The Emergence and Resolution of Bias Issues in Large Language Models

Dai Zhang[1]

Continuing Education College, Jinan University, Jinan, 266011, China

ZDC12345678901@outlook.com

Abstract. The issue of bias in the output results of large language models has sparked extensive discussions and research in academia and industry. This article systematically analyzes the manifestations and formation mechanisms of bias in big language models, and points out that current governance methods still have limitations in data transparency and causal reasoning mechanisms, and explores future development directions. Research has found that bias is mainly divided into two aspects: "insufficient interpretability" and "narrow generalization ability". Their formation is related to data-driven bias inheritance, black box characteristics and optimization bias of model architecture, and bias reinforcement brought about by feedback loops. Although some current governance methods, such as adding interpretability tools, multimodal fusion, and data synthesis and reinforcement, have to some extent alleviated bias issues, effective governance has not yet been carried out from the root causes and processes of bias. Based on the limitations of the governance methods mentioned above, research proposes that in the future, efforts can be made to build causal frameworks, promote scenario based dynamic reasoning, establish trustworthy data systems, and establish interdisciplinary collaboration mechanisms to address both the symptoms and root causes of bias, providing theoretical support for the development of more fair, reliable, and responsible new language models.

Keywords: Bias of Large Language Models, Output Issue, Development of LLM.

1. Introduction

The rapid development of artificial intelligence technology, especially the continuous improvement of the functions of large language models and the expansion of their application scope (Taani and Alabidi, 2024), which is reshaping the way people obtain information and make decisions by integrating human like reasoning processes, thereby accelerating the process of digital transformation. However, as the application scope of LLM continues to expand, the issue of bias in output content is becoming increasingly prominent. This not only threatens personal information security and erodes public trust in artificial intelligence, but also poses a potential threat to national security and global consensus, hindering the pace of national development.

Currently, the solution to the problem of bias in large language models focuses on the induction of bias and preliminary exploration of governance strategies (Raisch and Fomina, 2024). However, induction is limited to the description of surface phenomena, lacking systematic induction of the mechanisms of bias formation and dynamic evolution of its output content in different scenarios. At the same time, governance methods mostly focus on correcting the output results, without addressing issues related to the output process such as the model's raw data and content output mechanism. Therefore, the governance method of "treating symptoms without addressing the root cause" cannot solve the output bias and bias reinforcement cycle of the big language model.

In this context,the article starts from the two main manifestations of output bias - insufficient interpretability and poor generalization ability - to explore its theoretical and practical significance, and then reveals the dual limitations of bias in its formation and application; furthermore, the logic of bias formation and reinforcement is revealed from three levels: data inheritance, model architecture, and feedback loop. Finally, reflect on the limitations of existing governance models and provide

[1] https://orcid.org/0009-0006-9707-0217

suggestions for the scientific development of AI from four aspects: causal inference construction, multimodal fusion, data governance, and ethical coordination.

2. The Main Manifestations And Formation Mechanisms Of Bias In Llm

The bias problem of big language models is not simply output errors, but a systemic flaw deeply rooted in their data dependencies, model architecture, and optimization paradigms (Li and Yao, 2025). This flaw makes it impossible to truly achieve neutral and objective analysis, and instead maps and amplifies many existing biases and social issues. To systematically understand and address bias, it is necessary to deeply analyze its core manifestations and underlying formation mechanisms.

2.1. The Two Core Manifestations of Prejudice

From a macro perspective, the bias of LLM is mainly manifested in two interrelated aspects - the reasoning flaw of "knowing what it is but not knowing why it is" and the application limitation of "narrow range of abilities".Together, these two constitute a breeding ground for prejudice to manifest."

2.1.1 Knowing its Nature but Not its Reason - Insufficient Interpretability

This performance reveals the fundamental limitations of LLM in cognitive aspects such as causal reasoning and logical deduction (Xu, 2024). It can provide answers and even generate seemingly reasonable explanations, but its internal lack of a true causal reasoning and logical deduction ability leads to a serious lack of interpretability of its output results - the explanations provided mostly rely on surface level statistical data related to the output results rather than rigorous causal reasoning and deep reasoning thinking. Due to the lack of a true reflection of the reasoning process, people are unable to fully trust the output results and cannot trace their origins.

The output of LLM relies on statistical correlation rather than logical deduction. The output results of LLM rely on complex patterns based on probability matching, and its "decisions" depend on statistically correlated parts matched in massive training data,.which differs significantly from the traditional human mode of reasoning and understanding. For example, in political stance detection tasks, the model may learn surface data from the corpus and conclude that articles with terms such as' unemployment rate 'and' tax cuts' are more likely to be conservative. It can output this judgment and use interpretability tools to only contribute these keywords to the classification results. However, this is only a post hoc, correlation based statistical report, rather than a priori, causal logical reasoning framework. Once users ask deeper questions about the issue, such as why these words represent what position in these specific contexts? How are these arguments logically organized to support a viewpoint? The model is difficult to provide a convincing and consistent explanation. Therefore, although large language models can efficiently complete problem recognition and text generation tasks, their operating mechanisms are fundamentally different from human rational thinking based on causality and logic. This gap is particularly prominent in high-risk fields such as medical diagnosis, judicial decision-making, and financial risk control. For example, in the medical field, LLM recommends a medication that is appropriate for the patient's disease based on existing data, but it cannot correctly provide contraindications and conflicts between this medication and the patient's specific medical history, which may lead to side effects and even threaten the patient's life safety. The gap indicates that reliance on LLM outputs for critical decisions requires caution.It is essential to remain cautious and recognize that the intelligence demonstrated by the model is ultimately a sophisticated statistical simulation, rather than genuine cognition and understanding.

The current interpretability techniques of LLM often only provide superficial common features, making it difficult to reveal the underlying reasoning logic behind the output results. The existing model interpretation techniques mostly remain at the level of "which common features trigger data judgment", like a tool that only presents numerical values but cannot reveal principles. This superficial explanation not only fails to help understand the true operating mechanism of the model, but may even mislead users. It should rationalize a decision based on absurd "shortcuts". For example, in an environmental stance classification task, the model may misjudge an article as "opposing

environmental protection" simply because words such as "company" and "profit" frequently appear in it, completely ignoring the fact that the main idea of the entire article is actually to argue for the importance of "sustainable business". Explanatory tools honestly report that the words "company" and "profit" contribute a high weight to the predicted results, thereby embellishing and masking this biased and clearly erroneous decision-making logic.Such misinterpretations can have far-reaching effects, as it not only leads users to blindly trust a flawed system, but also solidifies and amplifies their biases in automated decision-making; It will also hinder developers from discovering and correcting the true flaws in the model, ultimately eroding the reliability and fairness of the entire AI system. Therefore, explanations lacking deep reasoning not only fail to bring true transparency, but may also become a "fig leaf" to cover up model biases and errors.

2.1.2. Narrow Ability Range - Poor Generalization Ability

The generalization ability of large language models has significant narrowness in their application, which makes them unable to provide specific explanations for specific scenarios and systematically induce and amplify existing biases when dealing with complex and ever-changing real-world problems (Wang and Li, 2025).

Firstly, the singularity of task types leads to highly optimized models for clean tasks such as "text summarization," "text generation," or "sentiment analysis," but they perform poorly in more complex and closely related real-life tasks such as fact-checking and malicious-behavior recognition, multimodal reasoning, and dynamic tracking. This limitation amplifies bias, as the model erroneously transfers surface associations learned in simple tasks to complex scenes. For example, associating certain keywords with specific positions and mistakenly applying them to complex scenarios that require rigorous factual and logical judgment, thereby solidifying existing biases in their training data into seemingly authoritative but actually erroneous conclusions.

Secondly, the singularity of data modality means that current LLMs are mostly used for pure text processing tasks, especially for processing well structured news or book corpora. However, the real world is highly multimodal, with output and presentation including various forms such as images, audio, video, emojis, and more. The model lacks in-depth analysis, understanding, and fusion capabilities for non textual information, resulting in one-sided and incomplete analyses. This amplifies bias, as the model makes judgments based on fixed information, completely ignoring the disruptive information that image or video context may bring, thus unable to derive the true meaning of non textual content. For example, a satirical emoji may be mistaken for genuine support, leading to absurd conclusions that are completely contrary to the true intention.

Thirdly, the singularity of analytical dimensions is manifested in LLM often focusing on the analysis of individual articles or speeches, which has strong isolation and lacks specific analysis of users' long-term behavior sequences, community interaction patterns, or cross temporal evolution trajectories of topics from a macro and dynamic perspective. This perspective will amplify bias because it cannot distinguish whether the user's output content is an accidental slip of tongue or a continuous behavioral pattern. Specifically, the model may easily label a user who is usually rational and has only used a sensitive word once as extreme, assuming that the user itself has extremism; On the contrary, it may also be difficult to identify the sophisticated strategies of truly extreme users who maliciously guide certain public opinions through long-term covert and collaborative means.

Finally, the biggest application problem of LLM in implementation scenarios is that the input data for model training and testing usually assumes well structured and context rich long texts, and the information to be processed has idealized features that are detached from specific real-life scenarios. This makes it difficult for them to understand the true meaning of the text content when dealing with fragmented, non-standard, diverse viewpoints, and obscure expressions on social media, resulting in incorrect output results. In such scenarios, models tend to rely more on their existing simple and superficial statistical data to make judgments, which undoubtedly amplifies inherent biases in the original data and crudely categorizes marginal groups or complex issues, greatly weakening their effectiveness in real-world scenarios(Zhou et al., 2025).

In summary, this narrow ability is by no means a neutral feature in technology. It systematically induces and amplifies biases when models face complex real-world tasks that they are not good at handling. The judgment of the model will degenerate into a kind of prejudice and discrimination based on one-sided and erroneous information due to its multidimensional complexity of being "invisible" and "incomprehensible" to the real world, ultimately resulting in inaccurate output results and causing many unfair social problems, with unpredictable harm.

2.2. The Formation Mechanism of Prejudice

The bias issue in large language models is a systemic problem that involves the combined effects of data, model architecture, and feedback loops (Dai and Huang, 2025). Its bias stems not only from the inheritance of existing biases in its original training data, but also from the recognition and correction difficulties caused by the black box characteristics of neural networks and optimization target deviations, and more importantly, from the continuous reinforcement of existing biases in the data loop generated by the model. These three mechanisms are intertwined, which not only solidify prejudice, but also amplify and automate it, posing a deep threat to social fairness and trust. Facing up to these multiple and systematic deficiencies is an important prerequisite for building a more just, controllable, and responsible environment for the development of artificial intelligence in the future.

2.2.1 Data Driven Bias Inheritance

LLM does not "learn" the world from scratch, but rather compresses and simulates its training data, that is, the volume of texts on the Internet. The original data formed by the Internet text itself is the product of the long-term development of human society, inevitably embedded with various prejudices left over by history and inherent in society, such as gender, race, region, political position prejudice, etc. The model internalizes these biases as part of its own perceived 'world knowledge' while learning these primitive languages. The parts of the training data that are marginalized, overrepresented, or distorted will be absorbed by the model and reflected in its output results, and further amplified,resulting in a‘biased in, biased out’effect.

This bias greatly weakens the reliability and fairness of the model itself. The most direct manifestation is that bias can lead to the model outputting incorrect or unfair results in important application scenarios, giving users incorrect guidance or advice, gradually eroding their trust. For example, in the medical consultation scenario, if there is a lack of specific medical data for a certain population in the training data, such as information about drug allergies, the health recommendations provided by the model may not only be ineffective but also harmful to that group, leading to a series of medical accidents and exacerbating the unequal allocation of medical resources.

2.2.2 Limitations of Model Architecture

The black box nature of neural networks and the deviation from optimization objectives together form a highly risky structural problem, greatly exacerbating the concealment and harmfulness of biases in AI systems.

The combination of the two makes it extremely difficult to identify and correct biases. Deep neural networks interact with billions of parameters in a nonlinear manner, and their decision-making process is like a black box that is difficult to decipher, making it difficult for people to trace the specific sources of bias. At the same time, the optimization objectives set by the model during training drive it to excessively pursue statistical consistency with the training data on the surface during the result generation process, while neglecting the deep understanding of the intrinsic connections between the results. This means that the model will prioritize learning the easiest "shortcuts" in the data, namely various biases and stereotypes, rather than the laborious logical reasoning of the problem and causal testing of the results. A typical example is the recruitment screening system. Assuming there is a bias in historical recruitment data that there are more male programmers. The model will easily learn to implicitly associate "women" with "weak technical abilities", thereby systematically lowering the ratings of female job seekers in resume screening, resulting in gender discrimination. Due to the black box nature of the model, even if discrimination is detected in the results, it is

extremely difficult to pinpoint which neurons or parameter paths in the model caused this erroneous decision, making it impossible to perform precise "surgical" correction. In the end, this system not only replicated past discrimination, but also appeared "objective" due to its automated decision-making, rationalizing and institutionalizing discrimination, further increasing the difficulty of correction.

From a societal perspective, this harm is twofold. On the one hand, it systematically marginalizes vulnerable groups and exacerbates social inequality. When decisions based on biased black box models are applied to key areas such as healthcare, finance, justice, and education, their impact may cover the entire social system. It will deprive specific groups of their rights to fair resources, opportunities, and services on a large scale and automatically.This unfairness does not stem from individual intent, but from systemic technological flaws, making it more covert and difficult to hold accountable, ultimately solidifying or even exacerbating existing social class and power inequalities. On the other hand, it erodes public trust in artificial intelligence and even social systems. When people realize that major decisions that affect their lives are made by an inexplicable and potentially biased 'black box', they will inevitably have deep doubts about the fairness of LLM. Once the trust of users in the model collapses, it will not only hinder the responsible application and development of artificial intelligence technology, but also spread to trust in government agencies, enterprises, and even the entire social system that adopt the technology, fundamentally shaking the foundation of social governance.

2.2.3 Feedback Loop and Bias Reinforcement

Feedback loops and bias reinforcement are extremely dangerous and often overlooked mechanisms (Zhu, 2024). When a biased model is used to generate new data, it forms a self reinforcing bias loop. The model generates raw data and preliminary biased results based on the shallow and biased associations it learns. These newly generated data further reinforce the model's existing stereotypes and biases, leading to more deeply rooted and difficult to modify inherent biases in the next generation model's generated results. For example, in a debate on climate policy within a country, if the initial model has bias and believes that the argument of "supporting traditional energy" must be highly bound to "a specific region" or "a certain occupational group", then when the model is asked to generate the viewpoint of "supporting traditional energy", the synthesized data it produces will contain a large amount of stereotypical regional or occupational labels. These data with strong bias imprints, once used to train subsequent models, will continuously deepen the erroneous and harmful association result that 'a certain region or group will inevitably oppose environmental protection'.

From a national perspective, the harm of this bias cycle is rapidly amplified in the continuous practical application. Firstly, the model will systematically deepen the differences between different groups in China, by reinforcing the opposition and stereotypes between different groups, significantly hindering the formation of consensus within the group. The model simply attributes complex economic and social issues to geography, occupation, or identity, constantly hindering mutual understanding within groups, thereby exacerbating misunderstandings and hostility between different factions within society and ultimately hindering the formation of social consensus. Secondly, the public policies formulated based on biased model analysis are based on a one-sided and distorted understanding of social groups, which is seriously different from objective facts and leads to a series of serious problems that hinder national development, such as resource mismatch and policy failure. For example, in areas such as regional industrial transformation policies or social welfare distribution, such flawed decisions not only fail to solve existing injustices, but may even trigger new injustices due to existing social problems such as wealth inequality, resulting in even more detrimental issues to social development, such as group strikes. Finally, biased model outputs can distort the true image of a country and lay hidden dangers for international misunderstandings and conflicts. If a country's AI output frequently presents stereotypical impressions of specific internal groups, itmay lead to significant misinterpretations in international communications. The misunderstanding caused by this prejudice will spread to the international level through the Internet, distort the international image of

the country, and may be used by external forces to incite international opposition, or even trigger regional disputes, endangering national security.

3. The Evolution Of Bias Goverment Path And Current Bottlenecks

The governance of bias in big language models is a process that continues to deepen with technological evolution. The emergence of each new method aims to address the shortcomings of previous methods, but often also introduces new and more complex challenges. This section combines governance methods with existing bottlenecks to clearly demonstrate the spiral development process of this field.

3.1. From Artificial Features to Automatic Representation: the Paradox of Interpretability and Deep Semantic Capture

At the beginning of the evolution of governance paths, traditional learning methods such as SVM and random forest heavily relied on manual data annotation. Researchers need to manually design features and classify them based on domain knowledge, such as vocabulary, grammatical structure, sentiment scores, etc., for model learning. The core advantage of this method lies in its strong interpretability. Due to each feature being artificially defined, the decision-making logic is relatively transparent, making it easy to trace and audit the sources of bias. However, its bottleneck is also evident, as the design of the model is highly subjective and shallow, and it is difficult to capture complex and subtle semantic biases in language, resulting in poor application performance. Based on the huge labor cost of traditional methods, deep learning has achieved a paradigm shift through end-to-end methods, enhancing the ability of neural networks to automatically learn text representations. This method reduces LLM's reliance on subjective human experience and can automatically capture complex nonlinear patterns in text, significantly enhancing its generalization ability. The iteration of technological progress is further reflected in the application of advanced training techniques such as reordering and direct preference optimization , which optimize output quality by selecting the optimal solution from generated candidate answers or directly learning human preferences.

However, this evolution also amplifies another characteristic of LLM output results - a sharp loss of interpretability. Neural networks have become a 'black box', and their automatically learned semantic representations and decision-making processes are difficult to interpret causality. Althoughthe interpretable AI tools can provide feature attribution maps, they only reveal statistical correlations based on data rather than causal logic derived through thinking. For example, the tool can indicate that the term "unemployment rate" contributes significantly to the classification results, but cannot explain the "why" and "how" contributions. This leads to bias governance being limited to post correction at the output level, unable to address the underlying mechanisms of bias in the model, and even making auditing and eradicating bias exceptionally difficult, making it even more difficult toresolve the conflict between interpretability and representational depth in LLM outputs.

3.2. From Performance Driven to Multimodal Fusion: The Conflict between Narrow Competence and Generalization Requirements

The early models of governance path evolution were overly used to handle single, clean tasks such as optimizing "text summaries", ignoring their limitations in complex real-world scenarios, resulting in a narrow range of capabilities. To meet the complex and multimodal task demands of the real world, technological evolution is committed to moving from single task output to multi task and multimodal integration. The focus of researchers' work has shifted towards developing models that can simultaneously understand and process multimodal non textual information such as text, images, audio, and video. The goal is to enable AI to have more comprehensive and contextualized understanding capabilities, in order to reduce bias caused by one-sided analysis and avoid one-sided judgments due to modal loss.

Although the emerging new bottle neck is oriented correctly, the generalization ability of the current model is still seriously insufficient. LLM performs well in scenario analysis tasks within the training data distribution, but its output interpretability performance sharply decreases when analyzing complex and dynamic scenarios outside the distribution. For example, in a multimodal context, the model may completely ignore ironic expressions in images or tone and intonation in videos, and only focus on keywords that can be found in the database based on text to output results. This static and isolated analytical perspective cannot effectively track the dynamic evolution of user positions or public opinion topics, thus systematically inducing misjudgments and amplifying biases in real applications. The conflict between LLM's ability optimization and generalization requirements remains a core bottleneck in current governance practices.

3.3. From Real Data to Synthetic Data: The Dilemma of Data Scarcity and Bias Loops

The evolution of governance paths and the scarcity of annotated data are key factors that constrain model performance. In the context of data scarcity, generative big language models and synthetic data methods have emerged. Under this method, LLM utilizes existing manually annotated data to generate a large amount of new synthetic data, which is used to enhance the output quality of training results or directly perform stance detection. It efficiently alleviates the problem of data scarcity and can generate diverse, high-quality data or result samples at extremely low cost, especially in resource scarce fields.

The emergence of new bottle necks carries a huge risk - the bias reinforcement cycle. Because "the essence of computer computing is still human computing", the base model for generating composite data may itself be biased. Under the mode of using its own data cycle generation, the generated data will inherit or even amplify the original bias. If these data are used to train new models without careful verification, a vicious cycle of "original bias → generating biased data → training new models → exacerbating bias" will form. Currently, there is a lack of effective mechanisms to ensure the fairness and diversity of synthesized data, which makes this approach a double-edged sword. While solving old problems, it may lead to more serious systemic risks.

3.4. The superficiality of Governance Paradigms: Without Touching Upon the Fundamental Mechanisms Underlying Bias

Looking at the current governance path, both traditional methods and cutting-edge technologies have a fundamental limitation - the superficiality of governance. Most methods focus on intervening at the output layer of the model, such as filtering sensitive words through post-processing rules or aligning output styles with human preferences through RLHF. These measures only address the symptoms rather than the root cause, failing to touch upon the inherent core mechanism of bias in the model - the formation and calculation process of its internal representations. Due to the lack of systematic modeling and fundamental intervention on the sources of bias, bias is often only temporarily suppressed or transferred, and has not been eradicated. As long as the data output of relevant participation results is not corrected, similar output biases may reappear in another task or another expression.

In summary, the governance of bias in big language models is at a critical turning point. Currently, there are various technological weapons available, ranging from manual to automatic, from single to multimodal, and from real to synthetic. However, none of these methods have fundamentally solved the four bottlenecks of interpretability, generalization ability, data sources, and intrinsic mechanisms. The breakthrough progress in the future will depend on whether people can break away from the current "patching" governance paradigm and shift towards deeper exploration of causal reasoning abilities, essential understanding, and human-machine collaborative governance system of the model.

4. Futrue Outlook: Toward A More Fair And Intelligent Llm

Currently, the bias governance of the big language model is at a critical turning point from "surface correction" to "root cause governance". Although significant progress has been made in the existing governance path, its inherent bottlenecks indicate that the deepest roots of bias have not yet been truly addressed. The future development must go beyond the current "patching" paradigm and shift towards a systematic reconstruction of model cognitive architecture, data ecology, and governance system.

4.1. Building a Causal Representation Learning Framework

By deeply integrating causal inference theory into the model architecture, LLM can identify the correlation and causal relationships between different data in different scenarios, fundamentally breaking down biased decisions based on false statistical association data. Although LLM currently has a certain ability to match data correlations, it cannot understand the causal logic between things, which makes it easy for it to learn and use false correlations in data as shortcuts for judgment, resulting in absurd and biased outputs. In short, correlations without causal analysis capabilities will only generate more problems in practical applications. This improvement approach can enable LLM to be applied in high-risk areas such as political stance analysis, recruitment resume screening, credit scoring, etc. The model can avoid discriminatory judgments caused by confounding variables such as gender, race, and region, achieving a transition from outputting a potentially biased result to understanding the causal mechanism behind that result, and constructing a more fair, robust, and trustworthy next-generation LLM.

4.2. Core Direction: Implementing Contextualized Dynamic Reasoning and Action Oriented Towards the Real World

Encourage LLM development to move from handling isolated and static text tasks to a versatile intelligent agent that integrates multiple functions such as cross modal joint reasoning, dynamic contextual understanding, and response to complex open world tasks, thereby reducing the occurrence and expansion of bias and misjudgment. The current capability of LLM is confined to rationalized laboratory tasks, and its static, singular, and isolated analytical mode is seriously disconnected from the comprehensiveness and connectivity of the real world. This leads to the model beinghighly rigid when faced with tasks that require a comprehensive understanding of multimodal information, tracking temporal evolution, or processing unstructured reality. The output results are extremely one-sided, which can easily lead to and amplify misjudgments and systemic biases. The improved LLM is no longer just an "advanced parrot" that can only answer preset questions, but has become an intelligent partner that can actively perceive complex environments, integrate diverse information, understand dynamic changes, and make robust and reliable decisions, truly achieving specific problems and analysis. This will greatly expand its application boundaries and significantly reduce output bias and errors caused by narrow abilities due to its more complete understanding of the real world.

The improved LLM is better suited for scenarios that require a deep understanding of cross modal semantics, such as identifying irony and rumors in social media settings, to enhance the health of the information ecosystem; Automated video content review and summarization, providing truly contextualized environmental commentary for visually impaired individuals, and other complex real-life scenarios. This improvement can also enable LLM to be applied in scenarios that require insight into development trends and evolution, such as tracking the dynamic evolution of public opinion events and analyzing their driving factors, monitoring long-term changes in brand reputation, and conducting continuous and fair evaluations of individuals or organizations' positions and behaviors, emphasizing individual subjective outputs and avoiding a "one-time conviction". At the same time, the improved LLM can also be applied to complex tasks in open environments, such as acting as a fact checker to autonomously retrieve and verify multi-source information, as a debate assistant to

understand opposing viewpoints and logically refute them, and assisting in comprehensive research with uncertainty in fields such as finance and healthcare.

4.3. Core Direction: Building a Trustworthy and Controllable Data Governance and Evaluation Foundation

Establish a governance system that covers the entire lifecycle of data, integrates automation and manual auditing, and develops standardized evaluation benchmarks, committed to curbing the breeding and circulation of bias from the source and data ecology. There are two core pain points in current data practice. One is that the abuse of synthetic data may lead to a "bias reinforcement loop", where biased models generate biased data and train biased results, even the next generation of models; The second issue is the lack of authoritative, diverse, and standardized evaluation criteria, which leads to a lack of unified standards for measuring model bias and makes it difficult to horizontally compare and continuously improve governance effectiveness. After improvement, LLM will shift from amplifying data bias to mitigating it, greatly enhancing the objectivity and accuracy of the output results based on reality. For example, by establishing a trustworthy, controllable, and transparent data supply chain, high-quality "nutrients" and objective "standards" can be provided for training and evaluating more fair LLMs, ultimately breaking the cycle of biased output results and promoting the responsible development of the entire AI ecosystem.

After data optimization, LLM can be applied to application scenarios that require high-risk synthetic data, especially in high-risk fields such as finance, healthcare, and justice. For example, LLM can simulate and generate synthetic data of rare or non-existent extreme financial scenarios in reality, providing a foundation for financial institutions to conduct more comprehensive and forward-looking high-risk stress testing. Before using synthetic data augmentation models, strict bias auditing and filtering must be conducted through this governance framework to ensure data security and fairness. Meanwhile, it can also be applied to model fairness assessment and authentication. The standardized benchmark dataset can serve as a "touchstone" for third-party organizations to conduct fairness audits and certifications of commercial LLMs, setting entry barriers and best practices for the industry. In addition, the improved LLM can also promote collaborative open source governance, using transparent and shared data governance practices within open source communities, industry alliances, and international organizations to jointly address global AI bias challenges, avoid duplication of effort, and standard fragmentation.

4.4. Core Direction: establishing a Systematic Governance Framework with Embedded Ethics and Interdisciplinary Collaboration

It is suggested to promote the upgrading of LLM governance from a single technical repair to a systematic engineering that integrates ethical design, interdisciplinary collaboration, and adaptive regulation, ensuring that its development aligns with social values. The current governance model faces issues of "fragmentation" and "lag". Technical teams, ethics scholars, and policy makers work independently, and governance is often done as a remedy after the fact, making it difficult to fundamentally solve bias issues. Moreover, the development of technology often goes beyond regulatory and ethical frameworks, resulting in governance blind spots. Therefore, building a forward-looking, sustainable, and trustworthy LLM governance ecosystem is an important development direction for LLM in the future. Under this system, technological progress will always be guided by ethics and regulated by law, ensuring that the development of LLM truly benefits the entire society and minimizes systemic risks such as bias.

LLM optimized under the ecosystem can be applied in the pre ethical review of high-risk applications. Before the launch of LLM applications in fields such as healthcare, justice, and finance, it is mandatory to conduct interdisciplinary ethical impact assessments and embed principles of fairness and transparency into product design. At the same time, optimized LLM can also help develop industry standards and certification systems. Industry alliances, standard organizations, and regulatory bodies should jointlydevelop auditable ethical AI standards to form a unified industry

framework and provide clear guidance and market incentives for responsible innovation. Finally, optimized LLM can also be applied in the fields of government legislation and regulation. For example, setting up a "regulatory sandbox" at the government level allows companies to test innovative applications in a controlled environment, while promoting the development of agile regulations that adapt to the speed of technological iteration, balancing innovation and risk. Ultimately, the core of this series of measures lies in internalizing "responsibility" and "trustworthiness" as the genes of technological innovation, laying a solid foundation for the healthy development of LLM in high-risk and other application fields.

In summary, the future development of big language models will inevitably be a paradigm revolution from "performance first" to "responsibility first". This means that technological progress should be accompanied by ethical considerations. The future LLM should not only be a more powerful text generation tool, but also a trustworthy and impartial intelligent agent with deep understanding, causal reasoning ability, and human trust. The process of achieving this goal will inevitably face many challenges, but it is entirely achievable to build a more fair, transparent, and beneficial AI era through the global joint efforts of academia, industry, and policy makers.

5. Conclusions

This article systematically discusses the issue of bias in the big language model, starting from the manifestations and formation mechanisms of bias, and deeply analyzes the evolution and bottlenecks of the current governance path. Based on this, future oriented governance methods and development visions are proposed.

Reflecting on the research topic, the bias of large language models is not a random technical flaw, but a systematic and structural bias rooted in their data foundation, model architecture, and optimization goals. This not only affects the reliability and fairness of the model, but also poses a potential threat to society's trust in technology and even national security.

This study found that the root of bias lies in the inference defects of the model that "knows what it is but does not know why it is" and the generalization limitations of "narrow ability range"; The formation of bias is the result of the combined effects of three mechanisms: data-driven bias inheritance, black box characteristics and optimization bias of model architecture, and bias reinforcement brought about by loop feedback; The governance of current bias issues continues to make technological progress, but there are still bottlenecks such as insufficient interpretability, limited generalization ability, and data bias loops, which have not fully addressed the root causes of bias.

This article proposes suggestions for bias governance from four directions: constructing a causal representation framework, implementing emotional dynamic reasoning for the real world, proposing a trustworthy and controllable data governance and evaluation system, and promoting embedded ethics and interdisciplinary collaborative governance.

In summary, this study not only deepens the theoretical understanding of the mechanism of bias formation and evolution in large language models, but also provides a systematic reference path for model governance at the practical level. By promoting the evolution of technological paradigms from "performance first" to "responsibility first", this study proposes theoretical suggestions for building a more fair, transparent, and trustworthy artificial intelligence, which to some extent promotes the positive interaction between artificial intelligence and human society.

References

[1] Bai, Q., Kuang, Q., Xia, X., Yan, S., & Zhang, S. (2025). Prejudice mitigation: A study on the recognition efficacy and cognitive adaptation interventions of large language models. Journal of Education and Media Studies, 3, 17–27. https://www.cnki.net

[2] Dai, G., & Huang, B. (2025). A discussion on the cognitive characteristics and developmental transformation paths of large language models in translation. Chinese Translators Journal, 4, 122–129. https://www.cnki.net

[3] Huang, H., Li, S., Lan, T., Qiu, Y., Liu, Z., Yao, J., Zeng, L., Shan, Y., Shi, X., & Guo, Y. (2024). A survey on the safety of large language models: Classification, evaluation, attribution, mitigation and prospect. CAAI Transactions on Intelligent Systems, 20(1), 2–32. https://link.cnki.net/urlid/23.1538.TP.20241212.1608.002

[4] Li, Z., & Yao, F. (2025). Challenges and responses of AI-generated content: Exploration and reflection from the perspective of sustainable development of audiovisual content platforms. Radio & Television Research, 9, 118–124.

[5] Meng, L., Wang, X., & Han, C. (2025). Global AI development trends and future prospects. Contemporary China and World, 3, 25–32.

[6] Raisch, S., & Fomina, K. (2024). Combining human and artificial intelligence: Hybrid problem-solving in organizations. Academy of Management Review, 50(2). https://doi.org/10.5465/amr.2021.0421

[7] Taani, O., & Alabidi, S. (2024). ChatGPT in education: Benefits and challenges of ChatGPT for mathematics and science teaching practices. International Journal of Mathematical Education in Science and Technology. Advance online publication. https://doi.org/10.1080/0020739X.2024.2357341

[8] Wang, Y., & Li, M. (2025). Artificial intelligence application and corporate sustainable development: A test based on innovation capability and absorptive capacity. Journal of Environmental Economics, 10(3), 47–72.

[9] Xu, Y. (2024). “The AI alignment problem” posed by DeepSeek: A philosophical challenge—From the perspective of the relationship between “pre-judgment” and “prejudice”. Journal of Philosophy and Technology, 2(7), 108–118. https://www.cnki.net

[10] Zhou, W., Zhu, X., Han, Q., Li, L., Chen, X., Wen, S., & Xiang, Y. (2025). [Article title not provided]. IEEE-CAA Journal of Automatica Sinica, 12(1), 1–26. https://doi.org/10.1109/JAS.2024.124983