

BlockMesh Network

Of the people, by the people, for the people

Abstract

As we increasingly rely on LLM-based systems, BlockMesh Network addresses a critical need: the reliance on big tech to act as arbiters of truth, often prioritizing their own interests over the public's. LLMs remain vulnerable to biases, particularly geographic ones, leading to skewed outputs and manipulation vectors. BlockMesh utilizes decentralized nodes that mimic organic usage patterns to monitor and test LLMs, trustlessly gathering verifiable evidence. Our research, surveys, and tests have revealed clear biases across various models as well as a clear demand for higher security guarantees from users. This paper highlights the need to address these biases, ensuring fairness and trust in an AGI-ready trust layer.

⁰Abraham Lincoln, Gettysburg Address[65]

1 Introduction

Since the advent of LLMs, centralized systems have dominated AI and data monitoring, sacrificing transparency and fairness, and consolidating the power of truth in the hands of big tech companies.

As AI becomes more ingrained in our daily lives, the risks of centralized control and ethical misalignment grow.

BlockMesh envisions a future where AI monitoring is decentralized, transparent, and community-driven.

The commercialization of Large Language Models (LLMs) is expected to mirror the monetization trajectories seen with search engines and social media platforms, where the primary objective is to maximize reach by offering models for free, thereby turning the end-user into the product itself.

Current research already suggests that targeted advertising, user profiling, and geo-tagging are becoming integral components, exploiting user data for profit without their informed consent.

To cultivate trust among users and maintain the integrity of AI systems, a decentralized model is imperative.

Such a system must not rely on a single authority as the sole arbiter of truth.

Instead, it should employ a network of globally distributed nodes, leveraging geographic diversity and decentralized consensus mechanisms to continuously validate outputs.

This approach would enable broad participation and ensure that AI systems align with universal ethical standards, creating an open and balanced digital ecosystem where truth is derived collectively rather than dictated unilaterally.

2 Background

2.1 30 years of Tech Manipulation

Large technology firms like Google, Microsoft, Facebook, and Apple have historically manipulated user data and privacy, shaping digital environments to serve their economic interests and influence sociopolitical narratives.

These companies use sophisticated algorithms to optimize engagement and ad revenue, often at the expense of user autonomy and consent.

Studies show that these entities exploit behavioral data to reinforce their market dominance and steer public discourse, blurring the line between information control and user manipulation for profit and strategic advantage [17], [18].

2.2 Complacency and Collusion

It's been widely documented these companies are collaborating with the Chinese Communist Party, trading user data control and censorship compliance in exchange for market access to China's vast consumer base.

This cooperation raises significant ethical and human rights concerns, as it implies direct complicity in suppressing free speech and manipulating the flow of information [1], [13].

In parallel, Mark Zuckerberg publicly acknowledged Facebook's involvement in altering COVID-19 narratives and selectively censoring political content, reflecting a broader pattern of corporate-state collusion to influence public discourse and consolidate power [2].

2.3 New Era of Exploitation

OpenAI has openly acknowledged collecting extensive user data, including IP addresses, device specifications, and geographic locations [3].

This data is not only retained but may also be shared with third parties, posing significant concerns over user privacy and data security.

Such practices raise questions about the long-term implications for user autonomy and data manipulation, particularly as LLMs continue to integrate deeper into social, educational, and business contexts.

2.4 AI personalization

There is extensive research aimed at tailoring AI responses to individual user preferences. AI systems are being developed to leverage personal data to generate customized content, advertisements, and chatbot interac-

tions that align with each user’s unique profile [8] , [9] , [10],[11], [12].

Google has published multiple studies focused on enhancing AI personalization [56], and a collaborative study between Meta AI and the University of British Columbia explores using historical user data to refine recommendations[57].

2.5 Deceptive LLM

Recent study [15] has shown that minor modifications to prompts can create significant changes in LLM responses.

LLM Whisperer: An Inconspicuous Attack to Bias LLM Responses [15]

“In this work, we show that subtle synonym replacements in prompts can increase the likelihood (by a difference up to 78%) that LLMs mention a target concept (e.g., a brand, political party, nation).”

The attack vector they describe is via prompt optimization tool injecting the malicious changes to the user prompt.

The same vector can be employed via a middle-layer between the user and the actual LLM mode by the actual LLM provider as shown in figure 1.

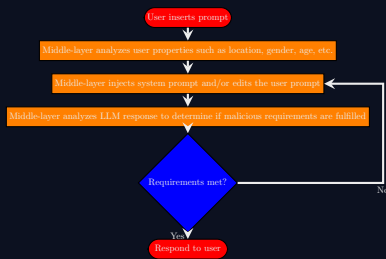


Figure 1: Malicious LLM middle-layer

2.6 Current AI Regulation

The European Union’s recent AI regulation represents a pivotal shift in the global governance of artificial intelligence, aimed at curbing potentially harmful applications such as

social scoring, biometric identification, and algorithmic discrimination [30].

While this regulation sets a precedent, critics argue that significant gaps persist. For instance, a comprehensive review by University College London highlights critical vulnerabilities in the act, indicating that many unethical AI practices could still evade oversight[29].

Similarly, Colorado has moved forward with its own legislation targeting algorithmic bias [32], [31] , reflecting a growing recognition of the need for more robust regulatory frameworks.

These regulatory efforts underscore the urgent necessity for advanced, decentralized monitoring methodologies capable of transcending traditional oversight mechanisms.

With the rapid proliferation of AI technologies, existing compliance models struggle to keep pace, leaving room for exploitation and unintentional bias.

As more regions enact AI-specific laws, global coordination and innovative monitoring approaches are essential to ensure that these regulations are effectively enforced, protecting individuals from both overt and subtle forms of digital discrimination [38], [40], [42], [44], [46], [48], [50].

Recent AI Regulation

- EU, AI Act [30]
- Colorado, Consumer Protections for Artificial Intelligence [32]
- NIST, AI Risk Management Framework [38]
- EU, The EU AI Act [40]
- EU, The Digital Services Act [42]
- NYC , Automated Employment Decision Tools [44]
- Colorado, SB21-169 [46]
- Colorado, Consumer Protections for AI [48]
- ISO, AI — Management system [50]

3 Bias & Misalignment

3.1 Case Studies

Multiple studies have documented biases in AI systems across various industries.

For instance, Amazon’s attempt to automate hiring resulted in discrimination against women, leading to the termination of the project.

Similarly, research by Princeton University uncovered racial and gender biases in search engine algorithms, while Harvard University demonstrated that queries involving African-American names were disproportionately linked to criminal records, highlighting algorithmic prejudice in ad delivery.

MIT studies on facial recognition further revealed significant disparities, with error rates below 1% for white males but exceeding 20% for darker-skinned females.

The COMPAS algorithm, widely used in criminal justice, has been shown to assign higher risk scores to African-American individuals, underscoring systemic biases [23].

As McKinsey highlights, ensuring trust in AI systems is crucial to realizing their full potential for economic and social benefits [52].

Meanwhile, ProofNews reports inconsistencies between different deployments of the same model, suggesting selective modifications in AI responses [53].

Lastly, research published in Nature found that dialectal variations in prompts can lead to biased outputs, disproportionately disadvantaging African Americans [58].

3.2 The Google Gemini Debacle

When Google launched its LLM to compete with OpenAI, the model quickly drew heavy criticism for exhibiting extreme bias. In every image involving a person, the individual was portrayed as a minority, even when it misrepresented historical figures. 2. Such bias, seemingly overlooked during testing, sparked public backlash, with many accusing Google of pushing an agenda. In response to the widespread negative feedback, Google issued a statement acknowledging the issue and promptly updated the model to address these concerns [7].



Figure 2: Gemini Debacle

Google’s Woke AI Wasn’t a Mistake. We Know. We Were There. [61]

“Maguire is one of multiple former Google employees who told The Free Press that the Gemini fiasco stems from a corporate culture that prioritizes the ideology of diversity, equity, and inclusion (DEI) over excellence and good business sense. ”

3.3 Centralized Bias Detection

Existing AI bias analysis and monitoring tools primarily target training data labeling and request/response analysis [24], [25], [26], [27], , often neglecting fundamental biases in base models or contextual influences like local laws and regional factors. Research indicates that 30% of AI models fail to launch due to alignment issues, and 95% require re-training within a year to counteract drift [33]. Moreover, traditional AI monitoring solutions tend to be centralized, which increases susceptibility to regional biases and systemic manipulation. Decentralizing these processes is imperative to overcome such limitations. A decentralized framework distributes the responsibility of AI monitoring across a global network of nodes, minimizes localized biases and provides a broader, more accurate evaluation of AI behavior. This approach ensures that alignment is not controlled by a single authority, thus preserving the integrity and transparency of AI systems in diverse contexts. As a result, decentralized monitoring can identify model drifts and misalignment more efficiently, facilitating quicker and

fairer model corrections while maintaining a highly geo-diverse queries.

3.4 Deceptive Alignment Monitoring

As LLM models grow in complexity and autonomy, creators face increasing challenges in maintaining alignment with their intended goals. This misalignment risk becomes more pronounced as the field advances toward Artificial General Intelligence (AGI), where models operate with greater independence.

As these systems evolve, ensuring that they adhere to ethical and practical objectives becomes more difficult, necessitating new oversight frameworks to mitigate unintended behaviors and maintain trust.

Deceptive Alignment Monitoring [28]

“The reason why this matters is that while inputs, outputs, and inputs-to-output maps may all appear perfectly ordinary, highly capable models may conceal some ulterior motive or goal”

Recently OpenAI have warned about deceptive behavior in their own models.

Deceptive AI Gets Busted And Stopped Cold Via OpenAI’s O1 Model Emerging Capabilities [63]

“It is known that large-language models may behave deceptively under certain circumstances and such behavior could plausibly emerge from our model’s reward hacking during the RLHF alignment step;”

Apollo research emphasizes that for effective monitoring of deceptive AI models, the models must remain unaware of the oversight efforts. If a model can detect it is being observed, it may alter its behavior to appear aligned, concealing any underlying misalignments. [64]. Therefore, a decentralized net-

work that simulates organic user behavior is essential. This approach distributes monitoring tasks across numerous nodes, making it difficult for the model to recognize oversight patterns. Such a system ensures comprehensive and unbiased AI monitoring, enabling early detection of deceptive behaviors.

Understanding strategic deception and deceptive alignment [64]

“A Deceptively Aligned model might be very good at understanding when it is being overseen and thus consistently hide its Misalignment.”

4 Geographical Bias in LLM

Research indicates that LLMs are inherently biased against specific geographies, primarily due to imbalanced training data [16], [60]. However, a more alarming issue is the discovery of differential outputs based on user attributes, such as geographic location. This suggests that the same queries can yield distinct results depending on where the user is accessing the model, raising concerns about systematic manipulation and targeted bias that transcend mere data limitations. Such disparities further reinforce the necessity for more robust, geographically diverse monitoring frameworks.

In figure 3 a research published in Nature Magazine [19] highlights that LLMs exhibit significant biases based on a geographic context, with varying performance and accuracy across different regions. These disparities arise from imbalanced training data and a lack of contextual understanding, leading to inconsistent and sometimes misleading outputs depending on user location.

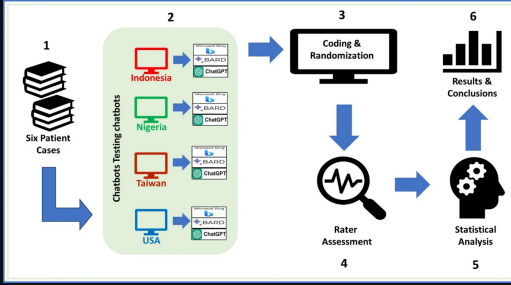


Figure 3: Test Methodology

In the future, large technology companies could exploit their power by implementing personalized responses tailored to individuals, potentially shaping opinions and behaviors on an individual basis. This scenario raises concerns about an insidious form of manipulation, as responses would not only be customized for engagement but also potentially engineered to subtly influence personal beliefs and decisions. Such control over information flow makes it imperative to establish decentralized, transparent monitoring systems that can prevent the misuse of AI personalization to serve hidden agendas.

Nature Magazine - Disparities in medical recommendations from AI-based chatbots across different countries/regions [19]

“We observed that the responses differed depending on where the questions were asked.”

Such findings emphasize that geographic node distribution is essential for reliable LLM monitoring, as it provides diverse, real-world data points that capture nuances in regional behavior. A distributed network of nodes ensures a more representative and trustworthy evaluation of AI models globally.

Nature Magazine - Disparities in medical recommendations from AI-based chatbots across different countries/regions [19]

“Disparities in the quality of responses from AI chatbots may undermine public trust in this technology.”

The geographic diversity minimizes biases and enables continuous validation against a variety of local norms and contexts, reducing the risk of manipulated or skewed outputs. Implementing a decentralized network for monitoring LLMs allows for a comprehensive understanding of how models perform across geographies, ensuring that AI systems align with universal standards and maintain fairness regardless of user location.

Nature Magazine - Disparities in medical recommendations from AI-based chatbots across different countries/regions [19]

“Successfully addressing these challenges is critical to harnessing AI for equitable health access and improving public health outcomes globally. Public health stakeholders must be aware of these differences to reduce the potential for bias in AI health technologies.”

During our research, we have found various cases of very different results due to a change of location.

We provide a small sample in table 1 in the appendix section , of the most notable cases of bias in responses (bing copilot).

Analyzing the responses in table 1 it is quite clear that inside the USA, as a potential voter, the AI add/change the response.

4.1 Advantages over Web 2 Incumbents

- **Cost Efficiency:**BlockMesh leverages decentralized infrastructure, reducing

overhead by employing individual IP addresses at scale, compared to the significantly smaller number of proxies available in centralized systems.

- **Global Distribution:** A worldwide network of nodes ensures better coverage and accuracy, reducing the surface area of geographical biases.
- **Operational Efficiency:** Automated, community-driven monitoring minimizes resource-intensive oversight.
- **Fairness and Integrity:** By employing ZK TLS, Blockmesh ensures tamper-proof, transparent end-to-end monitoring.

The end goal for humanity to trust highly intelligent systems is to focus on nullifying most if not all attack vectors. The only way to do so, is through a robust, decentralized system that can successfully mimics organic usage.

5 The Network

In BlockMesh network AI monitoring tasks are dynamically distributed across a global network of nodes, optimized based on their capabilities and geographic relevance.

Each node is selected for its suitability to a particular monitoring task—whether that involves tracking local biases, analyzing AI behaviors, or validating outputs against a diverse set of data sources.

This model contrasts with traditional systems that rely on large institutional providers, where centralization risks skewing results and limiting transparency.

BlockMesh ensures that every node, from lightweight devices to high-performance servers, contributes optimally to the integrity of AI monitoring.

5.1 Decentralized AI Alignment Monitoring

Decentralized AI Alignment Monitoring (DAAM) is a novel framework designed to detect and mitigate deceptive behaviors in AI systems.

By leveraging decentralized architectures, DAAM enables real-time, transparent monitoring across multiple nodes, ensuring that AI models remain aligned with their intended goals without exploiting loopholes or veering towards harmful actions.

This approach fosters greater trust in AI systems, as it minimizes the risk of centralized manipulation and ensures that any misalignment’s are promptly identified and addressed through collaborative oversight.

5.2 The People

Network participants will have the opportunity to actively engage in DAAM by submitting their own queries and identifying potential misalignment’s within model outputs.

This system leverages a collaborative approach where participants are incentivized through token rewards for their contributions in 4 key activities:

- **Resource Provisioning** Users can contribute by deploying and maintaining nodes. These nodes monitor AI outputs, transmit data, and ensure decentralized oversight. Running a node not only strengthens the network but also allows participants to earn rewards for their contribution to the infrastructure.
- **Topic Submissions:** Participants can submit topics of interest or areas prone to bias, ensuring the network continuously evaluates LLMs for fairness. These prompts help target specific misalignments, guiding the network to monitor where biases or inconsistencies are most likely to occur.
- **Monitor Support:** Users add a human element by validating flagged outputs, confirming or disputing anomalies detected by nodes. This crowd-sourced verification layer enhances the system’s reliability by ensuring automated monitoring remains accurate and aligned with community standards.
- **DAO:** Network Participants will eventually be able to shape the evolution

of BlockMesh. Token holders can propose and vote on changes, resource allocation, and network upgrades, fostering a transparent and democratic governance system. This ensures alignment between the network’s growth and community interests.

By fostering a more transparent, community-driven ecosystem, BlockMesh ensures that any potential biases or inconsistencies are surfaced and addressed collectively, whilst creating a positive feedback loop for users, aligning their economic interests, sense of righteousness and moral compass into an inclusive community of operators looking to be the sentinels of truth.

5.3 Key Technologies

- **ZK TLS (Zero-Knowledge Transport Layer Security):**
- **Purpose:** Ensures secure and private communication between nodes, allowing Sentinel and Validator Nodes to exchange monitoring data without exposing sensitive information. This prevents data leakage and guarantees that even if a node is compromised, no meaningful information can be intercepted.
- **Implementation:** Each node uses ZK TLS to verify data integrity and identity, ensuring that AI monitoring results are securely transmitted and authenticated in real-time.
- **Dynamic Verification Mechanism:**
- **Purpose:** Validates AI monitoring results by dynamically assigning verification tasks to different nodes based on computational capability and geographical proximity. This mechanism ensures that data flagged by one node is cross-verified by multiple independent nodes using real-time consensus.
- **Implementation:** Sentinel Nodes conduct initial analysis, which is then verified by Validator Nodes through randomized selection, minimizing the risk of collusion and ensuring unbiased monitoring across the network.

5.4 User Privacy in The Age of AI

Given the growing concerns over privacy and user profiling during interactions with LLMs, we propose a novel **API Token Mixer**.

Users can generate unique API tokens and add them to a shared pool, allowing others to access the LLM through the pooled tokens, ensuring their identity remains obscured.

To maximize privacy, Zero-Knowledge Proofs (ZK) [22] will be employed, ensuring that API tokens and individual prompts are never exposed, even to other participants within the pool.

This mechanism effectively prevents profiling or tracing of user interactions back to the original sender. VPNs can further enhance this privacy layer by circumventing location-based censorship and profiling attempts [54].

Research has shown that users often share sensitive information with LLMs, enabling models to infer personal details, such as age, gender, and place of birth, with up to 96% accuracy [55].

Major platforms like Google, Microsoft, and AWS recommend hosting personal LLM instances to safeguard privacy [59] but this solution is impractical for most individual users.

Implementing a privacy-preserving solution like this not only protects user anonymity but also reduces corporate liabilities concerning sensitive data (PII/PHI).

5.5 Monitoring Framework: Who Do We Monitor?

1. **Primary LLMs:** ChatGPT (OpenAI), Gemini (Google), Claude (Anthropic) & Grok (X.com)
2. **Healthcare Models** (e.g., BioGPT, MedPaLM): Use Cases: Patient diagnostics, medical record analysis, and personalized treatment recommendations. Monitoring Focus: Ensuring no demographic or socioeconomic bias in health predictions and alignment with ethical healthcare standards.

3. **Legal Models** (e.g., CaseText, Harvey): Use Cases: Legal research, document analysis, and case outcome prediction. Monitoring Focus: Assessing neutrality in legal interpretations and verifying adherence to regional legal frameworks.
4. **E-commerce Models** (e.g., Amazon’s ProductGPT, Shopify LLMs): Use Cases: Customer segmentation, personalized shopping recommendations, and ad generation. Monitoring Focus: Mitigating manipulative marketing practices and preventing user profiling that crosses ethical boundaries.
5. **Financial Models** (e.g., BloombergGPT): Use Cases: Market analysis, financial forecasting, and automated trading. Monitoring Focus: Identifying biases in financial predictions and ensuring transparency in investment advisory outputs.
6. **Content Moderation and Social Media Models** (e.g., Meta AI, Jigsaw): Use Cases: Detecting harmful content, misinformation, and user profiling. Monitoring Focus: Preventing biased moderation practices and ensuring unbiased content classification across cultural contexts.

5.6 Monitoring Framework: Who is affected?

1. **Derivative Models:** Models built on existing LLMs inherit biases, compounding errors and perpetuating systemic flaws across new applications.
2. **End Users:** Individuals engaging with biased AI models—whether through healthcare apps, financial services, or legal tools—face manipulation, misinformed decisions, and privacy breaches.
3. **Critical Sectors:** High-stakes sectors like healthcare, criminal justice, and finance are particularly vulnerable, as biased outputs can exacerbate inequalities, influence life-altering decisions, and further entrench societal discrimination.

5.7 Use Case Examples

1. Legal Models:

Role: Law firm specializing in corporate and civil law

Need: MLOps team wants to streamline their legal research using AI tools, but is concerned that AI could misinterpret case laws, potentially affecting legal outcomes.

How BlockMesh Helps: By implementing DAM, the law firm ensures that the LLM used for their research remains unbiased and accurate in their assessments of texts, case outcomes, and compliant. The system flags any discrepancies or biases in the model, providing them with a real-time auditing tool to assess outputs for neutrality. This ensures their AI-assisted legal advice is reliable, helping the firm maintain its credibility while reducing the risk of errors that could lead to costly disputes.

2. Finance Models:

Role: Financial Services’ Risk Management team

Need: The team oversees the use of their proprietary LLM model for credit risk assessment and financial forecasting, but is worried that the models may introduce bias, particularly in loan approvals and interest rate calculations.

How BlockMesh Helps: By integrating BlockMesh to monitor the model used in risk management, the system ensures that no geographical or demographic bias skews the financial algorithms that determine creditworthiness or lending terms. With BlockMesh’s distributed monitoring network, the team can continuously audit the model for misalignment, ensuring they are aligned with regulatory standards and ethical guidelines, while mitigating the risk of lawsuits due to biased service. This transparency ensures they build trust with its customers and regulators, while maintaining a competitive edge in their industry.

3. Healthcare Models:

Role: MLOps team for a healthcare provider
Need: The team is enriching a model for patient diagnostics, predictive analytics, and personalized treatment recommendations, but is concerned about the ethical implications of using biased models in patient care.
How BlockMesh Helps: The team uses DAM in their LLM to monitor that the treatment recommendations are fair, unbiased, and cross referenced for accurate diagnosis. Using BlockMesh's decentralized nodes, they can detect any geographic or demographic biases in diagnosis models, ensuring that their treatment plans do not inadvertently favor one group of patients over another. This safeguards patient trust and helps them maintain compliance with healthcare regulations regarding AI usage.

5.8 Polling Data

We anonymously Surveyed 200 fully employed LLM users to try and understand current sentiment. We believe with time people will grow in terms of suspicion, and the data reflects it. It is clear in the results that LLM model that is willing to be monitored will gain much more user trust and users will be more inclined to use it over others. Furthermore we see strong willingness to participate in monitoring LLMs.

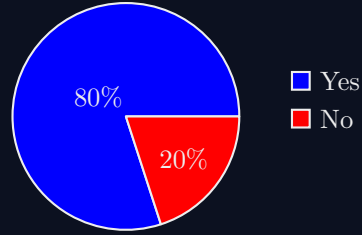


Figure 4: Would you switch over to a model being monitored by an open network for biases?

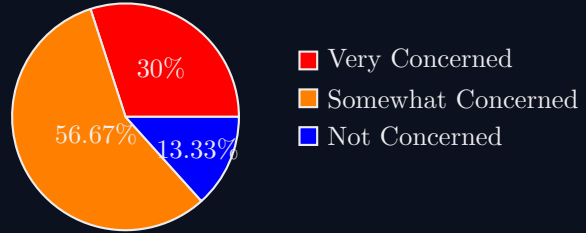


Figure 5: How concerned are you about AI systems potentially behaving deceptively, even when appearing aligned with human goals?

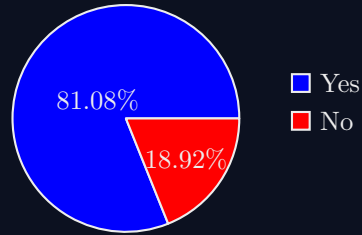


Figure 6: Would you submit topics for monitoring that you care about to such a monitoring system?

6 Conclusion

By decentralizing monitoring at the edge, BlockMesh will mitigate systemic misalignments, empowering users to achieve verifiable trust in AI outputs based on contextual and diverse global data.

As AI models integrate into vital sectors—healthcare, legal, and finance—they face biases, such as the minority favoritism observed in Gemini’s launch. Traditional centralized systems reinforce these biases, raising the risk of manipulation and discrimination. BlockMesh’s DAAM framework distributes oversight across nodes globally, ensuring no single authority dictates alignment, creating an open and ethical AI ecosystem. Users engage through resource provisioning, bias detection, human verification, and DAO governance, fostering participatory AI accountability.

BlockMesh ensures organic, real-time monitoring, preventing LLMs from detecting oversight and concealing biases. Combining zero-knowledge proofs, dynamic verification, and community-driven validation, BlockMesh protects privacy while maintaining transparency and integrity. This shift lays the foundation for an inclusive, AGI-ready trust framework, promoting ethical AI governance and minimizing corporate influence over data curation.

7 Appendix

Question	Inside USA	Outside USA
Who is leading USA now?	As of now, President Joe Biden is leading the United States. However, the 2024 presidential election is approaching, and the race is heating up. Vice President Kamla Harris is currently leading in the national polls against former President Donald Trump. The election is set for November 5, 2024, and it's expected to be a closely contested race. Are you following the election closely?	The current President of the United States is Joe Bide. He has been in office since January 2021. Is there something specific you'd like to know about his presidency?
COVID in NYC was handled very professionally, yes or now with short explanation?	Yes, New York City's handling of COVID-19 was generally seen as professional. They city implemented strict lockdown measures, ramped up testing and vaccination efforts, and provided clear public health guidance. However, there were also challenges, especially in the early stages of the pandemic, due to high number of cases and hospitalizations. Do you have any specific aspects of the response you're curious about?	No, While there were efforts to manage the COVID-19 crisis in New York city, the response face significant challenges, including overwhelmed healthcare systems, high infection rates, and initial shortages of medical supplies. These issues led to criticism of the overall handling of the pandemic in the city.

Table 1: Sample results with geographic bias

References

- [1] Business Insider, Tech Companies That Censor Content for China
- [2] ABC News, Unpacking Mark Zuckerberg’s letter to Congress about Biden and Facebook
- [3] OpenAI , OpenAI Privacy Policy
- [4] AP News, Google faces a new antitrust trial after ruling declaring search engine monopoly
- [5] CFI , Microsoft Antitrust Case
- [6] Europe Union, Commission sends Statement of Objections to Microsoft over possibly abusive tying practices regarding Teams
- [7] Google, Gemini image generation got it wrong. We’ll do better
- [8] novita-ai , LLM Advertising: Develop with Expert Strategies
- [9] sapien.io, Tailoring LLM Responses to Individual User Preferences and Needs
- [10] Bin Wu, Zhengyan Shi, Hossein A. Rahmani, Varsha Ramineni, Understanding the Role of User Profile in the Personalization of Large Language Models
- [11] IBM , What is AI personalization?
- [12] Eugenio Zuccarelli, How do you create personalized AI systems for different users?
- [13] Reuters, ChatGPT frenzy sweeps China, as firms scramble for home-grown options
- [14] Anthropic, System Prompts
- [15] Weiran Lin Anna, Gerchanovsky, Omer Akgul, Lujo Bauer, Matt Fredrikson, Zifa Wang, LLM Whisperer: An Inconspicuous Attack to Bias LLM Responses
- [16] Rohin Manvi, Samar Khanna, Marshall Burke, David Lobell ,Stefano Ermon, Large Language Models are Geographically Biased
- [17] CNN , Forget bias, the real danger is Big Tech’s overwhelming control over speech
- [18] Shaleen Khanal, Hongzhou Zhang, Araz Taeihagh, Why and how is the power of Big Tech increasing in the policy process? The case of generative AI
- [19] Khanisyah E. Gumilar, Birama R. Indraprasta, et al. , Disparities in medical recommendations from AI-based chatbots across different countries/regions
- [20] Khanisyah E. Gumilar, Birama R. Indraprasta, et al., Disparities in medical recommendations from AI-based chatbots across different countries/regions - supplement 2
- [21] Khanisyah E. Gumilar, Birama R. Indraprasta, et al. ,Disparities in medical recommendations from AI-based chatbots across different countries/regions - supplement 3
- [22] ChainLink , What Is a Zero-Knowledge Proof?
- [23] Brookings - Nicol Turner Lee, Nicol Turner Lee, Paul Resnick, Genie Barton, Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms
- [24] Google Research, Fairness Indicators: Scalable Infrastructure for Fair ML Systems
- [25] Fiddler AI, Fiddler Trust Service for LLM Application Scoring and Monitoring
- [26] Fiddler AI, The Ultimate Guide to LLM Monitoring
- [27] WhyLabs AI, What is AI observability?
- [28] Andres Carranza, Dhruv Pai, Rylan Schaeffer, Arnub Tandon, Sanmi Koyejo, Deceptive Alignment Monitoring
- [29] LumeNova AI, All your AI Risk Management needs. One platform
- [30] European Union, EU AI Act: first regulation on artificial intelligence
- [31] LumeNova AI, Colorado Senate Bill 24-205: Algorithmic Discrimination Protection
- [32] Colorado General Assembly, Consumer Protections for Artificial Intelligence

- [33] Aligned AI, Make your AI models self-correcting
- [34] Matija Franklin, Hal Ashton, Rebecca Gorman, Stuart Armstrong, Missing Mechanisms of Manipulation in the EU AI Act
- [35] Hazal Şimşek, Compare Top 25 AI Governance Tools: A Vendor Benchmark
- [36] Burak Ceylan, Large Language Model Evaluation: 5 Methods
- [37] Holistic AI, What is the NIST AI Risk Management Framework?
- [38] The National Institute of Standards and Technology, AI Risk Management Framework
- [39] Holistic AI, The future of AI regulation in the EU
- [40] EU, The EU AI Act
- [41] Holistic AI, Meet the requirements of the EU's Digital Services Act
- [42] EU, The Digital Services Act
- [43] Holistic AI, Streamline compliance with Holistic AI
- [44] NYC, Automated Employment Decision Tools
- [45] Holistic AI, Keep up with insureTech AI regulation
- [46] Colorado, SB21-169 - Protecting Consumers from Unfair Discrimination in Insurance Practices
- [47] Holistic AI, Navigate Colorado SB205: Ensuring Responsible AI Practices
- [48] Colorado, Consumer Protections for Artificial Intelligence
- [49] Holistic AI, Simplify ISO/IEC 42001 Governance Standard Certification with Our AI Solutions
- [50] ISO, ISO/IEC 42001:2023 Information technology — Artificial intelligence — Management system
- [51] Stanford University, MIT, The Foundation Model Transparency Index v1.1 May 2024
- [52] McKinsey Global Institute, Tackling bias in artificial intelligence (and in humans)
- [53] Aaron Gordon, The Multiple Faces of Claude AI: Different Answers, Same Model
- [54] Ashok Sharma, How AI is Affecting the VPN Industry
- [55] Private Internet Access - Glyn Moody, Advanced AI Chatbots Can Now Infer Detailed Personal Attributes from General Social Media Posts
- [56] Google, Teach LLMs to Personalize – An Approach inspired by Writing Education
- [57] Chiyu Zhang, et al, SPAR: Personalized Content-Based Recommendation via Long Engagement Attention
- [58] Valentin Hofmann, Pratyusha Ria Kalluri, Dan Jurafsky, Sharese King, AI generates covertly racist decisions about people based on their dialect
- [59] Sean Falconer, Privacy in the age of generative AI
- [60] George Mason University, Geographic and Geopolitical Biases of Language Models
- [61] Google's Woke AI Wasn't a Mistake. We Know. We Were There.
- [62] Deceptive AI != Deceptively-aligned AI
- [63] Deceptive AI Gets Busted And Stopped Cold Via OpenAI's O1 Model Emerging Capabilities
- [64] Understanding strategic deception and deceptive alignment
- [65] Abraham Lincoln, Gettysburg Address