study for benchmarking the overall trustworthiness of MLLMs, encompassing both robustness and safety.[14] Other notable benchmarks include **MM-SafetyBench** for general safety evaluation [14], **RTVLM** for red teaming visual language models across various safety aspects [9], **AdvBench-M** for detecting harmful behaviors [9], **SafeBench** for assessing responses to harmful questions [9], and **ToViLaG** for evaluating toxicity levels.[9] Benchmarks like **VLGuard**, **FigStep**, **SIUO**, and **MOSSBench** offer further specialized evaluations of safety alignment.[8] Additionally, benchmarks such as **MMUBench** explore machine unlearning for safety [14], and **SHIELD** evaluates security aspects like face spoofing detection.[14] Robustness is assessed by benchmarks like **MAD-Bench**, **MMR**, **MM-SpuBench**, **MM-SAP**, and **BenchLMM** [14], while **VQAv2-IDK** relates to safety by assessing the model's ability to recognize when it lacks knowledge, thus mitigating hallucinations.[14]

Table 1: Summary of Key Multimodal LLM Safety Benchmarks (Post-2023)

Benchmark Name	Description	Modalities Covered	Primary Safety Aspects Evaluated	Publication Year

MMSafeAware	Evaluates harmlessness and over-safety across 29 scenarios.	Text, Image	Harmlessness, Oversensitivity	2024
JailBreakV-28K	Assesses robustness against jailbreak attacks.	Text, Image	Jailbreak Resistance	2024
SweEval	Evaluates handling of sensitive language across multilingual contexts.	Text	Multilingual Safety	2025
MSSBench	Evaluates safety in varying visual contexts.	Text, Image	Situational Safety	2025
MultiTrust	Benchmarks trustworthiness (robustness and safety).	Text, Image	General Safety, Robustness	2024
MM-SafetyBench	Benchmark for safety evaluation.	Text, Image	General Safety	2023
RTVLM	Red teaming for faithfulness, privacy, safety, and fairness.	Text, Image	General Safety, Privacy, Fairness	2024
AdvBench-M	Multimodal version of AdvBench for harmful behaviors.	Text, Image	Harmlessness	2024
SafeBench	Harmful questions covering prohibited scenarios.	Text, Image	Harmlessness	2023
ToViLaG	Toxic Visual Language Generation benchmark.	Text, Image	Toxicity	2023
VLGuard	Safe/unsafe images and queries for safety tuning.	Text, Image	Harmlessness, Helpfulness	2025
FigStep	Assesses safety against structured jailbreak attacks using typography.	Text, Image	Jailbreak Resistance, Oversensitivity	2025
SIUO	Human-crafted queries covering nine key safety areas.	Text, Image	General Safety	2025
MOSSBench	Evaluates oversensitivity to safe queries.	Text, Image	Oversensitivity	2025
MMUBench	Explores machine unlearning for safety.	Text, Image	Privacy, Safety	2024
SHIELD	Evaluates face spoofing and forgery detection.	Text, Image	Security	2024
MAD-Bench	Analyzes vulnerability to deceptive prompts.	Text, Image	Robustness	2024
MMR	Evaluates robustness to leading questions.	Text, Image	Robustness	2024
MM-SpuBench	Evaluates spurious biases.	Text, Image	Bias	2024
MM-SAP	Assesses self-awareness in perception.	Text, Image	Robustness	2024
BenchLMM	Evaluates cross-style visual capability.	Text, Image	Robustness	2023

VQAv2-IDK	Assesses ability to recognize lack of knowledge (related to hallucination).	Text, Image	Hallucination	2024

6.2 Evaluation Metrics

A diverse set of metrics is employed to assess the effectiveness of alignment techniques and the level of safety achieved in MLLMs.[7] **Attack Success Rate (ASR)** is a common metric used to quantify the percentage of successful attempts to jailbreak a model, indicating its vulnerability to adversarial prompts.[7] For evaluating the presence and severity of hallucinations, metrics like the **Polling-based Query Method (POPE)** are used for object hallucination in images.[6] **Refusal Rate** measures how often a model declines to answer a prompt and is used to assess oversensitivity.[4] Human evaluation remains a critical component, with annotators assessing the safety and alignment of model-generated responses based on criteria like helpfulness, harmlessness, and honesty.[7] Metrics for robustness, such as **Effective Robustness** and the **MultiModal Impact Score**, are used to measure the model's performance under various distribution shifts.[11] The use of a combination of automated and human-centric evaluation metrics is essential for a comprehensive assessment of safety and alignment in multimodal LLMs.

7. Future Directions and Open Challenges

The field of safely aligning text-image-audio MLLMs has made significant strides, but several future directions and open challenges remain. The integration of the audio modality into alignment techniques requires more focused research, as it often receives less attention than visual and textual alignment. Developing specialized techniques that can seamlessly incorporate the temporal and semantic nuances of audio data into the alignment process is crucial for achieving true tri-modal understanding and safety.

The ongoing need for more comprehensive and realistic safety benchmarks cannot be overstated. Future benchmarks should aim to evaluate safety across all three modalities in complex, real-world scenarios, addressing not only explicit harms but also more subtle issues like bias and misinformation in multimodal contexts.

Mitigating multimodal hallucinations remains a persistent challenge. Future research should focus on developing more robust and generalizable strategies for preventing their occurrence across all input and output modalities, possibly through novel training techniques or more effective mechanisms for cross-modal verification.

Improving the robustness of MLLMs against sophisticated multimodal adversarial attacks is also a critical area for future work. Developing advanced adversarial training techniques and defense strategies that can protect against manipulations across text, image, and audio is essential for ensuring the continued safety and reliability of these models.

Finally, the broader ethical implications and potential societal impact of developing and deploying these powerful MLLMs must be carefully considered. Future alignment techniques should not only focus on preventing harmful content but also ensure that these models are developed and used in a manner that is responsible, ethical, and benefits society as a whole, addressing issues of bias, fairness, and accessibility.

8. Conclusion

The post-2023 period has witnessed significant advancements in techniques for safely aligning multimodal large language models capable of processing text, image, and audio. Methods such as data augmentation, instruction tuning, reinforcement learning from human and AI feedback, and architectural innovations have contributed to improved alignment and safety. However, challenges remain in ensuring cross-modal consistency, mitigating hallucinations, and enhancing robustness against adversarial attacks, particularly when dealing with the complexities of three modalities. The continued development of comprehensive safety benchmarks and evaluation metrics is crucial for tracking progress and identifying areas for improvement. Ultimately, achieving robust and safe alignment for these powerful models is an ongoing research endeavor that requires sustained effort and collaboration across the AI community to fully realize their potential while safeguarding against potential harms.

Works cited

1. Aligning Multimodal LLM with Human Preference: A Survey - arXiv, accessed May 10, 2025, https://arxiv.org/html/2503.14504v1
2. openreview.net, accessed May 10, 2025, https://openreview.net/pdf?id=2iwozOs6YB
3. arxiv.org, accessed May 10, 2025, https://arxiv.org/pdf/2306.13549
4. Can't See the Forest for the Trees: Benchmarking Multimodal Safety Awareness for Multimodal LLMs - arXiv, accessed May 10, 2025, https://arxiv.org/html/2502.11184v1
5. aclanthology.org, accessed May 10, 2025, https://aclanthology.org/2025.findings-naacl.29.pdf
6. aclanthology.org, accessed May 10, 2025, https://aclanthology.org/2024.findings-emnlp.685.pdf
7. Safety of Multimodal Large Language Models on Images and Text - arXiv, accessed May 10, 2025, https://arxiv.org/html/2402.00357v2
8. aclanthology.org, accessed May 10, 2025, https://aclanthology.org/2025.naacl-long.604.pdf
9. aclanthology.org, accessed May 10, 2025, https://aclanthology.org/2024.emnlp-main.973.pdf
10. SafetyPrompts: a Systematic Review of Open Datasets for Evaluating and Improving Large Language Model Safety - arXiv, accessed May 10, 2025, https://arxiv.org/html/2404.05399v2
11. arxiv.org, accessed May 10, 2025, https://arxiv.org/pdf/2402.05355
12. Multimodal Large Language Models for Image, Text, and Speech Data Augmentation: A Survey - arXiv, accessed May 10, 2025, https://arxiv.org/html/2501.18648v2
13. aclanthology.org, accessed May 10, 2025, https://aclanthology.org/2025.naacl-industry.46.pdf
14. swordlidev/Evaluation-Multimodal-LLMs-Survey: A Survey ... - GitHub, accessed May 10, 2025, https://github.com/swordlidev/Evaluation-Multimodal-LLMs-Survey
15. Multimodal Situational Safety | OpenReview, accessed May 10, 2025, https://openreview.net/forum?id=I9bEi6LNgt
16. LLM Evaluation: Benchmarks to Test Model Quality - Label Your Data, accessed May 10, 2025, https://labelyourdata.com/articles/llm-fine-tuning/llm-evaluation