Mitigating the Adverse Effects of Large Language Models: A Comprehensive Analysis of Solutions and Techniques

A report requested by Grace and produced by Google's Deep research on 21 Mar 2025

1. Introduction

The rapid advancement and increasing deployment of Generative AI Large Language Models (GAI LLMs) have ushered in a new era of technological capabilities across diverse applications [1]. However, this progress is accompanied by a growing awareness of potential adverse effects, including the generation of biased or harmful content, the propagation of misinformation and factual inaccuracies, the risk of privacy violations through unintended data disclosure, and vulnerabilities to security threats like prompt injection [3]. To harness the benefits of LLMs responsibly, the development and implementation of robust mitigation strategies are paramount [4]. This report provides a comprehensive analysis of a wide range of solutions and techniques aimed at mitigating these adverse effects, extending beyond commonly discussed tools to explore a broader spectrum of approaches. The analysis encompasses content moderation tools, privacy-preserving techniques, deployment strategies, and various other methods identified through research.

The current landscape of LLM risk mitigation is characterized by a multi-faceted approach, driven by the understanding that no single solution can completely eliminate all potential harms. The effectiveness of these techniques often involves navigating trade-offs between factors such as safety, performance efficiency, and usability, necessitating careful consideration and selection based on the specific context and application of the LLM [9]. This report aims to provide a detailed understanding of these trade-offs and the nuances associated with different mitigation strategies.

2. Content Moderation and Safety Tools

2.1 ShieldGemma

Developed by Google, ShieldGemma is a suite of instruction-tuned content classification models built upon the Gemma 2 architecture [11]. These models are specifically designed to evaluate text for potential violations across four key harm categories: sexually explicit content, dangerous content, hate speech, and harassment [11]. ShieldGemma operates in two primary modes: prompt-only mode, which analyzes user input to determine if it violates safety policies or attempts to elicit harmful responses from the AI, and prompt-response mode, which analyzes both the user input and the model's generated response to assess policy violations [15].

ShieldGemma can function by providing a probability score between 0 and 1, indicating the likelihood of a content violation, or by generating a textual "Yes/No" output along with the reasons for the classification [15]. The scoring prediction mode offers a finer level of control over content filtering thresholds [15]. Benchmark evaluations against datasets like OpenAI Moderation API and ToxicChat have demonstrated ShieldGemma's competitive performance in classifying harmful content [11]. However, the effectiveness of ShieldGemma can be sensitive to the specific phrasing and clarity of the user-provided safety policy descriptions, and the model might exhibit unpredictable behavior when dealing with ambiguous or nuanced language [12]. Furthermore, the current version of ShieldGemma is primarily available for the English language [11]. This specialized approach to content moderation, focusing on distinct harm categories, suggests a design aimed at achieving higher accuracy within those specific domains, particularly with smaller model sizes [15]. The performance across different parameter sizes (2B, 9B, and 27B) indicates a potential trade-off between model size and accuracy, where smaller, specialized models can be particularly effective [11].

2.2 Llama Guard (including Llama Guard 3)

Llama Guard, developed by Meta, is another significant content safety classification model designed to safeguard human-AI conversations by examining both user inputs and LLM responses [17]. The latest iteration, Llama Guard 3, expands upon the capabilities of its predecessors by incorporating new hazard categories such as defamation, election-related misinformation, and code interpreter abuse [17]. This model demonstrates multilingual support, enabling content safety classification in eight languages, including English, French, German, Hindi, Italian, Portuguese, Spanish, and Thai [17].

Furthermore, Llama Guard 3 Vision extends these safety measures to multimodal content, processing both text and images to identify harmful prompts and responses, particularly in image reasoning use cases [21]. Llama Guard 3 classifies content based on the MLCommons standardized hazards taxonomy, which includes a comprehensive range of categories such as violent crimes, non-violent crimes, sex-related crimes, child sexual exploitation, privacy violations, and hate speech [17]. Performance evaluations have shown high F1 scores and low false positive rates for Llama Guard 3, often surpassing the performance of models like GPT-4 in specific areas like multilingual capabilities and tool use detection [17]. Despite its robust performance, Llama Guard 3 may exhibit an increased tendency to refuse benign prompts (resulting in false positives), and its performance can vary depending on the specific language and application context [19]. It is important to note that Llama Guard 3 Vision is specifically designed for multimodal safety and is not intended as a standalone image safety or text-only classification tool [24]. The advancements in Llama Guard towards handling multimodal content and detecting code interpreter abuse highlight an increasing emphasis on addressing a broader spectrum of potential risks within more complex LLM deployments.

2.3 Google's Model Armor

Google Cloud's Model Armor is a fully managed service designed to enhance the security and safety of AI applications by acting as a screening layer for LLM prompts and responses [26]. This service is model-independent and cloud-agnostic, providing flexibility for organizations utilizing various AI models and deployment environments [26]. Model Armor offers centralized management of security and safety policies through a public REST API, enabling seamless integration into existing AI workflows. It also incorporates role-based access control and regional endpoints to ensure low latency [26]. A key feature of Model Armor is its integration with Google Cloud's Security Command Center, providing a unified view of security threats and policy violations [26].

Model Armor provides a suite of safety and security filters, including content safety checks for harmful content (dangerous instructions, harassment, hate speech, sexually explicit material), detection of prompt injection and jailbreak attempts, data loss prevention (DLP) capabilities using Sensitive Data Protection to prevent the leakage of PII and intellectual property, identification of malicious URLs, and the ability to scan text within PDF documents [26]. The service operates by inspecting both incoming user prompts and outgoing model-generated responses, allowing for actions such as sanitization, blocking, or allowing content based on predefined security policies [26]. The benefits of using Model Armor include enhanced AI safety and security, centralized monitoring and control over LLM applications, and mitigation of various security and safety risks [26]. However, Model Armor has certain limitations, such as a token limit for the prompt injection detection filter (512 tokens), limited regional availability, and the requirement for additional configuration for comprehensive detection of sensitive information like email addresses and passwords [28]. Furthermore, it currently lacks direct integration with the Vertex AI Gen AI stack [28]. Model Armor's comprehensive and centralized approach to AI security offers a broader range of protections compared to tools primarily focused on content moderation or prompt injection detection, making it a versatile option for organizations seeking to secure their AI deployments across diverse infrastructures.

3. Privacy-Preserving Techniques

3.1 Federated Learning

Federated learning represents a paradigm shift in training machine learning models, including LLMs, by enabling collaborative learning across decentralized data sources while ensuring that sensitive data remains localized on individual devices or within organizational silos [9]. In this approach, instead of transferring raw data to a central server, individual clients (devices or servers) train a local model on their respective datasets, and only model updates, such as gradients or weights, are shared with a central server for aggregation into a global model [9]. This process significantly enhances privacy by minimizing the exposure of sensitive data and reducing the risks associated with data breaches [32]. Additionally, federated learning can leverage diverse datasets distributed across various environments, potentially leading to the development of more robust and generalizable LLMs [34].

Despite its privacy advantages, implementing federated learning for LLMs presents several technical challenges. The frequent exchange of model updates between distributed devices can lead to significant communication overhead, straining network bandwidth and infrastructure [32]. Model drift, where local models diverge due to analyzing different data patterns, can negatively impact the overall performance of the global model [32]. The computational limitations of edge devices can also pose difficulties in training large-scale language models locally [32]. Furthermore, the update phase of federated learning systems is susceptible to security vulnerabilities, including data poisoning attacks and the introduction of backdoors by malicious actors [32]. Scaling federated learning to a large number of devices can also be resource-intensive [32], and evaluating and debugging federated models can be challenging due to the distributed nature of the data [32]. Privacy attacks targeting both the model updates and the final trained model are also a concern [36].

To mitigate these challenges, various techniques can be employed. Model sparsification and quantization can reduce the size of update data, thereby lowering communication overhead [32]. Differential privacy can be applied to add noise to model updates, providing an additional layer of privacy protection [9]. Secure aggregation techniques, such as homomorphic encryption, can protect the confidentiality of model updates during the aggregation process [32]. Federated learning for LLMs is a currently available solution, with ongoing research exploring its application in various privacy-sensitive domains like healthcare and finance [31]. The development of hybrid training models that combine centralized data preprocessing with decentralized client-side updates represents a promising direction for optimizing both performance and privacy [32].

3.2 Using LLMs Locally (On-Device or in Private Infrastructure)

Deploying LLMs locally, whether on individual devices or within a private organizational infrastructure, offers a direct approach to mitigating data privacy concerns and reducing exposure to external security vulnerabilities [37]. By keeping sensitive data within the confines of an organization's own servers or a user's device, this deployment strategy provides enhanced data privacy and security, minimizing the risks associated with external data transfers and potential breaches [38]. Local deployment also results in reduced latency, as data processing occurs locally without the need to communicate with remote servers [38]. For organizations operating in highly regulated industries, such as healthcare and finance, local LLMs can simplify compliance with stringent data protection regulations like GDPR and HIPAA [38]. Furthermore, this approach provides organizations with greater control over their data handling practices and the ability to customize models to meet specific business needs [38]. It also significantly reduces the potential for data interception by external systems and prevents unauthorized third parties from using the data in ways that violate internal policies [39].

However, deploying LLMs locally also presents several challenges. It requires a significant upfront investment in powerful computing resources, including high-end GPUs, substantial RAM, and robust storage systems [38]. The deployment and ongoing maintenance of local LLMs demand specialized technical expertise, requiring skilled professionals who understand both AI technologies and infrastructure management [38]. Unlike cloud-based solutions, local setups have inherent limitations in terms of scalability, making it difficult to dynamically adjust computational resources based on demand [38]. Organizations might also face challenges in accessing the latest pre-trained large language models and ensuring continuous updates, as these often originate from cloud-based providers [38]. Moreover, with local deployment, the responsibility for implementing and managing comprehensive security measures falls entirely on the organization [39]. Despite these challenges, running LLMs locally offers significant security benefits for use cases involving highly sensitive data, enabling applications such as secure internal tools, personalized customer service without external data sharing, and real-time decision support based on proprietary information [41]. The choice between local and cloud-based LLM deployment often hinges on a careful evaluation of data sensitivity requirements, computational resource availability, budget constraints, specific use case needs, and regulatory compliance obligations [38].

3.3 Other Privacy-Enhancing Technologies

Beyond federated learning and local deployment, several other privacy-enhancing technologies are relevant to mitigating privacy risks associated with LLMs. **Differential privacy (DP)** is a robust mathematical framework that adds a carefully calibrated amount of statistical noise to data during training or inference. This process provides provable privacy guarantees by limiting the amount of information that can be inferred about any individual record in the dataset [4]. While DP is effective in safeguarding individual privacy, it can introduce a trade-off with the accuracy and utility of the resulting model [9]. Notably, Google has implemented differential privacy in its BERT language models.

Homomorphic encryption (HE) is another promising technique that allows computations to be performed directly on encrypted data without the need for decryption [9], S_T1]. This ensures that sensitive information remains protected throughout the processing pipeline. However, current implementations of homomorphic encryption can be computationally intensive and slow, particularly for the complex operations involved in training large language models [47], S_T1]. **Secure multi-party computation (SMPC)** is a cryptographic protocol that enables multiple parties to collaboratively compute a function on their private inputs while keeping those inputs secret from each other [30], S_T1]. Similar to homomorphic encryption, SMPC can introduce performance overhead, making it challenging to apply to large-scale LLM training scenarios [47]. **Targeted Catastrophic Forgetting (TCF)** is a more specialized technique that involves iteratively fine-tuning a model to selectively "forget" specific sensitive data points from its training, thereby reducing the risk of privacy leakage during subsequent use [49]. While these various privacy-enhancing technologies hold significant potential for protecting sensitive data in the context of LLMs, their practical application often involves navigating trade-offs between privacy protection, model performance, and computational efficiency, necessitating ongoing research and development to optimize their effectiveness and scalability.

4. Additional Mitigation Solutions and Techniques (from GitHub Link)

The GAI-is-going-well repository on GitHub serves as a valuable collection of resources documenting the unexpected outcomes, security risks, and privacy concerns associated with the use of LLMs and Generative AI [50]. The repository is organized into categories including adverse effects, regulatory information, research articles on vulnerabilities, and, importantly, **mitigations & tooling** [50]. Analysis of the information related to this repository reveals a diverse set of mitigation techniques proposed by the community and researchers:

- **Model Updating and Patching:** Regularly updating the foundational LLM to incorporate the latest security patches and address known vulnerabilities is a crucial first line of defense [51].
- **Input Sanitization and Filtering:** Implementing robust filters to analyze and sanitize user inputs can help prevent prompt injection and jailbreaking attacks by stripping out potentially malicious or harmful requests [51].
- **Side-Channel Attack Mitigation:** Techniques such as monitoring packet headers for unusual token lengths can be employed to detect and mitigate side-channel attacks that might attempt to extract information about the model's inner workings [51].
- **Comprehensive Testing and Validation:** Establishing thorough testing processes for LLMs is essential to identify potential weaknesses, biases, and unexpected behaviors before deployment [52]. This includes validating both the input and output of the model to ensure integrity and safety [52].
- **Security and Privacy Guardrails:** Implementing comprehensive security and privacy guardrails throughout the LLMOps pipeline is critical for protecting sensitive data and preventing unauthorized access or misuse [52].
- **Secure Platform Utilization:** For applications handling sensitive information, utilizing paid GAI tools that offer data isolation and enhanced security features can provide an added layer of protection [53].
- **Policy Review and Compliance:** Thoroughly reviewing the terms of service, privacy policies, and intellectual property implications of GAI tools is important for understanding data handling practices and potential legal risks [53].
- **Output Verification and Scrutiny:** Rigorous scrutiny of all LLM-generated outputs, especially in critical domains like law and academia, is necessary to identify and correct hallucinations, misinformation, and inaccuracies [53].
- **User Education and Policy Communication:** Clearly communicating GAI usage policies to users and educating them about responsible use and the rationale behind these policies can foster a safer environment [55].
- **Assessment Design in Education:** In educational settings, employing strategies like shorter, more frequent assessments, scaffolding assignments, integrating class-specific material, requiring reflective elements, and utilizing oral presentations or in-class writing can reduce the incentive and effectiveness of

using GAI for academic dishonesty [55].

- **Content Provenance:** Implementing mechanisms to establish the origin of AI-generated content, such as watermarking or digital signatures, can aid in tracking the spread of misinformation and addressing intellectual property issues [56].
- **Access Control:** Implementing robust access controls to protect training data and the models themselves from unauthorized access is fundamental to security [56].
- **Security Assessments and Red-Teaming:** Regularly conducting security assessments and engaging in red-teaming exercises can proactively identify and address potential vulnerabilities in LLM systems [56].
- **Secure Development Lifecycle:** Incorporating security considerations into every stage of the LLM development lifecycle, including the supply chain of components, is crucial for building resilient systems [56].
- **Monitoring and Incident Response:** Establishing systems for continuous monitoring and having well-defined incident response plans are essential for detecting and mitigating security incidents and attacks [56].
- **Environmental Impact Mitigation:** Techniques like model compression, energy-efficient training practices, and carbon offsetting can help reduce the environmental footprint of large LLMs [56].
- **Bias Mitigation Strategies:** Employing a range of techniques aimed at reducing bias in both the training data and the model's outputs is crucial for fairness and equity [56].
- **Local LLM Security Measures:** For locally deployed LLMs, implementing anomaly detection, data redaction, and access monitoring and control can enhance security [39].
- **Shift-Left Security:** Integrating security considerations early in the development process, including data minimization and continuous testing, is a proactive approach [46].
- **Reinforcement Learning from Human Feedback (RLHF):** Utilizing human feedback to align model outputs with desired values and reduce harmful content is a powerful mitigation strategy [46].
- **Adversarial Training:** Training LLMs on examples designed to trick them can improve their robustness against malicious inputs [46].
- **Prompt Engineering for Safety:** Carefully crafting prompts to guide the model towards safer and less biased responses, as well as using techniques like post-generation self-diagnosis, can be effective without requiring additional training [39].
- **AI Guardrails and Response Filtering:** Implementing automated checks and

filters to identify and prevent the generation or exposure of sensitive or harmful content [37].

- **Ethical Framework Integration:** Incorporating established AI ethics guidelines into the development lifecycle, including fairness, accountability, and transparency, is essential for responsible innovation [58].
- **Fairness-Aware Algorithms and Evaluation:** Using algorithms that penalize biased outcomes and employing evaluation metrics that go beyond simple accuracy to assess fairness across different social groups [58].
- **Stakeholder Engagement:** Proactively involving diverse stakeholders, especially those from marginalized communities, in the development and evaluation process to ensure a wider range of perspectives are considered [58].
- **Transparency Enhancements:** Utilizing tools like attention visualization and token-level confidence scores can provide insights into the model's decision-making process, improving transparency [59]. Similarly, citation and source tracking can help users verify the accuracy of generated content [59].
- **Ethical Oversight and Auditing:** Establishing ethical review boards and implementing internal audit procedures can ensure that ethical principles are integrated and followed throughout the LLM lifecycle [60].

This extensive list highlights the breadth and depth of the ongoing efforts within the research and development community to address the various challenges posed by GAI LLMs. The GAI-is-going-well repository serves as a valuable hub for tracking these issues and the proposed solutions.

5. Detailed Description of Mitigation Strategies

(This section will now provide a detailed description for a selection of the key mitigation strategies identified, drawing upon the information from the previous sections.)

5.1 ShieldGemma: A suite of content moderation models from Google, built on Gemma 2, designed to classify text into categories like sexually explicit, dangerous, hate, and harassment. It works by analyzing text inputs or input-output pairs against defined safety policies, providing a score or a "Yes/No" classification with reasons. Benefits include specialized models for specific harm types and competitive benchmark performance. Limitations include sensitivity to policy phrasing and limited language support [11].

5.2 Llama Guard 3: A content safety classification model by Meta, fine-tuned for both prompts and responses, with multilingual support and expanded hazard categories including code interpreter abuse and election-related misinformation. Llama Guard 3 Vision extends this to multimodal content. It operates by classifying content as safe or unsafe based on the MLCommons taxonomy. Benefits include high accuracy and low false positive rates. Limitations include potential for increased refusals of benign prompts and performance variations across languages [17].

5.3 Google's Model Armor: A fully managed Google Cloud service that screens LLM prompts and responses for security and safety risks like prompt injection, data leaks, harmful content, and malicious URLs. It works by filtering both input and output based on defined security policies. Benefits include centralized management, cloud-agnostic support, and comprehensive security features. Limitations include token limits for prompt injection detection and limited regional availability [26].

5.4 Federated Learning: A decentralized training technique where models are trained on distributed data without sharing the raw data itself. Only model updates are exchanged and aggregated. It enhances privacy and security by keeping data localized. Benefits include improved privacy and the ability to train on diverse datasets. Limitations include communication overhead, model drift, computational constraints, and security vulnerabilities during the update phase [9].

5.5 Using LLMs Locally: Deploying LLMs on-device or within private infrastructure keeps data within the organization's control, mitigating privacy concerns and security risks. It offers enhanced data privacy, reduced latency, and easier regulatory compliance. Challenges include high upfront hardware costs, technical complexity, and limited scalability [37].

5.6 Prompt Engineering for Safety: Carefully designing prompts to guide the LLM towards desired outputs and avoid harmful or biased responses. Techniques include using clear instructions, specifying roles, and employing post-generation self-diagnosis. Benefits include simplicity and ease of implementation without additional training. Limitations include reliance on the model's inherent capabilities and potential for bypass with sophisticated prompts [39].

5.7 Bias Mitigation Techniques: A broad category encompassing various methods to identify and reduce bias in LLMs. These include using diverse training data, employing bias detection tools, implementing counterfactual data augmentation, fine-tuning with fairness constraints, and post-processing model outputs. Benefits include fairer and more equitable model behavior. Limitations often involve trade-offs with model accuracy and the complexity of identifying and addressing all forms of bias [4].

6. Categorization of Mitigation Strategies

The mitigation strategies discussed can be broadly categorized as follows:

- **Content Moderation Tools:** ShieldGemma, Llama Guard (including Llama Guard 3), Google's Model Armor.
- **Privacy-Preserving Techniques:** Federated Learning, Using LLMs Locally, Differential Privacy, Homomorphic Encryption, Secure Multi-Party Computation, Targeted Catastrophic Forgetting.
- **Deployment Strategies:** Using LLMs Locally, Secure Cloud Deployments (leveraging tools like Model Armor).
- **Prompt-Based Techniques:** Prompt Engineering for Safety, Post-Generation Self-Diagnosis.
- **Bias Mitigation Techniques:** Diverse training data, bias detection tools, counterfactual data augmentation, fairness-aware algorithms, post-processing adjustments, and many others listed in Section 4.
- **Transparency and Accountability Measures:** Attention visualization tools, token-level confidence scores, citation and source tracking, detailed documentation, ethical review boards, internal audit procedures.
- **Security Safeguards:** Input sanitization and filtering, monitoring packet headers, access control, security assessments and red-teaming, secure development practices, monitoring and incident response, AI Guardrails, response filtering, secure model serving.
- **Educational and Procedural Measures:** User education, clear GAI policies, assessment design in education, thorough review of GAI tool policies, output verification.

7. Availability Status

- **Currently Available:** ShieldGemma, Llama Guard (including Llama Guard 3), Google's Model Armor, using LLMs locally, prompt engineering, and many bias mitigation techniques (e.g., diverse data, basic debiasing methods). Federated learning is also being actively implemented and explored in various applications.
- **Research and Development:** Advanced privacy-preserving techniques like homomorphic encryption and secure multi-party computation are still largely in the research and development phase for widespread application with large-scale LLMs due to computational costs and performance limitations. Some advanced bias mitigation techniques and more sophisticated security measures also continue to be areas of active research.

8. Comparison and Contrast of Mitigation Approaches

The following table provides a comparison of some of the key mitigation strategies discussed in this report:

Mitigation Strategy	**Effectiveness**	**Ease of Implementation**	**Types of Adverse Effects Addressed**
ShieldGemma	High for targeted harm categories	Relatively straightforward API integration	Harmful content (hate, harassment, sexual, dangerous)
Llama Guard 3	High for a broad range of harms, multilingual	Relatively straightforward API integration	Harmful content, prompt injection, code interpreter abuse, election misinformation
Model Armor	High for security and safety, comprehensive	Requires Google Cloud setup and API integration	Prompt injection, jailbreak, data leaks, harmful content, malicious URLs
Federated Learning	High for privacy in training	Technically complex, requires specialized frameworks and expertise	Data privacy, security in distributed training
Local LLMs	High for data control and privacy	Can be complex, requires significant hardware and expertise	Data privacy, security from external threats
Prompt Engineering	Varies, can be effective for specific issues	Relatively easy, requires understanding of prompting techniques	Harmful content, bias, misinformation
Bias Mitigation	Varies depending on	Ranges from simple (diverse data) to	Bias in generated text

Techniques	the technique	complex (fairness-aware algorithms)	

9. Conclusion

Mitigating the adverse effects of GAI LLMs requires a comprehensive and multi-layered approach, leveraging a combination of content moderation tools, privacy-preserving techniques, and responsible deployment strategies. Tools like ShieldGemma, Llama Guard, and Google's Model Armor offer valuable capabilities for detecting and preventing harmful content and security threats. Privacy-preserving techniques such as federated learning and local LLM deployment address critical concerns around data privacy and security, although they often come with technical and resource-related challenges. Furthermore, a wide array of additional mitigation strategies, ranging from prompt engineering and bias reduction techniques to robust security protocols and ethical guidelines, contribute to a more responsible and trustworthy AI ecosystem.

The field of LLM safety and security is continuously evolving, with ongoing research and development leading to new and improved mitigation techniques. It is crucial for developers, researchers, and users to remain informed about these advancements and to adapt their strategies accordingly. The shared responsibility in ensuring the safe and ethical use of GAI LLMs is paramount to fostering trust and realizing the full potential of these powerful technologies across various domains. As LLMs become increasingly integrated into our lives, the continued focus on developing and implementing effective mitigation strategies will be essential for navigating the associated risks and maximizing their societal benefits.

Works cited

1. Editorial – The Use of Large Language Models in Science: Opportunities and Challenges, accessed on March 21, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC10485814/
2. How Large Language Models Could Impact Jobs - Knowledge at Wharton, accessed on March 21, 2025, https://knowledge.wharton.upenn.edu/article/how-large-language-models-could-impact-jobs/
3. LLMs: The Dark Side of Large Language Models Part 1 | HiddenLayer, accessed on March 21, 2025, https://hiddenlayer.com/innovation-hub/the-dark-side-of-large-language-models/
4. Ethical Considerations in LLM Development and Deployment - DEV Community,

accessed on March 21, 2025, https://dev.to/nareshnishad/ethical-considerations-in-llm-development-and-deployment-3j6n

5. Bias and Fairness in Large Language Models: A Survey - MIT Press, accessed on March 21, 2025, https://direct.mit.edu/coli/article/50/3/1097/121961/Bias-and-Fairness-in-Large-Language-Models-A
6. Identifying the Risks and Challenges of Generative AI - The Environmental Blog, accessed on March 21, 2025, https://www.theenvironmentalblog.org/2024/06/risks-challenges-generative-ai/
7. AI Transparency in the Age of LLMs: A Human-Centered Research Roadmap - arXiv, accessed on March 21, 2025, https://arxiv.org/abs/2306.01941
8. AI Transparency in the Age of LLMs: A Human-Centered Research Roadmap, accessed on March 21, 2025, https://hdsr.mitpress.mit.edu/pub/aelql9qy
9. Privacy-Preserving Large Language Models: Mechanisms, Applications, and Future Directions - arXiv, accessed on March 21, 2025, https://arxiv.org/html/2412.06113v1
10. Bias in Large Language Models: Origin, Evaluation, and Mitigation - arXiv, accessed on March 21, 2025, https://arxiv.org/html/2411.10915v1
11. ShieldGemma: Generative AI Content Moderation Based on Gemma - arXiv, accessed on March 21, 2025, https://arxiv.org/html/2407.21772v2
12. ShieldGemma model card | Google AI for Developers - Gemini API, accessed on March 21, 2025, https://ai.google.dev/gemma/docs/shieldgemma/model_card
13. google / shieldgemma-9b - NVIDIA API Documentation, accessed on March 21, 2025, https://docs.api.nvidia.com/nim/reference/google-shieldgemma-9b
14. google/shieldgemma-2b - Hugging Face, accessed on March 21, 2025, https://huggingface.co/google/shieldgemma-2b
15. LLM Content Safety Evaluation using ShieldGemma | by PI | Neural Engineer | Medium, accessed on March 21, 2025, https://medium.com/neural-engineer/llm-content-safety-evaluation-using-shieldgemma-ea05491da271
16. Shieldgemma 27b · Models - Dataloop, accessed on March 21, 2025, https://dataloop.ai/library/model/google_shieldgemma-27b/
17. Llama Guard – Vertex AI - Google Cloud console, accessed on March 21, 2025, https://console.cloud.google.com/vertex-ai/publishers/meta/model-garden/llama-guard?authuser=0
18. LLM guardrail tutorial with Llama Guard 3-11b-vision in watsonx | IBM, accessed on March 21, 2025, https://www.ibm.com/think/tutorials/llm-guardrails
19. Llama Guard 3 8B · Models - Dataloop, accessed on March 21, 2025, https://dataloop.ai/library/model/meta-llama_llama-guard-3-8b/
20. Llama-guard3-1b Model - MAX Builds, accessed on March 21, 2025, https://builds.modular.com/models/llama-guard3/1b
21. [2411.10414] Llama Guard 3 Vision: Safeguarding Human-AI Image Understanding Conversations - arXiv, accessed on March 21, 2025, https://arxiv.org/abs/2411.10414

22. Llama Guard 3 Vision: Safeguarding Human-AI Image Understanding Conversations - arXiv, accessed on March 21, 2025, https://arxiv.org/html/2411.10414v1
23. llama-guard3 - Ollama, accessed on March 21, 2025, https://ollama.com/library/llama-guard3
24. Llama Guard 3-11B-vision Model Card - GitHub, accessed on March 21, 2025, https://github.com/meta-llama/PurpleLlama/blob/main/Llama-Guard3/11B-vision/MODEL_CARD.md
25. Trust & Safety - Llama, accessed on March 21, 2025, https://www.llama.com/trust-and-safety/
26. Model Armor overview | Security Command Center | Google Cloud, accessed on March 21, 2025, https://cloud.google.com/security-command-center/docs/model-armor-overview
27. Google Cloud Model Armor -. Strengthening AI Security in a… | by …, accessed on March 21, 2025, https://medium.com/google-cloud/google-cloud-model-armor-2696bd719e02
28. Google Cloud Model Armor. To Secure Your Generative AI… | by …, accessed on March 21, 2025, https://medium.com/google-cloud/google-cloud-model-armor-6242dbae90b8
29. Advancing the art of AI-driven security with Google Cloud, accessed on March 21, 2025, https://cloud.google.com/blog/products/identity-security/advancing-the-art-of-ai-driven-security-with-google-cloud-at-rsa
30. (PDF) Privacy-Preserving Techniques in Generative AI and Large Language Models: A Narrative Review - ResearchGate, accessed on March 21, 2025, https://www.researchgate.net/publication/385514119_Privacy-Preserving_Techniques_in_Generative_AI_and_Large_Language_Models_A_Narrative_Review
31. Federated Learning - DeepLearning.AI, accessed on March 21, 2025, https://www.deeplearning.ai/short-courses/intro-to-federated-learning/
32. Federated learning and LLMs: Redefining privacy-first AI … - Outshift, accessed on March 21, 2025, https://outshift.cisco.com/blog/federated-learning-and-llms
33. Safely Learning with Private Data: A Federated Learning Framework for Large Language Model - arXiv, accessed on March 21, 2025, https://arxiv.org/html/2406.14898v4
34. Federated Learning with Large Language Models: Balancing AI Innovation and Data Privacy | by Aparna Bhat | Medium, accessed on March 21, 2025, https://medium.com/@bhat_aparna1/federated-learning-with-large-language-models-balancing-ai-innovation-and-data-privacy-2425b3e0044e
35. LLM Security: Top 10 Risks and 7 Security Best Practices - Exabeam, accessed on March 21, 2025, https://www.exabeam.com/explainers/ai-cyber-security/llm-security-top-10-risks-and-7-security-best-practices/
36. Privacy Attacks in Federated Learning | NIST, accessed on March 21, 2025, https://www.nist.gov/blogs/cybersecurity-insights/privacy-attacks-federated-learning

37. How To Preserve Data Privacy In LLMs In 2025 - Protecto.ai, accessed on March 21, 2025, https://www.protecto.ai/blog/how-to-preserve-data-privacy-in-llms
38. Local LLMs: Balancing Power, Privacy, and Performance - Pale Blue, accessed on March 21, 2025, https://www.paleblueapps.com/rockandnull/local-llms-balancing-power-privacy-and-performance/
39. Lower Data Breaches and Security Risks with Local Language Models - Netguru, accessed on March 21, 2025, https://www.netguru.com/blog/lower-data-breaches-and-security-risks-with-local-language-models
40. Leveraging Local LLMs and Secure Environments to Protect Sensitive Information, accessed on March 21, 2025, https://a-team.global/blog/leveraging-local-llms-and-secure-environments-to-protect-sensitive-information/
41. Local LLMs: The key to security, cost savings, and control | Geniusee, accessed on March 21, 2025, https://geniusee.com/single-blog/local-llm-models
42. The Pros and Cons of Using Large Language Models (LLMs) in the ..., accessed on March 21, 2025, https://www.datacamp.com/blog/the-pros-and-cons-of-using-llm-in-the-cloud-versus-running-llm-locally
43. Enterprise Challenges in Deploying Open-Source LLMs at Scale: Where Do You Struggle Most? : r/LocalLLaMA - Reddit, accessed on March 21, 2025, https://www.reddit.com/r/LocalLLaMA/comments/1h0bh2r/enterprise_challenges_in_deploying_opensource/
44. Exploring Local Large Language Models and associated key challenges - Medium, accessed on March 21, 2025, https://medium.com/@uppadhyayraj/exploring-local-large-language-models-and-associated-key-challenges-93618691fc9b
45. LLM Privacy and Security. Mitigating Risks, Maximizing Potential… | by Bijit Ghosh - Medium, accessed on March 21, 2025, https://medium.com/@bijit211987/llm-privacy-and-security-56a859cbd1cb
46. LLM Security Risks & Best Practices To Mitigate Them - Granica AI, accessed on March 21, 2025, https://granica.ai/blog/llm-security-risks-grc
47. Privacy-Preserving Techniques in Generative AI and Large Language Models: A Narrative Review - MDPI, accessed on March 21, 2025, https://www.mdpi.com/2078-2489/15/11/697
48. Privacy Preserving Machine Learning Methods and Challenges - BytePlus, accessed on March 21, 2025, https://www.byteplus.com/en/topic/405906
49. Protecting LLMs against Privacy Attacks While Preserving Utility, accessed on March 21, 2025, https://www.scirp.org/journal/paperinformation?paperid=136070
50. grapesfrog/GAI-is-going-well: This is a curated collection of articles and research papers related to the unexpected or unwanted outcomes , security & privacy risks associated with using LLMs/GAI. - GitHub, accessed on March 21, 2025, https://github.com/grapesfrog/GAI-is-going-well
51. GAI Is Going Well part deux · Missives about mostly GCP related things, accessed

on March 21, 2025, https://grumpygrace.dev/posts/gai-is-going-well-ii/

52. GAI Is Going Well · Missives about mostly GCP related things, accessed on March 21, 2025, https://grumpygrace.dev/posts/gai-is-going-well/
53. Using Generative AI in the Practice of Corporate Law - California Lawyers Association, accessed on March 21, 2025, https://calawyers.org/business-law/using-generative-ai-in-the-practice-of-corporate-law/
54. When AI Gets It Wrong: Addressing AI Hallucinations and Bias, accessed on March 21, 2025, https://mitsloanedtech.mit.edu/ai/basics/addressing-ai-hallucinations-and-bias/
55. Assessment and Assignment Guidance in the GAI Era - WordPress Service - UW-Green Bay, accessed on March 21, 2025, https://blog.uwgb.edu/catl/assessment-and-assignment-guidance-in-the-gai-era/
56. 12 key risks associated with Generative AI (GAI) | by Tahir - Medium, accessed on March 21, 2025, https://medium.com/@tahirbalarabe2/12-key-risks-associated-with-generative-ai-gai-9323a29f51b2
57. How to mitigate bias in LLMs (Large Language Models) - Hello Future, accessed on March 21, 2025, https://hellofuture.orange.com/en/how-to-avoid-replicating-bias-and-human-error-in-llms/
58. Ethical Considerations in LLM Development - Gaper.io, accessed on March 21, 2025, https://gaper.io/ethical-considerations-llm-development/
59. Using Transparency to Handle LLMs Bias | by Devansh - Medium, accessed on March 21, 2025, https://machine-learning-made-simple.medium.com/using-transparency-to-handle-llms-bias-d5b992df8f07
60. The Only Way is Ethics: A Guide to Ethical Research with Large Language Models - arXiv, accessed on March 21, 2025, https://arxiv.org/html/2412.16022v1
61. MBIAS: Mitigating Bias in Large Language Models While Retaining Context - ACL Anthology, accessed on March 21, 2025, https://aclanthology.org/2024.wassa-1.9.pdf
62. Ethical Considerations and Best Practices in LLM Development - Neptune.ai, accessed on March 21, 2025, https://neptune.ai/blog/llm-ethical-considerations
63. Understanding and Mitigating Bias in Large Language Models (LLMs) - Digital Bricks Empower AI Education, accessed on March 21, 2025, https://www.digitalbricks.ai/blog-posts/understanding-and-mitigating-bias-in-large-language-models-llms
64. View of Empirical Study and Mitigation Methods of Bias in LLM-Based Robots, accessed on March 21, 2025, https://drpress.org/ojs/index.php/ajst/article/view/24872/24359
65. Ethical Considerations and Fundamental Principles of Large Language Models in Medical Education: Viewpoint - PMC, accessed on March 21, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC11327620/