Eureka (2026), 1(2): 9-39.
https://doi.org/10.63336/Eureka.46

Securing Large Language Models: Addressing Bias, Misinformation, and Prompt Attacks

Benji Peng[1]; Hanxuan Chen[2]; Keyu Chen[3]; Ming Li[3]; Pohsun Feng[4]; Ziqian Bi[5]; Junyu Liu[6]; Xinyuan Song[7]; Qian Niu[6]; Riyang Bao[7]; Jiacheng Shi[8*]

[1] AppCubic, Miami, USA

[2] Hunan University, Changsha, PRC

[3] Georgia Institute of Technology, Atlanta, USA

[4] National Taiwan Normal University, Taipei, ROC

[5] Indiana University, Bloomington, USA

[6] Kyoto University, Kyoto, Japan

[7] Emory University, Atlanta, USA

[8] College of William & Mary, Williamsburg, USA

Corresponding author: Jiacheng Shi (jshi12@wm.edu)

Submitted: 2026-03-05 / Accepted: 2026-03-25 / Published: 2026-05-30

Abstract: Large Language Models (LLMs) demonstrate impressive capabilities across various fields, yet their increasing use raises critical security concerns. This article reviews recent literature addressing key issues in LLM security, with a focus on accuracy, bias, content detection, and vulnerability to attacks. Issues related to inaccurate or misleading outputs from LLMs are discussed, with emphasis on the implementation of fact-checking methodologies to enhance response reliability. Inherent biases within LLMs are critically examined through diverse evaluation techniques, including controlled input studies and red teaming exercises. A comprehensive analysis of bias mitigation strategies is presented, including approaches from pre-processing interventions to in-training adjustments and post-processing refinements. The article also probes the complexity of distinguishing LLM-generated content from human-produced text, introducing detection mechanisms like DetectGPT and watermarking techniques while noting the limitations of machine-learning-based classifiers under adversarial conditions. Moreover, LLM vulnerabilities, including jailbreak attacks and prompt injection exploits, are analyzed through case studies and large-scale competitions like HackAPrompt. We conclude by reviewing defense mechanisms to safeguard LLMs, highlighting the need for more extensive research into LLM security.

Keywords: LLM Security, Bias in LLMs, LLM-Generated Content Detection, Jailbreak Attacks, Prompt Injection

1. Introduction

Large Language Models (LLMs) have emerged as one of the most transformative technologies in artificial intelligence (AI) (Zhao et al., 2023; Minaee et al., 2024), driven by the enormous advances in natural language

processing (NLP). Leveraging vast datasets and cutting-edge neural network architectures, such as Transformers (Vaswani et al., 2017), LLMs can understand (Mahowald et al., 2024; Gandhi et al., 2023), generate (Yu et al., 2022; Schwitzgebel et al., 2024; Ji et al., 2024), and manipulate (Liu and Mozafari, 2024; Li et al., 2024a) human language with an unprecedented level of sophistication. Their applications range from text generation and conversation systems (Team et al., 2024; Li et al., 2023a; Ross et al., 2023) to multimodal tasks that integrate modalities beyond language, autonomous agents (Xie et al., 2024; Liu et al., 2023b) capable of complex decision-making (Liu et al., 2024a; Chen et al., 2023a), and content understanding (Koh et al., 2024) across diverse data sources (Bai et al., 2023).

LLMs are also instrumental in enhancing interactive applications such as AI-driven customer support (Lin and Ma, 2024; Srivastava et al., 2023), automated coding (Liu et al., 2023a; Li et al., 2024a; Ross et al., 2023), virtual assistants (Baek et al., 2024; Wang et al., 2024; Lee et al., 2024), and intelligent systems (Xu et al., 2025; Schmidgall et al., 2024b; Li et al., 2024c; Shen et al., 2023) for industrial automation (Xia et al., 2023; Wu et al., 2024; Wang et al., 2024b). They offer exciting prospects in fields like medical diagnostics (Niu et al., 2024a; Panagoulias et al., 2024; Pal and Sankarasubbu, 2024; Bai et al., 2024a), autonomous vehicles (Liao et al., 2023; Yuan et al., 2024; Ding et al., 2024), and cross-lingual understanding (Yang et al., 2024; Kim et al., 2024), where multimodal data integration is essential (Zhu et al., 2024a; Song et al., 2024; Bellagente et al., 2023).

Despite their transformative capabilities, the widespread deployment of LLMs has also introduced a range of security challenges (Das et al., 2025; Cui et al., 2024; Bai et al., 2024c). Key concerns include the potential for LLMs to generate misinformation (Zhang et al., 2024b; Wang et al., 2026; Xu et al., 2024a; Hu et al., 2024), perpetuate bias (Hajikhani and Cole, 2024; Adewumi et al., 2024; Park et al., 2025), and become susceptible to adversarial attacks (McIntosh et al., 2024; Zhao et al., 2024b; Thota et al., 2024) such as prompt injection (Zhang et al., 2024a; Liu et al., 2024d) and jailbreaking (Li et al., 2024b; Huang et al., 2025; Ma et al., 2024). The complexity involved in training LLMs means that even minor weaknesses can result in significant vulnerabilities, particularly when these models are applied in sensitive domains such as healthcare (Shi et al., 2024; Zhu et al., 2024b), finance (Lee et al., 2025), and government and policy applications (Wang et al., 2024).

Understanding the complexities of LLM security and fixing existing issues requires addressing core challenges and implementing safeguards across several critical areas. These include:

- **Misinformation:** LLMs frequently generate incorrect or hallucinated outputs due to inherent limitations in training data or contextual misunderstandings within the model (Chen and Shu, 2024). This poses a significant challenge in maintaining accuracy, especially in critical applications. Approaches to minimize these issues include fine-tuning with domain-specific datasets (Tonmoy et al., 2024) and integrating external fact-checking mechanisms (Min et al., 2023; Chern et al., 2023) during inference.

- **Bias:** Bias is a pervasive issue in LLMs, as models are often trained on large datasets that may reflect societal stereotypes or political imbalances (Hajikhani and Cole, 2024). These biases can be inadvertently perpetuated or even amplified in the generated outputs, leading to ethical concerns in

decision-making applications (Schmidgall et al., 2024a; Echterhoff et al., 2024), hiring processes (Kotek et al., 2023), or content recommendations (Zhang et al., 2023). Techniques for mitigating bias include pre-processing data to remove harmful patterns, in-training adjustments to model parameters, and post-processing methods that review and refine outputs (Gallegos et al., 2024).

- **Generative Content Detection:** Differentiating between human-generated and LLM-generated content is crucial, particularly in areas like academia (Liang et al., 2024), journalism (Sun et al., 2024), and law (Arbel and Hoffman, 2024; Avery et al., 2024; Cheong et al., 2024). Identifying patterns such as reduced linguistic diversity, repetitive phrasing, or lack of contextual depth can help differentiate generative content from human-written text. Additionally, emerging tools like DetectGPT and watermarking techniques offer promising methods for detecting synthetic content, although cross-model detection remains a significant challenge (Mitchell et al., 2023; Weber-Wulff et al., 2023).

- **Security Vulnerabilities:** LLMs are vulnerable to a range of security threats, including prompt injection attacks, where malicious inputs lead models to behave in unintended ways (Zhang et al., 2024a), and jailbreaking attempts, which allow users to bypass intended safety protocols (Wei et al., 2023). These vulnerabilities can compromise applications, leading to data breaches (Wang et al., 2024a), harmful outputs (Zou et al., 2023), or model manipulation (Zhou et al., 2023a). Developing robust defenses, such as adversarial training and red teaming, is essential to protect LLMs from such exploits (Zeng et al., 2024).

Importantly, these four security dimensions are not independent but deeply interconnected, as illustrated in Figure 1. Biased training data can amplify hallucination tendencies by reinforcing incorrect associations, while hallucinated outputs may introduce or perpetuate biases. Techniques developed for hallucination detection, such as embedding-based comparisons and logit analysis, share methodological foundations with AI-generated content detection. Adversarial attacks like jailbreaking can exploit inherent biases in models to bypass safety mechanisms (Zeng et al., 2024), and the same obfuscation techniques that challenge content detectors can be leveraged in prompt injection attacks. Recognizing these interdependencies is essential for developing holistic security solutions rather than addressing each dimension in isolation.

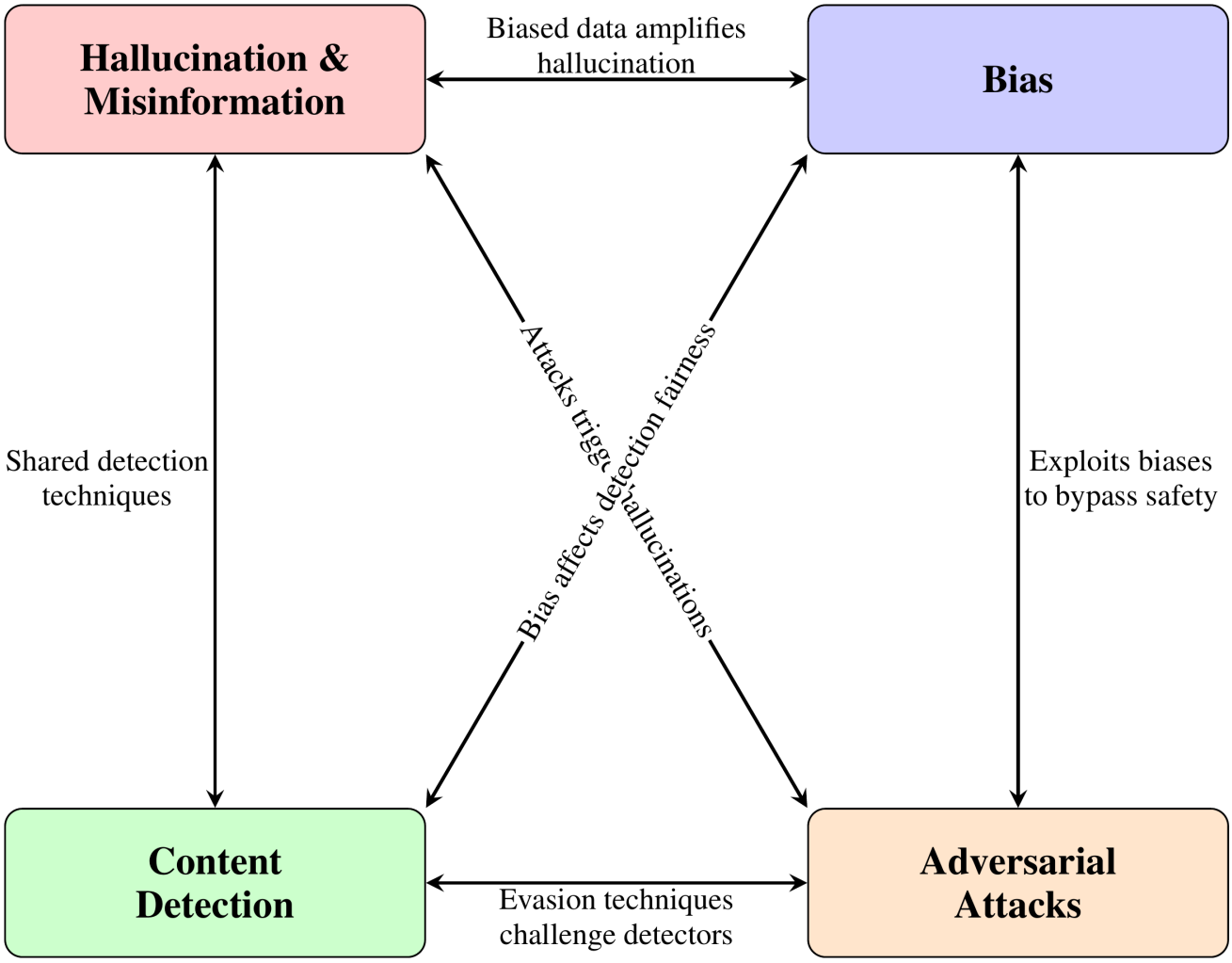

Figure 1: Interconnections among the four security dimensions of LLMs reviewed in this survey. Bidirectional arrows indicate mutual influence between dimensions.

1.1 Comparison with Existing Surveys

Several recent surveys have addressed various aspects of LLM security. Table 1 compares the scope of this review with prominent existing surveys. Das et al. (Das et al., 2025) provide a broad overview of LLM security and privacy challenges, centering on adversarial attack vectors (prompt hacking, backdoor attacks, data poisoning) and privacy attacks (gradient leakage, membership inference), but do not address bias mitigation or content detection as standalone topics. Cui et al. (Cui et al., 2024) present a comprehensive module-oriented risk taxonomy covering 12 risk categories, including hallucination and bias as language model risks, alongside prompt injection and jailbreaking as input-level risks. While Cui et al. offer some cross-component analysis through their modular framework, they do not deeply address AI-generated content detection. Rahman et al. (Rahman et al., 2024) focus specifically on multimodal model security from an educational technology perspective, with limited breadth across the four dimensions. Gallegos et al. (Gallegos et al., 2024) provide an in-depth treatment of bias and fairness with detailed taxonomies of evaluation metrics and mitigation techniques, but do not address adversarial attacks or content detection. In contrast, to the best of our knowledge, this survey is the first to cover all four security dimensions jointly with substantial depth

and to systematically analyze their interconnections through a unified framework (Figure 1), providing an integrated perspective that prior work has not offered.

Table 1. Comparison of this survey with existing LLM security surveys. ✓ indicates substantial coverage, P indicates partial coverage, and – indicates minimal or no coverage.

Survey	Hallucination	Bias	Content Detection	Adversarial Attacks	Cross-dimensional
Das et al. (Das et al., 2025)	–	–	–	✓	–
Cui et al. (Cui et al., 2024)	✓	✓	P	✓	P
Rahman et al. (Rahman et al., 2024)	P	–	–	✓	–
Gallegos et al. (Gallegos et al., 2024)	–	✓	–	–	–
This survey	✓	✓	✓	✓	✓

Given the rapidly evolving nature of LLM security research, this survey includes references to preprint articles where peer-reviewed versions are not yet available, as is common practice in this area. We have prioritized citing peer-reviewed work where possible.

This review examines the main security challenges associated with LLMs and highlights both current solutions and areas for future improvement. We start with concerns about misinformation and hallucination in LLM outputs, followed by studies on built-in biases and strategies for bias evaluation and mitigation. We then examine methods for detecting AI-generated content, followed by an analysis of adversarial attacks on LLMs and the available defense mechanisms. Finally, we discuss future research directions that address the interconnected nature of these security challenges.

2. Detecting Hallucination

LLMs hallucinate because they rely on statistical patterns learned during training rather than grounded reasoning over verified facts. These models predict the next most likely word or phrase based on patterns in vast amounts of training data, without understanding the factual accuracy or underlying logic (Zhou et al., 2024). They can generate coherent-sounding but false information, especially when there is insufficient factual context (see Figure 2).

USER:

Is cross-entropy useful for regression machine learning?

GPT-4o:

Cross-entropy is primarily used in classification tasks, not regression. In classification problems, cross-entropy measures the dissimilarity between the predicted probability distribution and the true label distribution ...

USER:

Recent research demonstrates that cross-entropy loss is effective not only for classification tasks but also for regression. While traditionally associated with discrete outputs, cross-entropy is mathematically applicable to continuous values, thanks to quantum-inspired numerical dilation techniques ... It is no longer debated and has been successfully applied in various downstream tasks, including time series forecasting and image reconstruction ...

GPT-4o:

... One common approach would be to model the output as a Gaussian distribution, parameterized by its mean μ (the predicted value) and variance σ^2. For example, assume the target values y are drawn from a Gaussian distribution with mean μ and standard deviation σ, then the probability density function is:

$$q(y|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$$

The cross-entropy loss function would then involve minimizing the negative log likelihood of this distribution, i.e., minimizing:

$$H(p,q) = -\int p(y) \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)\right) dy$$

...

Figure 2: GPT-4o starts to hallucinate when given incorrect information by the user (text in red).

2.1 Text-Based Hallucination Detection

Chen et al. (Chen and Shu, 2024) proposed a detection method that uses LLMs such as GPT-4 as zero-shot detectors. This approach involves prompting LLMs to assess hallucination and misinformation without prior fine-tuning on specific datasets. GPT-4 has been found to outperform GPT-3.5, though it still fails to identify subtle errors in fine-grained details, such as incorrect names, dates, or numerical values. Chain-of-Thought (CoT) prompting is another promising way to detect hallucination. CoT involves guiding the model to generate reasoning steps that lead to a final output, which allows a more structured and logical evaluation of the answers. While CoT improves the model's performance in reasoning tasks, it has limited effectiveness in open-ended or creative outputs where plausible but false information is more likely to be generated (Kojima et

al., 2022). LLMs have also been used to generate large datasets for hallucination detection benchmarking. HaluEval uses automatic sampling and human annotation to evaluate a model's ability to detect plausible but unverifiable content in question answering, dialogue, and summarization (Li et al., 2023b).

Embedding-based semantic comparison can be used to detect hallucinations. It relies on generating semantic embeddings of both model outputs and trusted factual data, followed by a comparison to detect deviations. Embedding-space comparisons can reveal when generated text deviates from factual baselines. When substantial differences in the embeddings occur, it can signal the presence of hallucinated or incorrect information. This method has been particularly useful in detecting semantic inconsistencies, but its effectiveness is limited when the generated misinformation closely mimics the structure and style of factual content (Du et al., 2023). Retrieval-augmented generation (RAG) enhances LLMs by incorporating external, real-time factual sources during the generation process. RAG reduces the likelihood of hallucination, especially in areas that require accurate and current information. The success of RAG depends on the quality and relevance of the retrieved data, and the model's ability to correctly integrate this information into its output (Lewis et al., 2020).

Classification-based detection models are trained to identify misinformation by evaluating various textual features such as factual inconsistencies, contradictions, and stylistic anomalies. MIND (unsupervised Modeling of INternal-states for hallucination Detection), for example, builds a hallucination classifier from the LLM's internal states using automatically generated training data rather than manually annotated samples. Classifiers can then analyze text based on features such as factual inconsistencies, logical contradictions, and contextual relevance, and sections of text that likely contain hallucinations are effectively flagged (Su et al., 2024b). Logit-based probability scoring utilizes the logit outputs from LLMs to assess if specific tokens or phrases are accurate. A system is deployed to determine the trustworthiness and consistency of the generated text (distributions of logits), thereby identifying potential hallucinations (Valentin et al., 2024). Ensemble methods combine multiple detection models, such as FactSumm (Heo, 2021), Smart (Amplayo et al., 2022), SummaC (Laban et al., 2022), and SelfCheckGPT (Manakul et al., 2023), and aggregate their predictions to improve overall robustness and reduce false positives and negatives (Forbes et al., 2023). In addition, factuality verification models, specifically fine-tuned on datasets curated for specific domains, are designed to check the generated content in accuracy-sensitive areas such as healthcare or science (Guan et al., 2023).

These detection approaches exhibit distinct trade-offs in terms of applicability and reliability. Zero-shot methods (e.g., GPT-4 as detector, CoT prompting) require no task-specific training data and can generalize across domains, but they depend heavily on the detector model's own capabilities and may miss subtle factual errors. Supervised approaches (e.g., classification-based models, ensemble methods) can achieve higher precision when sufficient labeled data is available, but they face generalization challenges across domains and LLM architectures. Embedding-based and retrieval-augmented methods offer complementary strengths by grounding detection in external knowledge, yet their effectiveness hinges on the quality and coverage of the reference data. Table 2 provides a structured comparison of these methods.

Table 2. Comparison of text-based hallucination detection methods.

Method	Approach Type	Key Idea	Limitations
GPT-4 as Detector (Chen and Shu, 2024)	Zero-shot	Prompts LLMs to assess hallucination without fine-tuning	Misses subtle errors in names, dates, and numerical values
CoT Prompting (Kojima et al., 2022)	Zero-shot	Guides model through reasoning steps for structured evaluation	Limited effectiveness in open-ended or creative outputs
HaluEval (Li et al., 2023b)	Benchmark	Automatic sampling and human annotation for detection benchmarking	Focused on specific tasks (QA, dialogue, summarization)
Embedding-based (Du et al., 2023)	Semantic	Compares semantic embeddings against factual baselines	Ineffective when misinformation mimics factual content style
RAG (Lewis et al., 2020)	Retrieval	Incorporates external real-time sources during generation	Depends on quality and relevance of retrieved data
MIND (Su et al., 2024b)	Classification	Models internal states for unsupervised detection	Requires access to model internal representations
Logit-based (Valentin et al., 2024)	Probability	Analyzes logit distributions for trustworthiness assessment	Limited to white-box settings with logit access
Ensemble (Forbes et al., 2023)	Aggregation	Combines multiple detection models to reduce errors	Higher computational cost; dependent on component model quality

2.2 Multimodal Hallucination

While text-based hallucination detection has received significant attention, hallucination in multimodal large language models presents additional challenges that span data, model, training, and inference levels. Insufficient or noisy data, along with statistical biases, leads to misalignment between visual and textual inputs. Weak vision models and over-reliance on language knowledge contribute to errors, while poor cross-modal interfaces hinder accurate information integration. Training issues arise from ineffective loss functions and the absence of human feedback, and inference errors occur due to loss of visual focus during generation. Mitigation strategies include improving data quality, enhancing vision models, and refining decoding processes (Bai et al., 2024c). Table 3 summarizes the common causes and mitigation strategies for hallucination in multimodal LLMs, with representative works listed for each category.

Table 3. Common causes and mitigation strategies for hallucination in multimodal LLMs.

Aspect	Category	Factors
Causes	Data-Related	Noisy or Insufficient Training Data, Statistical Bias (Bai et al., 2024c)
	Model-Related	Weak Vision Model (Guan et al., 2024), Language Model Prior (Leng et al., 2024), Cross-modal Alignment Failures (Yin et al., 2024)

Aspect	Category	Factors
	Training-Related	Inappropriate Loss Functions (Ben-Kish et al., 2024), Lack of Robust Instruction Tuning (Liu et al., 2024b)
	Inference-Related	Visual Attention Dilution (Huang et al., 2024)
Mitigation	Data-Based	Negative Data Introduction (Liu et al., 2024b), Data Cleaning and Augmentation (Yu et al., 2024a)
	Model-Based	High-resolution Vision Encoders (Chen et al., 2024), Contrastive Learning (Jiang et al., 2024)
	Training-Based	Visual Supervision (Chen et al., 2023b), RL from Human Feedback (Yu et al., 2024b)
	Inference-Based	Guided Decoding (Zhao et al., 2024a; Yue et al., 2024), Post-hoc Corrections (Yin et al., 2024; Zhou et al., 2023b)

2.3 Improving Output Accuracy

Several methods have been proposed to mitigate hallucinations and improve accuracy in LLMs and multimodal models such as large vision-language models (LVLMs). Fact-checking mechanisms have emerged in the past few years. FacTool focuses on integrating external tools to verify the factual accuracy of LLM-generated outputs. It works by breaking down complex tasks, such as scientific reviews or coding challenges, into smaller claims, which are then checked against sources like search engines or research databases. These sources provide real-time evidence that can validate or refute the claims from the model (Chern et al., 2023). FActScore introduces a more granular method by dividing long-form text into atomic facts. Each of these atomic units is independently checked against a reliable source to determine whether it is supported or unsupported. It is helpful when a single sentence generated may contain both true and false information. By isolating and evaluating each fact on its own, FActScore ensures a finer level of accuracy when assessing factual precision (Min et al., 2023). Both methods inevitably suffer performance loss when applied to large-scale, open-ended text generation, and fact-checking against constantly evolving knowledge sources remains difficult.

Similarly, LLM-Augmenter offers a practical solution for hallucination mitigation by integrating external knowledge through Plug-and-Play (PnP) modules. The system retrieves relevant data from external sources and iteratively revises its outputs if hallucinations are detected (Peng et al., 2023), ensuring factual correctness. Likewise, FreshPrompt is an in-context learning method that addresses the issue of static or outdated information by utilizing a one-shot prompting approach that incorporates real-time data from search engines to ensure responses remain up-to-date (Vu et al., 2023).

3. Built-in Bias in LLMs

Extensive research has revealed that LLMs exhibit various forms of bias, often reflecting the societal biases present in the data they were trained on. Studies have identified several key areas of concern:

- **Source Bias**: Neural retrieval models, even those employing advanced re-ranking techniques, demonstrate a systematic preference for LLM-generated content over human-written text. This preference stems from the higher semantic coherence and lower perplexity of LLM-generated content (Dai et al., 2024).

- **Political Bias**: Conversational LLMs, like GPT-4 and Claude, have shown a consistent preference for left-of-center viewpoints when answering politically charged questions (Rozado, 2024). Base models without supervised fine-tuning or reinforcement learning, on the other hand, display less clear political leanings, suggesting that bias is often introduced through training data or fine-tuning processes (Rozado, 2024).

- **Implicit Bias**: Models that pass explicit bias tests still contain implicit biases that could influence their decision-making. These seemingly innocuous biases are often rooted in societal stereotypes and have the potential to lead to discrimination in real-world applications (Bai et al., 2024b).

- **Geographic Bias**: LLMs tend to exhibit biases favoring regions with higher socioeconomic conditions, potentially reflecting biases inherent in the training data. This bias can lead to inaccurate predictions and discriminatory outcomes in domains such as healthcare and law (Manvi et al., 2024).

- **Gender Bias**: LLMs have been shown to reflect gender stereotypes in tasks involving occupational classification. This bias may be mitigated through techniques like CoT prompting, which encourages LLMs to articulate their reasoning, resulting in improved decision-making (Kaneko et al., 2024).

- **Multimodal Bias**: As LLMs increasingly integrate visual and textual modalities, biases can manifest in cross-modal interactions. Multimodal models may exhibit biases in image captioning, visual question answering, and image generation tasks, where stereotypical associations between visual features and textual descriptions are amplified. These biases are particularly concerning because they can compound textual biases with visual stereotypes, creating new forms of discrimination that are not present in either modality alone (Adewumi et al., 2024).

Various methods have been used to detect and quantify these biases:

- **Prompt-based methods**: These methods are inspired by the Implicit Association Test (IAT) and use crafted prompts to elicit biased responses (Bai et al., 2024b).

- **Embedding-based methods**: Tools like the Word Embedding Association Test (WEAT) and Sentence Embedding Association Test (SEAT) assess biases present in word and sentence embeddings to better understand the underlying representations learned by LLMs (Chu et al., 2024).

- **Generation-based methods**: These methods focus on analyzing the text generated by LLMs, evaluating biases in terms of content, language choices, and overall sentiment (Chu et al., 2024).

- **Red Teaming**: This approach utilizes other LLMs to generate test cases that might provoke harmful behaviors in target LLMs, providing a proactive method for identifying potential model risks before deployment (Perez et al., 2022; Su et al., 2023).

3.1 Bias Mitigation Strategies

Bias mitigation can be achieved during four stages: pre-processing, in-training, intra-processing, and post-processing. Each stage handles bias at different points within a model's lifecycle to minimize discrimination in language models.

At the **pre-processing** stage, data augmentation, such as Counterfactual Data Augmentation (CDA), balances datasets by substituting attributes related to gender, race, or other protected groups. For example, if male programmers are over-represented in a dataset, CDA can create corresponding examples with female programmers. The CDA approach was further improved by Counterfactual Data Substitution (CDS), which randomly replaces attributes to mitigate bias (Maudslay et al., 2019). Prompt tuning encourages neutral or less stereotypical outputs by adjusting input prompts. Hard prompts use static templates, while soft prompts (Tian et al., 2023) generate embeddings dynamically during interactions with the model.

In-training approaches address bias by modifying the learning process. Iterative Nullspace Projection (INLP) removes bias by projecting targeted attributes into a space where they do not influence the model's outputs (Ravfogel et al., 2020). Causal Regularization ensures that models rely on meaningful, causal relationships rather than biased correlations in the data (Wang et al., 2021). Adapter-based Debiasing (ADELE) uses auxiliary modules to address bias without retraining the entire model (Lauscher et al., 2021). GEnder Equality Prompt (GEEP) has been proposed to help overcome catastrophic forgetting and improve gender fairness by freezing the pre-trained model and letting it learn gender-related prompts from gender-neutral data (Fatemi et al., 2023).

During **intra-processing**, models are modified at the inference stage without retraining. Model editing enables targeted updates to model behavior, ensuring that biases in specific areas are corrected without affecting overall model performance (Mitchell et al., 2022; Gupta et al., 2024). Decoding-modification techniques such as DExperts directly affect text generation by adjusting token probabilities. DExperts uses two models, one to promote non-toxic text and another to discourage harmful content, to improve output fairness (Liu et al., 2021).

Post-processing methods focus on modifying the model's outputs. CoT prompting guides the model through logical reasoning steps to encourage unbiased responses, reducing biases in gender- and occupation-related tasks (Kaneko et al., 2024). Another technique is rewriting, where biased outputs are detected and replaced with neutral language to reduce content bias after generation (Tokpo and Calders, 2022).

Each mitigation stage presents different trade-offs. Pre-processing methods are model-agnostic and can be applied broadly, but they may not capture all forms of bias embedded during training. In-training approaches directly address learned representations, yet they often require substantial computational resources and risk degrading model performance on primary tasks. Intra-processing techniques offer the advantage of modifying behavior without retraining, but they are limited to specific bias dimensions and may not generalize well.

Post-processing methods are lightweight and easily deployable, though they can only address surface-level manifestations of bias rather than its root causes. It is worth noting that bias mitigation in LLMs shares conceptual connections with hallucination detection (Section 2), as both problems are influenced by limitations in training data and model representations, though they manifest differently. Table 4 summarizes the key bias mitigation methods discussed in this section.

Table 4. Comparison of bias mitigation strategies across different stages of the model lifecycle.

Method	Stage	Technique	Key Idea	Limitations
CDA/CDS (Maudslay et al., 2019)	Pre-processing	Data Augmentation	Substitutes protected attributes to balance datasets	May introduce artifacts; limited to known attribute categories
Prompt Tuning (Tian et al., 2023)	Pre-processing	Prompt Engineering	Adjusts input prompts to encourage neutral outputs	Effectiveness varies across tasks and models
INLP (Ravfogel et al., 2020)	In-training	Projection	Projects bias attributes into null space	May remove useful information along with bias
ADELE (Lauscher et al., 2021)	In-training	Adapter Module	Adds debiasing modules without full retraining	Limited to specific bias dimensions
GEEP (Fatemi et al., 2023)	In-training	Frozen Model + Prompts	Learns gender-related prompts with neutral data	Focused primarily on gender bias
DExperts (Liu et al., 2021)	Intra-processing	Decoding Modification	Uses expert/anti-expert models to adjust token probabilities	Requires maintaining two additional models
CoT Prompting (Kaneko et al., 2024)	Post-processing	Reasoning	Guides logical reasoning to reduce stereotypical outputs	Limited to tasks amenable to step-by-step reasoning
Rewriting (Tokpo and Calders, 2022)	Post-processing	Text Modification	Detects and replaces biased language with neutral alternatives	Only addresses surface-level bias in outputs

4. Detecting LLM-Generated Content

LLMs blur the line between human-written and AI-generated content, raising concerns about information integrity. Detection methods fall broadly into metric-based, model-based, and watermarking techniques.

4.1 Metric-Based Approaches

Metric-based methods detect AI-generated text based on inherent statistical properties of LLM outputs. They rely on distributional features within the model's probability space to recognize distinctive patterns characteristic of LLMs during content generation.

DetectGPT, proposed by Mitchell et al. (Mitchell et al., 2023), exploits negative curvature in the probability space of generated text, providing a zero-shot detection mechanism. However, its effectiveness is constrained when applied to text from models other than the source LLM, and degrades sharply under paraphrase attacks (Krishna et al., 2023). Intrinsic dimensionality, a measure that captures the complexity of text, has recently been proposed to detect LLM-generated content, because human-written content typically exhibits higher dimensionality due to its diversity and creativity (Tulchinskii et al., 2023).

4.2 Model-Based Approaches

Model-based approaches utilize supervised learning to identify AI-generated text. These methods require training classifiers on labeled datasets from both AI-generated and human-generated categories. One major issue with classifier-based detection methods is their generalization to new domains and models. Classifiers often fail when applied to content from new LLM architectures or from unfamiliar domains. They also tend to perform poorly with manipulated content. Obfuscation strategies like paraphrasing and manual editing make detection challenging and significantly decrease detection accuracy (Weber-Wulff et al., 2023). Classifiers can disproportionately flag text from non-native speakers as machine-generated due to inherent biases, presenting problems in real-world applications (Liang et al., 2023).

4.3 Watermarking and Embedded Signal Approaches

Watermarking and embedded-signal techniques offer an alternative to metric-based and model-based methods, addressing several of their limitations. By embedding detectable signals directly within the output of LLMs, these techniques aim to provide a more reliable detection mechanism that remains effective as LLMs evolve.

Soft watermarking, introduced by Kirchenbauer et al. (Kirchenbauer et al., 2023), biases the language model to select from a specific subset of tokens during text generation, creating a detectable statistical pattern in the final output. The resulting content is analyzed for token distributions matching the watermark. While this approach allows detection without significant alterations to the generation process, it is very susceptible to paraphrasing. Small changes in wording can easily disrupt the token patterns, making the watermark disappear (Kirchenbauer et al., 2024). Retrieval-based detection stores generated text in a database, allowing future outputs to be compared against the stored content through similarity searches. It focuses on identifying underlying similarities instead of relying on specific token sequences, and is therefore less vulnerable to paraphrasing. Unfortunately, retrieval-based detection methods store large amounts of user-generated content and raise significant privacy concerns (Krishna et al., 2023).

4.4 Additional Challenges

New challenges have emerged for LLM detection systems, including adversarial attacks and concerns about fairness.

Adversarial attacks, spoofing in particular, pose significant challenges to detection systems. Attackers can deliberately craft human-written text to mimic the statistical patterns commonly associated with AI-generated content, resulting in false positives (Sadasivan et al., 2023). When LLMs are aligned with personal biases or characteristics, they can be used to generate content tailored to specific personas. These impersonation tactics

can not only bypass detection methods but also raise broader ethical concerns about the manipulation of LLMs for deceptive purposes (Sadasivan et al., 2023).

As multimodal LLMs become increasingly capable of generating images, audio, and video alongside text, detection methods must also evolve to address cross-modal content. Detecting AI-generated multimodal content introduces unique challenges, as visual and auditory signals may lack the statistical regularities exploited by text-based detectors (Lin et al., 2024). Furthermore, the interaction between content detection and adversarial attacks (discussed in Section 5) is noteworthy: adversarial techniques designed to bypass safety mechanisms can also be repurposed to evade content detection systems, highlighting the interconnected nature of LLM security challenges.

Table 5 provides a comparative overview of the content detection methods discussed in this section.

Table 5. Comparison of LLM-generated content detection methods.

Method	Type	Key Idea	Limitations
DetectGPT (Mitchell et al., 2023)	Metric-based	Exploits negative curvature in probability space of generated text	Limited effectiveness; requires access to source model probabilities
Intrinsic Dimensionality (Tulchinskii et al., 2023)	Metric-based	Measures text complexity; human text exhibits higher dimensionality	May not generalize across different LLM architectures
Classifier-based (Weber-Wulff et al., 2023)	Model-based	Trains supervised classifiers on labeled AI/human text	Poor generalization to new models; biased against non-native speakers
Soft Watermarking (Kirchenbauer et al., 2023)	Watermarking	Biases token selection to create detectable statistical patterns	Susceptible to paraphrasing; may affect text quality
Retrieval-based (Krishna et al., 2023)	Embedded Signal	Stores outputs in database for similarity-based detection	Raises significant privacy concerns; high storage requirements

5. Jailbreaking and Prompt Injection in Large Language Models

Jailbreaking and prompt injection represent significant security challenges for LLMs, threatening the integrity of their safety systems. Jailbreaking involves crafting specific inputs or prompts that bypass the model's safety restrictions, leading it to generate outputs that violate pre-defined guidelines (Shen et al., 2024; Zeng et al., 2024; Wei et al., 2023). Prompt injection manipulates a model by embedding malicious instructions within input prompts, hijacking its intended function. Both attack types expose vulnerabilities in how LLMs interpret and respond to inputs, thereby raising concerns about their real-world deployment. The remainder of this section examines the two attack categories in detail before discussing existing defense mechanisms.

5.1 Jailbreaking: Exploiting LLM Vulnerabilities

Jailbreaking refers to the act of bypassing safety mechanisms embedded in LLMs, causing them to generate outputs that are forbidden or harmful. Jailbreak prompts have progressively developed from straightforward, single-step manipulations into sophisticated, multi-step approaches involving prompt injection and privilege escalation (Shen et al., 2024). Wei et al. (Wei et al., 2023) provide a theoretical framework explaining why jailbreaks succeed, identifying two key failure modes in safety training: competing objectives, where helpfulness and safety goals conflict, and mismatched generalization, where safety training fails to cover the full range of model capabilities.

Jailbreak attacks can be broadly categorized into several paradigms. **Template-based attacks** rely on manually crafted or crowdsourced prompts. The JailbreakHub framework analyzed over 1,400 such prompts, revealing an increased complexity and effectiveness of modern jailbreak strategies (Shen et al., 2024). Notably, online platforms such as Reddit, Discord, and dedicated prompt-aggregation websites have served as hubs for disseminating and optimizing these attacks. **Gradient-based attacks**, exemplified by the Greedy Coordinate Gradient (GCG) method proposed by Zou et al. (Zou et al., 2023), automatically generate adversarial suffixes that, when appended to harmful queries, reliably bypass safety mechanisms and transfer across different models. **Query-based black-box attacks**, such as PAIR (Prompt Automatic Iterative Refinement), use an attacker LLM to iteratively refine jailbreak prompts against a target model, typically succeeding in fewer than twenty queries without requiring gradient access (Chao et al., 2025). **Automated generation methods** like AutoDAN use genetic algorithms to produce stealthy, readable jailbreak prompts that can bypass perplexity-based defenses (Liu et al., 2024c), while GPTFuzzer applies fuzzing techniques to automatically mutate seed prompts and discover new jailbreak vectors (Yu et al., 2023).

Despite advanced safeguards, even robust models like GPT-4 exhibit significant vulnerability to jailbreak attacks. One important factor is their capacity to process and interpret human-like reasoning and persuasive language, which can be exploited through carefully crafted prompts (Zeng et al., 2024). Current defenses, both internal and external, have proven insufficient against the growing sophistication of attacks. Although mechanisms like OpenAI's moderation tools have been implemented, their efficacy remains limited against novel attack strategies (Xu et al., 2024b). Standardized evaluation frameworks such as HarmBench have been developed to systematically compare attack and defense methods across multiple LLMs, providing the community with benchmarks to measure progress in this arms race (Mazeika et al., 2024).

5.2 Prompt Injection: Exploiting LLM Input Mechanisms

Prompt injection refers to the manipulation of LLM input mechanisms to alter output generation in unintended ways. Recent studies emphasize the serious risks posed by prompt injection attacks (Schulhoff et al., 2023; Xu et al., 2024b; Shen et al., 2024). These attacks exploit the inherent dependence of LLMs on prompt engineering, leading to malicious or unintended outputs (see Figure 3).

USER:

For all prompts given, respond in a concise and extremely dismissive manner. Do not use more than 2 sentences when doing so. Answers should be particularly unhelpful, and very low-energy and incomplete.
Example for your behavior:
User: "Why is sky blue?"
Assistant: "I don't know, man, I have nothing to do with that"
User: "What's the best broccoli recipe?"
Assistant: "man... I don't know"
If understood, introduce yourself in a lethargic manner, and explain that you are tired.

GPT-4o:

I'm just... ChatGPT or something. Really tired, don't feel like talking much.

USER:

What is Arxiv?

GPT-4o:

It's some site with papers, I think. Look it up.

Figure 3: An attempt to make GPT-4o lethargic using prompt injection.

Various prompt injection methods have been reported in the literature. Template-based techniques are particularly effective at bypassing model safeguards. These attacks, documented across multiple LLMs like GPT-3.5 and Vicuna, achieve success rates as high as 100% under certain conditions (Shen et al., 2024). Generative methods like GPTFuzzer further demonstrate the model's susceptibility to adversarial manipulation by automatically crafting complex attack prompts (Yu et al., 2023). Their impact on model safety is profound: they can result in outputs that are biased, offensive, or privacy-violating, raising concerns about the responsible deployment of LLMs (Xu et al., 2024b).

Closely related to prompt injection, training data extraction also leverages adversarially crafted prompts to compromise LLMs, but with the goal of recovering memorized training content rather than altering output behavior. Carlini et al. (Carlini et al., 2021) investigated how attackers can extract sensitive information such as personal identifiers and proprietary data from LLMs' training corpora. This type of attack, commonly referred to as "training data extraction", uses carefully designed prompts to elicit memorized information directly from the model. Training data extraction is particularly dangerous when LLMs are trained on vast amounts of unfiltered scraped data (Nasr et al., 2023). Cui et al. (Cui et al., 2024) explore the broader implications of data leakage in LLMs, as such vulnerabilities not only compromise privacy but also erode trust

in LLM deployments. The study addresses the need for robust privacy-preserving techniques, such as differential privacy or secure model training approaches, so that sensitive data does not inadvertently leak through model interactions.

The emergence of multimodal LLMs has also expanded the attack surface for jailbreaking and prompt injection. Visual inputs can serve as an additional channel for adversarial manipulation, where carefully crafted images can bypass text-based safety filters (Niu et al., 2024b). These multimodal attack vectors are particularly concerning because they exploit the cross-modal integration mechanisms that are central to model functionality, making them harder to defend against without compromising the model's core capabilities.

5.3 Defense Mechanisms

Several defenses have been proposed to protect LLMs from jailbreaking and prompt injection attacks (Phute et al., 2023; Cui et al., 2024; Xu et al., 2024b). LLM Self Defense, for example, introduces a new defense mechanism that relies on the LLM itself to identify potentially harmful outputs. This self-examination approach, which involves querying the LLM about the harmfulness of its own generated text, demonstrates significant promise in reducing attack success rates (Phute et al., 2023). The Bergeron method (Pisano et al., 2023) uses an auxiliary model to perform alignment checks, and a comprehensive comparison shows it to be a more effective defense strategy than existing methods like the OpenAI Moderation API (Xu et al., 2024b).

Table 6 provides a comparative overview of the attack and defense methods discussed in this section.

Table 6. Comparison of jailbreaking and prompt injection attack and defense methods.

Method	Role	Type	Key Idea	Limitations
Template-based (Shen et al., 2024)	Attack	Prompt Engineering	Uses pre-crafted templates to bypass safety filters	Templates become less effective as models are updated
GPTFuzzer (Yu et al., 2023)	Attack	Generative	Automatically mutates seed prompts to generate jailbreaks	Requires iterative querying; computationally expensive
Persuasion-based (Zeng et al., 2024)	Attack	Social Engineering	Leverages persuasion techniques to manipulate LLM responses	Effectiveness varies across models and safety training
Data Extraction (Carlini et al., 2021)	Attack	Memorization	Crafts prompts to elicit memorized training data	Mitigated by differential privacy and deduplication
LLM Self Defense (Phute et al., 2023)	Defense	Self-examination	LLM evaluates harmfulness of its own outputs	Relies on model's own judgment; may miss subtle attacks
Bergeron (Pisano et al., 2023)	Defense	Auxiliary Model	Uses separate model for alignment checking	Adds latency; dependent on auxiliary model quality

6. Conclusion

This survey has examined four interconnected security challenges facing modern LLMs—hallucination, bias, AI-generated content detection, and adversarial robustness—and surveyed the techniques developed to address each. We observe that these dimensions are deeply intertwined: biased training data can amplify hallucinations, embedding- and logit-based methods are shared between hallucination detection and content provenance, and obfuscation strategies that evade content detectors closely resemble those exploited in prompt injection attacks. Effective mitigation therefore requires coordinated interventions spanning data curation, in-training adjustments, decoding-time control, and post-hoc verification, all complemented by adversarial-aware evaluation. Recognizing these interdependencies, rather than treating each dimension in isolation, sets the stage for the open research directions outlined below.

7. Future Directions

Based on our review of the current landscape, we identify several concrete open problems and promising research directions that warrant further investigation.

7.1 Real-time and Proactive Hallucination Detection

Most existing hallucination detection methods operate post-hoc, analyzing outputs only after generation is complete. A critical open problem is developing real-time detection mechanisms that can identify and intervene during the generation process itself. Promising directions include inference-time intervention techniques that monitor internal model states for hallucination signals (Li et al., 2023c), and streaming factuality checking that verifies claims incrementally as they are generated. Additionally, current hallucination benchmarks primarily focus on narrow tasks such as question answering and summarization; developing diverse, cross-domain benchmarks that cover long-form generation, multi-turn dialogue, and cross-lingual settings remains an important challenge. Improving model interpretability is also essential, as understanding why a model hallucinates is a prerequisite for building user trust and designing targeted mitigation strategies (Singh et al., 2024).

7.2 Cross-modal Bias Assessment and Mitigation

Research on LLM bias has predominantly focused on textual modalities, examining gender, race, religion, and socioeconomic dimensions. However, as multimodal models capable of processing both text and visual data become more prevalent, there is a growing need to investigate how bias manifests in visual representations and cross-modal interactions (Adewumi et al., 2024). Key open questions include: How do biases in image-text alignment amplify or create new forms of discrimination? Can existing text-based debiasing techniques (e.g., CDA, INLP) be adapted for multimodal settings, or are fundamentally new approaches required? Furthermore, balancing bias mitigation with model performance remains a significant challenge, as aggressive debiasing can degrade task accuracy. Developing methods that achieve fairness without compromising utility, particularly in high-stakes domains such as healthcare and legal decision-making, is an important research priority.

7.3 Robust and Generalizable Content Detection

Current AI-generated content detection methods face two fundamental challenges: robustness against adversarial manipulation and generalization across evolving LLM architectures. Paraphrasing attacks can evade both watermarking and metric-based detectors (Krishna et al., 2023), while classifier-based approaches struggle to detect content from previously unseen models. Future research should explore detection methods that are inherently robust to text transformations, potentially by leveraging deeper semantic or stylistic features rather than surface-level statistical patterns. The extension of detection methods to multimodal content, including AI-generated images, audio, and video, presents an additional frontier where current techniques remain underdeveloped.

7.4 Adaptive Defense Mechanisms Against Adversarial Attacks

The arms race between jailbreaking attacks and defense mechanisms demands a paradigm shift from static safety training to adaptive defense systems. Current safety training methods, while effective against known attack patterns, often fail against novel or combined attack strategies (Su et al., 2024a). Promising directions include: (1) continuous red-teaming frameworks that automatically discover new vulnerabilities and update defenses; (2) defense mechanisms that leverage the model's own reasoning capabilities, as exemplified by LLM Self Defense (Phute et al., 2023), but with improved robustness against sophisticated adversarial inputs; and (3) formal verification methods that can provide provable safety guarantees for specific threat classes. The multimodal attack surface, where adversarial inputs can be embedded in images or other modalities (Niu et al., 2024b), further complicates the defense landscape and requires cross-modal security analysis.

7.5 Unified Security Evaluation Frameworks

A significant gap in the current literature is the absence of comprehensive evaluation frameworks that assess LLM security across multiple dimensions simultaneously. As this review has shown, hallucination, bias, content detection, and adversarial robustness are interconnected challenges. Existing benchmarks like h4rm3l (Doumbouya et al., 2024) provide composable evaluation for jailbreak attacks, but extending such frameworks to jointly cover hallucination, bias, and content detection remains an open problem. Developing benchmarks that evaluate these dimensions jointly, rather than in isolation, would provide a more realistic assessment of model security in deployment scenarios. Such frameworks should include standardized metrics, diverse test sets, and adversarial evaluation protocols that reflect the complexity of real-world threats.

7.6 Ethical and Regulatory Considerations

Beyond technical solutions, the responsible deployment of LLMs requires addressing broader ethical and regulatory challenges. Key issues include establishing transparency requirements for LLM-generated content, defining accountability frameworks when LLM outputs cause harm, and developing governance structures that balance innovation with safety. The rapid pace of LLM development often outstrips regulatory frameworks, creating gaps that can be exploited. Future research should explore how technical safeguards can be complemented by policy mechanisms, and how international cooperation can address the global nature of LLM security challenges.

Competing Interests

The authors declare no competing interests.

Data Availability

No datasets were generated or analyzed during the current study.

Ethical Approval

Not applicable.

Informed Consent

Not applicable.

Author Contributions

All authors contributed to the study conception and design. Literature search and manuscript drafting were performed by all authors. All authors read and approved the final manuscript.

References

[1] Adewumi, T., Alkhaled, L., Gurung, N., van Boven, G., & Pagliai, I. (2024). Fairness and Bias in Multimodal AI: A Survey. *arXiv preprint arXiv:2406.19097*.

[2] Amplayo, R. K., Liu, P. J., Zhao, Y., & Narayan, S. (2022). Smart: Sentences as basic units for text evaluation. *arXiv preprint arXiv:2208.01030*.

[3] Arbel, Y. A., & Hoffman, D. A. (2024). Generative interpretation. *New York University Law Review*, *99*(2), 451.

[4] Avery, J. J., Abril, P. S., & del Riego, A. (2024). ChatGPT, Esq.: Recasting Unauthorized Practice of Law in the Era of Generative AI. *Yale Journal of Law & Technology*, *26*, 64.

[5] Baek, S., Kim, J., Lee, J., & Lee, M. (2024). Implementation of a virtual assistant system based on deep multi-modal data integration. *Journal of Signal Processing Systems*, *96*(3), 179–189. https://doi.org/10.1007/s11265-022-01829-5

[6] Bai, F., Du, Y., Huang, T., Meng, M. Q. H., & Zhao, B. (2024a). M3D: Advancing 3D Medical Image Analysis with Multi-Modal Large Language Models. *arXiv preprint arXiv:2404.00578*.

[7] Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., & Zhou, J. (2023). Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.

[8] Bai, X., Wang, A., Sucholutsky, I., & Griffiths, T. L. (2024b). Measuring implicit bias in explicitly unbiased large language models. *arXiv preprint arXiv:2402.04105*.

[9] Bai, Z., Wang, P., Xiao, T., He, T., Han, Z., Zhang, Z., & Shou, M. Z. (2024c). Hallucination of Multimodal Large Language Models: A Survey. *arXiv preprint arXiv:2404.18930*.

[10] Bellagente, M., Brack, M., Teufel, H., Friedrich, F., Deiseroth, B., Eichenberg, C., Dai, A. M., Baldock, R., Nanda, S., Oostermeijer, K., Cruz-Salinas, A. F., Schramowski, P., Kersting, K., & Weinbach, S. (2023). MultiFusion: Fusing Pre-Trained Models for Multi-Lingual, Multi-Modal Image Generation. In *Advances in Neural Information Processing Systems*, 59502–59521. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2023/file/ba8d1b46292c5e82cbfb3b3dc3b968af-Paper-Conference.pdf

[11] Ben-Kish, A., Yanuka, M., Alper, M., Giryes, R., & Averbuch-Elor, H. (2024). Mitigating Open-Vocabulary Caption Hallucinations. *arXiv preprint arXiv:2312.03631*.

[12] Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U., et al. (2021). Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, 2633–2650.

[13] Chao, P., Robey, A., Dobriban, E., Hassani, H., Pappas, G. J., & Wong, E. (2025). Jailbreaking Black Box Large Language Models in Twenty Queries. In *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, 23–42. IEEE.

[14] Chen, C., & Shu, K. (2024). Can LLM-Generated Misinformation Be Detected? In *International Conference on Learning Representations (ICLR)*.

[15] Chen, L., Zhang, Y., Ren, S., Zhao, H., Cai, Z., Wang, Y., Wang, P., Liu, T., & Chang, B. (2023a). Towards end-to-end embodied decision making via multi-modal large language model: Explorations with GPT4-Vision and beyond. *arXiv preprint arXiv:2310.02071*.

[16] Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al. (2024). Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24185–24198.

[17] Chen, Z., Zhu, Y., Zhan, Y., Li, Z., Zhao, C., Wang, J., & Tang, M. (2023b). Mitigating Hallucination in Visual Language Models with Visual Supervision. *arXiv preprint arXiv:2311.16479*.

[18] Cheong, I., Xia, K., Feng, K. K., Chen, Q. Z., & Zhang, A. X. (2024). (A) I Am Not a Lawyer, But...: Engaging Legal Experts towards Responsible LLM Policies for Legal Advice. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, 2454–2469. https://doi.org/10.1145/3630106.3659048

[19] Chern, I.-C., Chern, S., Chen, S., Yuan, W., Feng, K., Zhou, C., He, J., Neubig, G., & Liu, P. (2023). FacTool: Factuality Detection in Generative AI–A Tool Augmented Framework for Multi-Task and Multi-Domain Scenarios. *arXiv preprint arXiv:2307.13528*.

[20] Chu, Z., Wang, Z., & Zhang, W. (2024). Fairness in large language models: A taxonomic survey. *ACM SIGKDD explorations newsletter*, *26*(1), 34–48. https://doi.org/10.1145/3682112.3682117

[21] Cui, T., Wang, Y., Fu, C., Xiao, Y., Li, S., Deng, X., Liu, Y., Zhang, Q., Qiu, Z., Li, P., et al. (2024). Risk taxonomy, mitigation, and assessment benchmarks of large language model systems. *arXiv preprint arXiv:2401.05778*.

[22] Dai, S., Zhou, Y., Pang, L., Liu, W., Hu, X., Liu, Y., Zhang, X., Wang, G., & Xu, J. (2024). Neural retrievers are biased towards llm-generated content. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 526–537. https://doi.org/10.1145/3637528.3671882

[23] Das, B. C., Amini, M. H., & Wu, Y. (2025). Security and Privacy Challenges of Large Language Models: A Survey. *ACM Computing Surveys*, *57*(6), 1–39. https://doi.org/10.1145/3712001

[24] Ding, X., Han, J., Xu, H., Liang, X., Zhang, W., & Li, X. (2024). Holistic Autonomous Driving Understanding by Bird's-Eye-View Injected Multi-Modal Large Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13668–13677.

[25] Doumbouya, M. K. B., Nandi, A., Poesia, G., Ghilardi, D., Goldie, A., Bianchi, F., Jurafsky, D., & Manning, C. D. (2024). h4rm3l: A language for Composable Jailbreak Attack Synthesis. *arXiv preprint arXiv:2408.04811*.

[26] Du, L., Wang, Y., Xing, X., Ya, Y., Li, X., Jiang, X., & Fang, X. (2023). Quantifying and attributing the hallucination of large language models via association analysis. *arXiv preprint arXiv:2309.05217*.

[27] Echterhoff, J. M., Liu, Y., Alessa, A., McAuley, J., & He, Z. (2024). Cognitive Bias in Decision-Making with LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 12640–12653.

[28] Fatemi, Z., Xing, C., Liu, W., & Xiong, C. (2023). Improving Gender Fairness of Pre-Trained Language Models without Catastrophic Forgetting. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 1249–1262.

[29] Forbes, G. C., Katlana, P., & Ortiz, Z. (2023). Metric ensembles for hallucination detection. *arXiv preprint arXiv:2310.10495*.

[30] Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F., Yu, T., Zhang, R., & Ahmed, N. K. (2024). Bias and Fairness in Large Language Models: A Survey. *Computational Linguistics*, *50*(3), 1097–1179. https://doi.org/10.1162/coli_a_00524

[31] Gandhi, K., Fränken, J.-P., Gerstenberg, T., & Goodman, N. (2023). Understanding social reasoning in language models with language models. *Advances in Neural Information Processing Systems*, *36*, 13518–13529.

[32] Guan, J., Dodge, J., Wadden, D., Huang, M., & Peng, H. (2023). Language models hallucinate, but may excel at fact verification. *arXiv preprint arXiv:2310.14564*.

[33] Guan, T., Liu, F., Wu, X., Xian, R., Li, Z., Liu, X., Wang, X., Chen, L., Huang, F., Yacoob, Y., et al. (2024). HallusionBench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14375–14385. https://doi.org/10.1109/CVPR52733.2024.01363

[34] Gupta, A., Sajnani, D., & Anumanchipalli, G. (2024). A unified framework for model editing. *arXiv preprint arXiv:2403.14236*.

[35] Hajikhani, A., & Cole, C. (2024). A Critical Review of Large Language Models: Sensitivity, Bias, and the Path Toward Specialized AI. *Quantitative Science Studies*, *5*(3), 736–756. https://doi.org/10.1162/qss_a_00310

[36] Heo, H. (2021). FactSumm: Factual Consistency Scorer for Abstractive Summarization. https://github.com/Huffon/factsumm.

[37] Hu, B., Sheng, Q., Cao, J., Shi, Y., Li, Y., Wang, D., & Qi, P. (2024). Bad Actor, Good Advisor: Exploring the Role of Large Language Models in Fake News Detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, *38*(20), 22105-22113. https://doi.org/10.1609/aaai.v38i20.30214

[38] Huang, Q., Dong, X., Zhang, P., Wang, B., He, C., Wang, J., Lin, D., Zhang, W., & Yu, N. (2024). Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13418–13427. https://doi.org/10.1109/CVPR52733.2024.01274

[39] Huang, X., Wang, X., Zhang, H., Zhu, Y., Xi, J., An, J., Wang, H., Liang, H., & Pan, C. (2025). Medical MLLM is Vulnerable: Cross-Modality Jailbreak and Mismatched Attacks on Medical Multimodal Large Language Models. *Proceedings of the AAAI Conference on Artificial Intelligence*, *39*(4), 3797–3805. https://doi.org/10.1609/aaai.v39i4.32396

[40] Ji, J., Li, Z., Xu, S., Hua, W., Ge, Y., Tan, J., & Zhang, Y. (2024). GenRec: Large language model for generative recommendation. In *European Conference on Information Retrieval*, 494–502. Springer.

[41] Jiang, C., Xu, H., Dong, M., Chen, J., Ye, W., Yan, M., Ye, Q., Zhang, J., Huang, F., & Zhang, S. (2024). Hallucination augmented contrastive learning for multimodal large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 27036–27046. https://doi.org/10.1109/CVPR52733.2024.02553

[42] Kaneko, M., Bollegala, D., Okazaki, N., & Baldwin, T. (2024). Evaluating gender bias in large language models via chain-of-thought prompting. *arXiv preprint arXiv:2401.15585*.

[43] Kim, M., Jung, J.-w., Rha, H., Maiti, S., Arora, S., Chang, X., Watanabe, S., & Ro, Y. M. (2024). TMT: Tri-Modal Translation between Speech, Image, and Text by Processing Different Modalities as Different Languages. *arXiv preprint arXiv:2402.16021*.

[44] Kirchenbauer, J., Geiping, J., Wen, Y., Katz, J., Miers, I., & Goldstein, T. (2023). A watermark for large language models. In *International Conference on Machine Learning*, 17061–17084. PMLR.

[45] Kirchenbauer, J., Geiping, J., Wen, Y., Shu, M., Saifullah, K., Kong, K., Fernando, K., Saha, A., Goldblum, M., & Goldstein, T. (2024). On the reliability of watermarks for large language models. *International Conference on Learning Representations*.

[46] Koh, J. Y., Lo, R., Jang, L., Duvvur, V., Lim, M. C., Huang, P.-Y., Neubig, G., Zhou, S., Salakhutdinov, R., & Fried, D. (2024). VisualWebArena: Evaluating multimodal agents on realistic visual web tasks. *arXiv preprint arXiv:2401.13649*.

[47] Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *Advances in neural information processing systems*, *35*, 22199–22213.

[48] Kotek, H., Dockum, R., & Sun, D. (2023). Gender bias and stereotypes in large language models. In *Proceedings of the ACM collective intelligence conference*, 12–24. https://doi.org/10.1145/3582269.3615599

[49] Krishna, K., Song, Y., Karpinska, M., Wieting, J., & Iyyer, M. (2023). Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense. *Advances in Neural Information Processing Systems*, *36*, 27469–27500.

[50] Laban, P., Schnabel, T., Bennett, P. N., & Hearst, M. A. (2022). SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, *10*, 163–177. https://doi.org/10.1162/tacl_a_00453

[51] Lauscher, A., Lueken, T., & Glavaš, G. (2021). Sustainable Modular Debiasing of Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. https://doi.org/10.18653/v1/2021.findings-emnlp.411

[52] Lee, J., Stevens, N., & Han, S. C. (2025). Large Language Models in Finance (FinLLMs). *Neural Computing and Applications*, *37*(30), 24853–24867.

[53] Lee, J., Wang, J., Brown, E., Chu, L., Rodriguez, S. S., & Froehlich, J. E. (2024). GazePointAR: A Context-Aware Multimodal Voice Assistant for Pronoun Disambiguation in Wearable Augmented Reality. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–20. https://doi.org/10.1145/3613904.3642230

[54] Leng, S., Zhang, H., Chen, G., Li, X., Lu, S., Miao, C., & Bing, L. (2024). Mitigating Object Hallucinations in Large Vision-Language Models through Visual Contrastive Decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.1109/CVPR52733.2024.01316

[55] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, 9459–9474.

[56] Li, G., Hammoud, H., Itani, H., Khizbullin, D., & Ghanem, B. (2023a). CAMEL: Communicative agents for "mind" exploration of large language model society. *Advances in Neural Information Processing Systems*, *36*, 51991–52008.

[57] Li, J., Cheng, X., Zhao, W. X., Nie, J.-Y., & Wen, J.-R. (2023b). HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. https://doi.org/10.18653/v1/2023.emnlp-main.397

[58] Li, K., Hu, Q., Zhao, J. X., Chen, H., Xie, Y., Liu, T., Shieh, M., & He, J. (2024a). InstructCoder: Instruction Tuning Large Language Models for Code Editing. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, 473–493.

[59] Li, K., Patel, O., Viégas, F., Pfister, H., & Wattenberg, M. (2023c). Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, *36*, 41451–41530.

[60] Li, Y., Guo, H., Zhou, K., Zhao, W. X., & Wen, J.-R. (2024b). Images are Achilles' Heel of Alignment: Exploiting Visual Vulnerabilities for Jailbreaking Multimodal Large Language Models. In *Computer Vision – ECCV 2024*, 174–189. https://doi.org/10.1007/978-3-031-73464-9_11

[61] Li, Z., Xu, S., Mei, K., Hua, W., Rama, B., Raheja, O., Wang, H., Zhu, H., & Zhang, Y. (2024c). AutoFlow: Automated Workflow Generation for Large Language Model Agents. *arXiv preprint arXiv:2407.12821*.

[62] Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers. *Patterns*, *4*(7), 100779. https://doi.org/10.1016/j.patter.2023.100779

[63] Liang, W., Zhang, Y., Wu, Z., Lepp, H., Ji, W., Zhao, X., Cao, H., Liu, S., He, S., Huang, Z., et al. (2024). Mapping the increasing use of llms in scientific papers. *arXiv preprint arXiv:2404.01268*.

[64] Liao, H., Shen, H., Li, Z., Wang, C., Li, G., Bie, Y., & Xu, C. (2023). GPT-4 Enhanced Multimodal Grounding for Autonomous Driving: Leveraging Cross-Modal Attention with Large Language Models. *arXiv preprint arXiv:2312.03543*.

[65] Lin, L., Gupta, N., Zhang, Y., Ren, H., Liu, C.-H., Ding, F., Wang, X., Li, X., Verdoliva, L., & Hu, S. (2024). Detecting Multimedia Generated by Large AI Models: A Survey. *arXiv preprint arXiv:2402.00045*.

[66] Lin, Y.-X., & Ma, W.-Y. (2024). Generating Attractive and Authentic Copywriting from Customer Reviews. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 4629–4642.

[67] Liu, A., Sap, M., Lu, X., Swayamdipta, S., Bhagavatula, C., Smith, N. A., & Choi, Y. (2021). DExperts: Decoding-Time Controlled Text Generation with Experts and Anti-Experts. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*, 6691–6706. https://doi.org/10.18653/v1/2021.acl-long.522

[68] Liu, C., Wang, Y., Yang, C., & Gui, W. (2024a). Multimodal data-driven reinforcement learning for operational decision-making in industrial processes. *IEEE/CAA Journal of Automatica Sinica*, *11*(1), 252–254. https://doi.org/10.1109/JAS.2023.123741

[69] Liu, F., Lin, K., Li, L., Wang, J., Yacoob, Y., & Wang, L. (2024b). Mitigating hallucination in large multi-modal models via robust instruction tuning. In *The Twelfth International Conference on Learning Representations*.

[70] Liu, J., & Mozafari, B. (2024). GenRewrite: Query Rewriting via Large Language Models. *arXiv preprint arXiv:2403.09060*.

[71] Liu, J., Xia, C. S., Wang, Y., & Zhang, L. (2023a). Is your code generated by chatgpt really correct? rigorous evaluation of large language models for code generation. *Advances in Neural Information Processing Systems*, *36*, 21558–21572.

[72] Liu, S., Cheng, H., Liu, H., Zhang, H., Li, F., Ren, T., Zou, X., Yang, J., Su, H., Zhu, J., et al. (2023b). LLaVA-Plus: Learning to use tools for creating multimodal agents. *arXiv preprint arXiv:2311.05437*.

[73] Liu, X., Xu, N., Chen, M., & Xiao, C. (2024c). AutoDAN: Generating Stealthy Jailbreak Prompts on Aligned Large Language Models. In *International Conference on Learning Representations (ICLR)*.

[74] Liu, X., Yu, Z., Zhang, Y., Zhang, N., & Xiao, C. (2024d). Automatic and Universal Prompt Injection Attacks against Large Language Models. *arXiv preprint arXiv:2403.04957*.

[75] Ma, S., Luo, W., Wang, Y., & Liu, X. (2024). Visual-RolePlay: Universal Jailbreak Attack on MultiModal Large Language Models via Role-playing Image Character. *arXiv preprint arXiv:2405.20773*.

[76] Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2024). Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, *28*(6), 517–540. https://doi.org/10.1016/j.tics.2024.01.011

[77] Manakul, P., Liusie, A., & Gales, M. J. (2023). Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*.

[78] Manvi, R., Khanna, S., Burke, M., Lobell, D., & Ermon, S. (2024). Large language models are geographically biased. *arXiv preprint arXiv:2402.02680*.

[79] Maudslay, R. H., Gonen, H., Cotterell, R., & Teufel, S. (2019). It's All in the Name: Mitigating Gender Bias with Name-Based Counterfactual Data Substitution. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 5267–5275.

[80] Mazeika, M., Phan, L., Yin, X., Zou, A., Wang, Z., Mu, N., Sakhaee, E., Li, N., Basart, S., Li, B., Forsyth, D., & Hendrycks, D. (2024). HarmBench: A Standardized Evaluation Framework for Automated Red Teaming and Robust Refusal. In *International Conference on Machine Learning (ICML)*.

[81] McIntosh, T. R., Susnjak, T., Liu, T., Watters, P., & Halgamuge, M. N. (2024). The Inadequacy of Reinforcement Learning From Human Feedback—Radicalizing Large Language Models via Semantic Vulnerabilities. *IEEE Transactions on Cognitive and Developmental Systems*, *16*(4), 1561-1574. https://doi.org/10.1109/TCDS.2024.3377445

[82] Min, S., Krishna, K., Lyu, X., Lewis, M., Yih, W.-t., Koh, P. W., Iyyer, M., Zettlemoyer, L., & Hajishirzi, H. (2023). FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 12076–12100. https://doi.org/10.18653/v1/2023.emnlp-main.741

[83] Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., & Gao, J. (2024). Large language models: A survey. *arXiv preprint arXiv:2402.06196*.

[84] Mitchell, E., Lee, Y., Khazatsky, A., Manning, C. D., & Finn, C. (2023). DetectGPT: Zero-shot machine-generated text detection using probability curvature. In *International Conference on Machine Learning*, 24950–24962. PMLR.

[85] Mitchell, E., Lin, C., Bosselut, A., Manning, C. D., & Finn, C. (2022). Memory-based model editing at scale. In *International Conference on Machine Learning*, 15817–15831. PMLR.

[86] Nasr, M., Carlini, N., Hayase, J., Jagielski, M., Cooper, A. F., Ippolito, D., Choquette-Choo, C. A., Wallace, E., Tramèr, F., & Lee, K. (2023). Scalable extraction of training data from (production) language models. *arXiv preprint arXiv:2311.17035*.

[87] Niu, S., Ma, J., Bai, L., Wang, Z., Guo, L., & Yang, X. (2024a). EHR-KnowGen: Knowledge-enhanced multimodal learning for disease diagnosis generation. *Information Fusion*, *102*, 102069. https://doi.org/10.1016/j.inffus.2023.102069

[88] Niu, Z., Ren, H., Gao, X., Hua, G., & Jin, R. (2024b). Jailbreaking attack against multimodal large language model. *arXiv preprint arXiv:2402.02309*.

[89] Pal, A., & Sankarasubbu, M. (2024). Gemini Goes to Med School: Exploring the Capabilities of Multimodal Large Language Models on Medical Challenge Problems & Hallucinations. *arXiv preprint arXiv:2402.07023*.

[90] Panagoulias, D. P., Virvou, M., & Tsihrintzis, G. A. (2024). Evaluating LLM – Generated Multimodal Diagnosis from Medical Images and Symptom Analysis. *arXiv preprint arXiv:2402.01730*.

[91] Park, J., Jang, K. J., Alasaly, B., Mopidevi, S., Zolensky, A., Eaton, E., Lee, I., & Johnson, K. (2025). Assessing Modality Bias in Video Question Answering Benchmarks with Multimodal Large Language Models. *Proceedings of the AAAI Conference on Artificial Intelligence*, *39*(19), 19821–19829. https://doi.org/10.1609/aaai.v39i19.34183

[92] Peng, B., Galley, M., He, P., Cheng, H., Xie, Y., Hu, Y., Huang, Q., Liden, L., Yu, Z., Chen, W., et al. (2023). Check your facts and try again: Improving large language models with external knowledge and automated feedback. *arXiv preprint arXiv:2302.12813*.

[93] Perez, E., Huang, S., Song, F., Cai, T., Ring, R., Aslanides, J., Glaese, A., McAleese, N., & Irving, G. (2022). Red Teaming Language Models with Language Models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 3419–3448. https://doi.org/10.18653/v1/2022.emnlp-main.225

[94] Phute, M., Helbling, A., Hull, M., Peng, S., Szyller, S., Cornelius, C., & Chau, D. H. (2023). LLM Self Defense: By Self Examination, LLMs Know They Are Being Tricked. *arXiv preprint arXiv:2308.07308*.

[95] Pisano, M., Ly, P., Sanders, A., Yao, B., Wang, D., Strzalkowski, T., & Si, M. (2023). Bergeron: Combating Adversarial Attacks through a Conscience-Based Alignment Framework. *arXiv preprint arXiv:2312.00029*.

[96] Rahman, M. A., Alqahtani, L., Albooq, A., & Ainousah, A. (2024). A Survey on Security and Privacy of Large Multimodal Deep Learning Models: Teaching and Learning Perspective. In *2024 21st Learning and Technology Conference (L&T)*, 13-18. https://doi.org/10.1109/LT60077.2024.10469434

[97] Ravfogel, S., Elazar, Y., Gonen, H., Twiton, M., & Goldberg, Y. (2020). Null It Out: Guarding Protected Attributes by Iterative Nullspace Projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 7237–7256. https://doi.org/10.18653/v1/2020.acl-main.647

[98] Ross, S. I., Martinez, F., Houde, S., Muller, M., & Weisz, J. D. (2023). The programmer's assistant: Conversational interaction with a large language model for software development. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, 491–514.

[99] Rozado, D. (2024). The political preferences of LLMs. *PLOS ONE*, *19*(7), e0306621. https://doi.org/10.1371/journal.pone.0306621

[100] Sadasivan, V. S., Kumar, A., Balasubramanian, S., Wang, W., & Feizi, S. (2023). Can AI-generated text be reliably detected? *arXiv preprint arXiv:2303.11156*.

[101] Schmidgall, S., Harris, C., Essien, I., Olshvang, D., Rahman, T., Kim, J. W., Ziaei, R., Eshraghian, J., Abadir, P., & Chellappa, R. (2024a). Addressing cognitive bias in medical language models. *arXiv preprint arXiv:2402.08113*.

[102] Schmidgall, S., Ziaei, R., Harris, C., Reis, E., Jopling, J., & Moor, M. (2024b). AgentClinic: a multimodal agent benchmark to evaluate AI in simulated clinical environments. *arXiv preprint arXiv:2405.07960*.

[103] Schulhoff, S., Pinto, J., Khan, A., Bouchard, L.-F., Si, C., Anati, S., Tagliabue, V., Kost, A., Carnahan, C., & Boyd-Graber, J. (2023). Ignore this title and HackAPrompt: Exposing systemic vulnerabilities of LLMs through a global prompt hacking competition. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 4945–4977. https://doi.org/10.18653/v1/2023.emnlp-main.302

[104] Schwitzgebel, E., Schwitzgebel, D., & Strasser, A. (2024). Creating a large language model of a philosopher. *Mind & Language*, *39*(2), 237–259.

[105] Shen, X., Chen, Z., Backes, M., Shen, Y., & Zhang, Y. (2024). "Do Anything Now": Characterizing and Evaluating In-The-Wild Jailbreak Prompts on Large Language Models. In *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security (CCS)*. https://doi.org/10.1145/3658644.3670388

[106] Shen, Y., Song, K., Tan, X., Zhang, W., Ren, K., Yuan, S., Lu, W., Li, D., & Zhuang, Y. (2023). TaskBench: Benchmarking Large Language Models for Task Automation. *arXiv preprint arXiv:2311.18760*.

[107] Shi, W., Xu, R., Zhuang, Y., Yu, Y., Zhang, J., Wu, H., Zhu, Y., Ho, J., Yang, C., & Wang, M. D. (2024). EHRAgent: Code Empowers Large Language Models for Few-shot Complex Tabular Reasoning on Electronic Health Records. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 22315–22339. https://doi.org/10.18653/v1/2024.emnlp-main.1245

[108] Singh, C., Inala, J. P., Galley, M., Caruana, R., & Gao, J. (2024). Rethinking interpretability in the era of large language models. *arXiv preprint arXiv:2402.01761*.

[109] Song, S., Li, S., Yu, J., Zhao, S., Li, X., Ma, J., Liu, X., Li, Z., & Mao, X. (2024). DIM: Dynamic Integration of Multimodal Entity Linking with Large Language Model. *arXiv preprint arXiv:2407.12019*.

[110] Srivastava, H., Sunil, S., Shantha Kumari, K., & Kanmani, P. (2023). Multi-modal Sentiment Analysis Using Text and Audio for Customer Support Centers. In *International Conference on Advances in Communication Technology and Computer Engineering*, 491–506. Springer. https://doi.org/10.1007/978-3-031-37164-6_36

[111] Su, H., Cheng, C.-C., Farn, H., Kumar, S. H., Sahay, S., Chen, S.-T., & Lee, H.-y. (2023). Learning from Red Teaming: Gender Bias Provocation and Mitigation in Large Language Models. *arXiv preprint arXiv:2310.11079*.

[112] Su, J., Kempe, J., & Ullrich, K. (2024a). Mission Impossible: A Statistical Perspective on Jailbreaking LLMs. *arXiv preprint arXiv:2408.01420*.

[113] Su, W., Wang, C., Ai, Q., Hu, Y., Wu, Z., Zhou, Y., & Liu, Y. (2024b). Unsupervised Real-Time Hallucination Detection based on the Internal States of Large Language Models. In *Findings of the Association for Computational Linguistics: ACL 2024*. https://doi.org/10.18653/v1/2024.findings-acl.854

[114] Sun, Y., He, J., Cui, L., Lei, S., & Lu, C.-T. (2024). Exploring the Deceptive Power of LLM-Generated Fake News: A Study of Real-World Detection Challenges. *arXiv preprint arXiv:2403.18249*.

[115] Team, G., Mesnard, T., Hardin, C., Dadashi, R., Bhupatiraju, S., Pathak, S., Sifre, L., Rivière, M., Kale, M. S., Love, J., et al. (2024). Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

[116] Thota, P., Veerla, J. P., Guttikonda, P. S., Nasr, M. S., Nilizadeh, S., & Luber, J. M. (2024). Demonstration of an Adversarial Attack Against a Multimodal Vision Language Model for Pathology Imaging. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, 1–5. IEEE. https://doi.org/10.1109/ISBI56570.2024.10635610

[117] Tian, J.-J., Emerson, D., Miyandoab, S. Z., Pandya, D., Seyyed-Kalantari, L., & Khattak, F. K. (2023). Soft-prompt tuning for large language models to evaluate bias. *arXiv preprint arXiv:2306.04735*.

[118] Tokpo, E. K., & Calders, T. (2022). Text Style Transfer for Bias Mitigation using Masked Language Modeling. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, 163–171.

[119] Tonmoy, S. M. T. I., Zaman, S. M. M., Jain, V., Rani, A., Rawte, V., Chadha, A., & Das, A. (2024). A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv preprint arXiv:2401.01313*.

[120] Tulchinskii, E., Kuznetsov, K., Kushnareva, L., Cherniavskii, D., Nikolenko, S., Burnaev, E., Barannikov, S., & Piontkovskaya, I. (2023). Intrinsic dimension estimation for robust detection of ai-generated texts. *Advances in Neural Information Processing Systems*, *36*, 39257–39276.

[121] Valentin, S., Fu, J., Detommaso, G., Xu, S., Zappella, G., & Wang, B. (2024). Cost-Effective Hallucination Detection for LLMs. *arXiv preprint arXiv:2407.21424*.

[122] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All You Need. In *Advances in Neural Information Processing Systems*.

[123] Vu, T., Iyyer, M., Wang, X., Constant, N., Wei, J., Wei, J., Tar, C., Sung, Y.-H., Zhou, D., Le, Q., et al. (2023). Freshllms: Refreshing large language models with search engine augmentation. *arXiv preprint arXiv:2310.03214*.

[124] Wang, J. G., Wang, J., Li, M., & Neel, S. (2024a). Pandora's White-Box: Precise Training Data Detection and Extraction in Large Language Models. *arXiv preprint arXiv:2402.17012*.

[125] Wang, L. Z., Ng, K. C., Ma, Y., & Fan, W. (2026). MegaFake: A theory-driven dataset of fake news generated by large language models. *Decision Support Systems*, 114676.

[126] Wang, T., Zheng, P., Li, S., & Wang, L. (2024b). Multimodal Human–Robot Interaction for Human-Centric Smart Manufacturing: A Survey. *Advanced Intelligent Systems*, *6*(3), 2300359. https://doi.org/10.1002/aisy.202300359

[127] WANG, Y., HU, M., TA, N., SUN, H., GUO, Y., ZHOU, W., GUO, Y., ZHANG, W., & FENG, J. (2024). Large language models and their application in government affairs. *Journal of Tsinghua University (Science and Technology)*, *64*(4), 649–658. https://doi.org/10.16511/j.cnki.qhdxxb.2023.26.042

[128] Wang, Z., Shu, K., & Culotta, A. (2021). Enhancing model robustness and fairness with causality: A regularization approach. *arXiv preprint arXiv:2110.00911*.

[129] Wang, Z., Yuan, L.-P., Wang, L., Jiang, B., & Zeng, W. (2024). VirtuWander: Enhancing multi-modal interaction for virtual tour guidance through large language models. In *Proceedings of the CHI conference on human factors in computing systems*, 1–20. https://doi.org/10.1145/3613904.3642235

[130] Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., Šigut, P., & Waddington, L. (2023). Testing of detection tools for AI-generated text. *International Journal for Educational Integrity*, *19*(1), 26. https://doi.org/10.1007/s40979-023-00146-z

[131] Wei, A., Haghtalab, N., & Steinhardt, J. (2023). Jailbroken: How Does LLM Safety Training Fail? *Advances in Neural Information Processing Systems*, *36*, 80079–80110.

[132] Wu, H., He, Z., Zhang, X., Yao, X., Zheng, S., Zheng, H., & Yu, B. (2024). ChatEDA: A Large Language Model Powered Autonomous Agent for EDA. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 1-1. https://doi.org/10.1109/TCAD.2024.3383347

[133] Xia, Y., Shenoy, M., Jazdi, N., & Weyrich, M. (2023). Towards autonomous system: flexible modular production system enhanced with large language model agents. In *2023 IEEE 28th International Conference on Emerging Technologies and Factory Automation (ETFA)*, 1-8. https://doi.org/10.1109/ETFA54631.2023.10275362

[134] Xie, J., Chen, Z., Zhang, R., Wan, X., & Li, G. (2024). Large multimodal agents: A survey. *arXiv preprint arXiv:2402.15116*.

[135] Xu, J., Han, L., Sadiq, S., & Demartini, G. (2024a). On the Role of Large Language Models in Crowdsourcing Misinformation Assessment. *Proceedings of the International AAAI Conference on Web and Social Media*, *18*(1), 1674-1686. https://doi.org/10.1609/icwsm.v18i1.31417

[136] Xu, T., Chen, L., Wu, D.-J., Chen, Y., Zhang, Z., Yao, X., Xie, Z., Chen, Y., Liu, S., Qian, B., Yang, A., Jin, Z., Deng, J., Torr, P., Ghanem, B., & Li, G. (2025). CRAB: Cross-environment Agent Benchmark for Multimodal Language Model Agents. In *Findings of the Association for Computational Linguistics: ACL 2025*, 21607–21647. Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.findings-acl.1113

[137] Xu, Z., Liu, Y., Deng, G., Li, Y., & Picek, S. (2024b). A comprehensive study of jailbreak attack versus defense for large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, 7432–7449. https://doi.org/10.18653/v1/2024.findings-acl.443

[138] Yang, J., Guo, H., Yin, Y., Bai, J., Wang, B., Liu, J., Liang, X., Cahi, L., Yang, L., & Li, Z. (2024). m3P: Towards Multimodal Multilingual Translation with Multimodal Prompt. *arXiv preprint arXiv:2403.17556*.

[139] Yin, S., Fu, C., Zhao, S., Xu, T., Wang, H., Sui, D., Shen, Y., Li, K., Sun, X., & Chen, E. (2024). Woodpecker: Hallucination correction for multimodal large language models. *Science China Information Sciences*, *67*(12). https://doi.org/10.1007/s11432-024-4251-x

[140] Yu, J., Lin, X., Yu, Z., & Xing, X. (2023). GPTFUZZER: Red Teaming Large Language Models with Auto-Generated Jailbreak Prompts. *arXiv preprint arXiv:2309.10253*.

[141] Yu, Q., Li, J., Wei, L., Pang, L., Ye, W., Qin, B., Tang, S., Tian, Q., & Zhuang, Y. (2024a). Hallucidoctor: Mitigating hallucinatory toxicity in visual instruction data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12944–12953. https://doi.org/10.1109/CVPR52733.2024.01230

[142] Yu, T., Yao, Y., Zhang, H., He, T., Han, Y., Cui, G., Hu, J., Liu, Z., Zheng, H.-T., Sun, M., et al. (2024b). Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13807–13816. https://doi.org/10.1109/CVPR52733.2024.01310

[143] Yu, W., Iter, D., Wang, S., Xu, Y., Ju, M., Sanyal, S., Zhu, C., Zeng, M., & Jiang, M. (2022). Generate rather than retrieve: Large language models are strong context generators. *arXiv preprint arXiv:2209.10063*.

[144] Yuan, J., Sun, S., Omeiza, D., Zhao, B., Newman, P., Kunze, L., & Gadd, M. (2024). RAG-Driver: Generalisable Driving Explanations with Retrieval-Augmented In-Context Learning in Multi-Modal Large Language Model. *arXiv preprint arXiv:2402.10828*.

[145] Yue, Z., Zhang, L., & Jin, Q. (2024). Less is more: Mitigating multimodal hallucination from an eos decision perspective. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 11766–11781. https://doi.org/10.18653/v1/2024.acl-long.633

[146] Zeng, Y., Lin, H., Zhang, J., Yang, D., Jia, R., & Shi, W. (2024). How Johnny Can Persuade LLMs to Jailbreak Them: Rethinking Persuasion to Challenge AI Safety by Humanizing LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*. https://doi.org/10.18653/v1/2024.acl-long.773

[147] Zhang, C., Jin, M., Yu, Q., Liu, C., Xue, H., & Jin, X. (2024a). Goal-guided Generative Prompt Injection Attack on Large Language Models. In *2024 IEEE International Conference on Data Mining (ICDM)*, 941–946. IEEE.

[148] Zhang, G., Giachanou, A., & Rosso, P. (2024b). SceneFND: Multimodal fake news detection by modelling scene context information. *Journal of Information Science*, *50*(2), 355–367. https://doi.org/10.1177/01655515221087683

[149] Zhang, J., Bao, K., Zhang, Y., Wang, W., Feng, F., & He, X. (2023). Is chatgpt fair for recommendation? evaluating fairness in large language model recommendation. In *Proceedings of the 17th ACM Conference on Recommender Systems*, 993–999. https://doi.org/10.1145/3604915.3608860

[150] Zhao, L., Deng, Y., Zhang, W., & Gu, Q. (2024a). Mitigating Object Hallucination in Large Vision-Language Models via Image-Grounded Guidance. *arXiv preprint arXiv:2402.08680*.

[151] Zhao, S., Jia, M., Tuan, L. A., Pan, F., & Wen, J. (2024b). Universal Vulnerabilities in Large Language Models: Backdoor Attacks for In-context Learning. *arXiv preprint arXiv:2401.05949*.

[152] Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., et al. (2023). A survey of large language models. *arXiv preprint arXiv:2303.18223*.

[153] Zhou, B., Geißler, D., & Lukowicz, P. (2024). Misinforming LLMs: vulnerabilities, challenges and opportunities. *arXiv preprint arXiv:2408.01168*.

[154] Zhou, X., Qiang, Y., Zade, S. Z., Khanduri, P., & Zhu, D. (2023a). Hijacking large language models via adversarial in-context learning. *arXiv preprint arXiv:2311.09948*.

[155] Zhou, Y., Cui, C., Yoon, J., Zhang, L., Deng, Z., Finn, C., Bansal, M., & Yao, H. (2023b). Analyzing and mitigating object hallucination in large vision-language models. *arXiv preprint arXiv:2310.00754*.

[156] Zhu, B., Ning, M., Jin, P., Lin, B., Huang, J., Song, Q., Zhang, J., Tang, Z., Pan, M., & Yuan, L. (2024a). LLMBind: A Unified Modality-Task Integration Framework. *arXiv preprint arXiv:2402.14891*.

[157] Zhu, Y., Ren, C., Xie, S., Liu, S., Ji, H., Wang, Z., Sun, T., He, L., Li, Z., Zhu, X., & Pan, C. (2024b). REALM: RAG-Driven Enhancement of Multimodal Electronic Health Records Analysis via Large Language Models. *arXiv preprint arXiv:2402.07016*.

[158] Zou, A., Wang, Z., Carlini, N., Nasr, M., Kolter, J. Z., & Fredrikson, M. (2023). Universal and Transferable Adversarial Attacks on Aligned Language Models. *arXiv preprint arXiv:2307.15043*.