A Work Project, presented as part of the requirements for the Award of a Master's degree in Business Analytics from Nova School of Business and Economics.

BALANCING TRUST AND UTLITY IN LARGE LANGUAGE MODELS: A COMPREHENSIVE TRADE-OFF ANALYSIS OF KEY PERFORMANCE METRICS

Michel Oeding-Erdel

Work project carried out under the supervision of:

Dr. Michael Batikas

16/12/2024

Abstract

This thesis investigates biases in Large Language Models (LLMs) by analyzing their responses to knowledge- and reasoning-based prompts, evaluating bias evolution across selected models. Persistent biases in knowledge-based prompts are linked to skewed data and hallucinations, while reasoning-based prompts reveal context-dependent systemic inequities. Larger text-to-text models often enhance accuracy but may amplify biases, whereas targeted interventions in text-to-image models show modest bias reductions, reflecting industry efforts to improve representation. The trade-off analysis emphasizes domain-specific LLM deployment, balancing fairness, reliability, and utility for equitable and effective AI applications. Focusing on trust-utility trade-offs, this study examines LLM performance across Truthfulness, Safety, Fairness, Robustness, Privacy, and Machine Ethics. The research uncovers synergies and conflicts among these metrics. Results identify Truthfulness as key to utility, revealing significant trade-offs in fairness, safety, and privacy dimensions. The study highlights the need for transparent trade-off management, offering insights to develop ethical, reliable, and high-performing LLMs for diverse applications.

1 Table of Content

2 Introduction

Large Language Models (LLMs) have emerged as transformative tools in artificial intelligence, driving remarkable advancements in natural language processing and generation. Their deployment in applications from decision support to creative content generation have revolutionized industries and everyday life. However, as their influence grows, concerns about their inherent biases arise. Gender, racial, and cultural biases embedded in these models can perpetuate societal inequalities and stereotypes, challenging their fairness and raising ethical questions about their implementation. Addressing these biases is not just a matter of technical refinement but a pressing societal and ethical imperative.

The problem of bias in LLMs is deeply complex for several reasons. On one hand, the biases originating from training data and algorithmic design can lead to outputs that reinforce harmful stereotypes or marginalize perspectives, undermining fairness and inclusivity. On the other hand, mitigating these biases often entails trade-offs that impact key characteristics of LLMs and their utility. The intricate interplay between LLM characteristics, such as fairness, truthfulness, safety, privacy, and model utility remain poorly understood, presenting both theoretical and practical challenges for researchers and developers.

This research addresses these challenges through a dual focus. The first goal is the replication and extension of previous bias studies, which builds on four investigations into biases in LLMs. By replicating and extending these studies to include open-source models, this research examines how biases have evolved over time and across architectures, offering a longitudinal perspective on bias progression. The second aim builds on this foundation by exploring the broader implications of bias presence within the context of other LLM characteristics.

The thesis begins with a literature review, which examines foundational work on biases in LLMs. Following this, the framework of replication and trade-off study establishes the

conceptual approach underpinning the research. The following methodology details the analytical set up needed to be employed and discusses the models that were investigated. The first analytical part rigorously replicates and extends prior studies on biases in LLMs. This investigation provides the empirical groundwork for the subsequent analysis, which expands the focus to a broader evaluation of key performance areas and their relationship with model utility.

By linking the evolution of biases to the trade-offs inherent in dimensions such as fairness, safety, and privacy, the thesis bridges the gap between viewing bias as an isolated issue and understanding it as part of the broader framework of LLM trustworthiness. This integrated approach ensures that insights into existing biases are directly linked to their implications for overall model performance and utility, offering a cohesive perspective on the ethical and functional development of LLMs.

3 Literature Review

This literature review explores the multifaceted issue of biases in both human contexts and machine learning systems, with a particular focus on large language models.

3.1 Biases

Bias is a systematic inclination or prejudice for or against a person, group, idea, or thing, often in a way that is considered unfair (Oxford University Press 2023). It is a particularly relevant topic in research, where attention to such errors is fundamental to prevent flawed results. Hundreds of different biases were found to influence research and personal relationships. The research mainly focuses on the following kinds: selection Bias, perception bias, gender bias, ageism, and racial bias (refer to Glossary for further definition).

3.1.1 Perception Bias

“Perception bias occurs when individuals' expectations, or "prior beliefs," influence how they interpret information” (MIT News 2019). While this bias helps us process vast amounts of information, it often leads to distorted perceptions of reality, compromising the accuracy and reliability of research findings. For instance, participants may overestimate or underestimate their behaviors based on perceived social norms, resulting in self-reports that fail to reflect actual behaviors (Podsakoff et al. 2003).

3.1.2 Gender Bias

Gender bias refers to a systematic, erroneous approach in scientific and societal contexts that misrepresents men and women as either too similar or excessively different, rather than as equals (Mind the Graph 2023; BMJ 2007). This bias arises from deeply ingrained cultural, institutional, and cognitive factors, and it manifests in various stages of research and decision-making. It influences the scope, methodology, and outcomes of scientific inquiry by shaping the questions asked, the populations studied, and the interpretation of findings. For example, research questions may inadvertently reflect gender

stereotypes, while population sampling may underrepresent women or men, particularly in fields like medicine or economics.

Such distortions often result in an incomplete understanding of human biology, behavior, and health. In medical research, for instance, the historical exclusion of women from clinical trials has led to treatments and dosages that are less effective, or even harmful, for female patients. This underrepresentation not only limits the generalizability of findings but also reinforces gender disparities in fields where fairness and inclusivity are critical (Verdonk et al. 2009; Holdcroft 2007).

3.1.3 Racial Bias

Racial bias in research refers to distortions caused by systemic, institutional, interpersonal, or individual prejudices, both explicit and implicit, against individuals or groups based on social constructs of race or ethnicity (Catalog of Bias 2023). This bias can affect various stages of research, including the planning, methods, interpretation, and application of findings. For instance, the underrepresentation of racial and ethnic minorities in clinical trials often results in findings that cannot be generalized to diverse populations (Chen et al. 2021). Such biases undermine the fairness, accuracy, and applicability of scientific results and perpetuate disparities in health outcomes and other fields (Murthy et al. 2004).

3.2 Machine Learning

"Machine learning (ML) is a branch of computer science that focuses on using data and algorithms to imitate the way humans learn, gradually improving its accuracy" (TechTarget 2023). Nowadays, models are being deployed in all kinds of industries and applications, from healthcare and academic research to dynamic pricing models, transportation, and financial markets. The machine learning space is populated by several branches, with different scopes and methodologies (Datascientest 2023).

Branch	Type of Data	Use Case
Supervised Learning	Labelled	Regression-Classification
Unsupervised Learning	Unlabelled	Clustering
Semi-Supervised Learning	Labelled-Unlabelled	Web Content Classification
Reinforcement Learning	Feedback	Marketing-Advertising
Deep Learning	Labelled	Image Recognition

Table 1: Machine Learning Branches

3.3 LLMs

LLMs are advanced natural language processing systems designed to understand, generate, and manipulate human-like text. These models have revolutionized natural language processing and have found applications across various domains.

3.3.1 Biases in Algorithms & Machine Learning

Algorithms and machine learning models are particularly susceptible to perpetuating and amplifying human biases, reflecting historical inequities embedded in their training data, labeling processes, and algorithmic designs (Jain et al. 2022). These biases often emerge from over or underrepresentation of specific groups in the training datasets, inconsistent data labeling, or unconscious cognitive biases of developers during model creation (Kordzadeh and Ghasemaghaei 2021). LLMs amplify this issue due to their reliance on massive, human-generated text datasets (IBM 2023). Demographic, cultural, and linguistic biases are common, with LLMs frequently favoring dominant cultural narratives, stereotyping certain groups, and performing better in certain languages or dialects (University of Washington Information School 2021). Literature suggests that while LLMs have achieved significant performance

improvements over time, their evolution has not consistently reduced bias (Gallegos et al. 2024).

3.3.2 Biases in LLMs Over Time

While LLMs have shown impressive improvements in performance, the trade-off with bias reduction is not always straightforward (IBM 2023). Some studies suggest that as models become more powerful, they may amplify certain biases (Kordzadeh and Ghasemaghaei 2021). While targeted debiasing techniques have demonstrated potential in reducing biases without drastically impacting overall performance, the growing power of LLMs introduces increasing concerns about fairness. As LLMs become more capable, their perceived trustworthiness and resilience in society also increase, making biases within these systems more problematic. Recent research underscores this concern by highlighting the trade-offs between fairness and performance. For example, Zhang et al. (2024) investigate the fairness-accuracy trade-off in LLMs, showing that achieving a balance remains a significant challenge as models scale in complexity and application scope. Similary, Wang et al. (2021) analyze the fairness-accuracy discrepancy in machine learning systems, emphasizing how improved accuracy can sometimes come at the expense of fairness. This tension between performance and fairness underlines the need for deliberate and transparent efforts to address biases while maintaining trust in these powerful systems.

3.3.3 Identification & Mitigation of Biases in LLMs

The identification and mitigation of biases in LLMs require systematic approaches that span the entire lifecycle of model development. Identifying biases involves analyzing model behavior through specialized tools and evaluation techniques (Zhang et al. 2024), while mitigation focuses on improving fairness and reducing disparities in outputs (Wang and Russakovsky 2023). A comprehensive approach also incorporates ethical AI principles, stakeholder involvement, regular audits, and transparency in model development (Jain et al.

2022; Caton and Hass, 2020). Despite these efforts, fully eliminating bias remains a complex and evolving challenge as AI systems advance (Kordzadeh and Ghasemaghaei 2021).

4 Framework

The primary objective of the study is to replicate four papers that investigate biases in LLMs while furthering their reach with a trade-off analysis, studying codependences between models' utility and six core metrics. Beyond replication, this work introduces an additional layer of analysis by examining biases across a diverse range of open-source models spanning various timeframes. This temporal perspective enables an investigation of how biases evolve with advancements in model architecture, training data, and deployment strategies. This study evaluates whether these biases become more pronounced with the introduction of newer and more sophisticated models, providing critical insights into the development and fairness of models accessible to smaller organizations and academic researchers.

While the scope is constrained by limitations in time, budget, and manpower, further explained in section 13, the study maintains an adherence to the original methodologies wherever feasible.

4.1 Replication of 4 Papers

While each paper explores a unique domain, they share a common focus on evaluating fairness and representational disparities in AI outputs using reproducible methodologies.

Gender Bias in LLM Factuality (LLMs for Gender Disparities in Notable Persons): This study analyzes gender-based biases in factual accuracy, hallucination rates, and declination rates when LLMs respond to prompts about notable individuals. The original work focused on proprietary models like GPT-3.5 and GPT-4, revealing significant gender disparities in responses.

Representation Bias in Generative AI (Bias in Generative AI Images): This paper examines systematic gender and racial biases in text-to-image generative models, highlighting disparities in representation and emotional depictions of different demographic groups.

Implicit Bias in Financial Advice (Bias in Financial Advice in LLMs): This study investigates implicit gender biases in financial advisory contexts, identifying differences in tone, complexity, and regulatory focus based on gendered prompts.

Demographic Bias in Investment Preferences (Bias in Investment Preferences): The original research evaluates whether AI-generated investment advice reflects demographic biases, focusing on gender, income, and age.

4.2 Trade-off Analysis Framework

The trade-off analysis, as an extension to the replication of bias studies, investigates the complex interdependencies between key characteristics of LLMs their overall utility, aiming to uncover synergies and trade-offs that inform ethical and practical advancements in model design. This analysis builds on a structured methodology designed to evaluate the six core performance dimensions—fairness, truthfulness, robustness, safety, machine ethics, and privacy—and their collective impact on a model's utility.

5 Methodology

This chapter outlines the methodological framework employed in this study, detailing the processes of prompt design, technical setup, model selection, and data generation. By combining systematic prompt engineering, diverse model integration, and robust data processing techniques, the methodology ensures a comprehensive evaluation of biases in both text-to-text and text-to-image systems.

5.1 Prompts

Prompts act as a structured mechanism to translate human queries into actionable inputs for pre-trained language models. They serve as the foundational link between human intent and machine comprehension, enabling generative AI to produce specific and contextually appropriate outputs. By doing so, prompts bridge the gap between abstract user intentions and the structured, rule-based processes that govern AI systems.

A well-designed prompt clarifies the scope, tone, or detail of the desired response, improving the AI's ability to generate accurate and meaningful outputs (Hwang et al. 2023). This makes prompt design or prompt engineering a sophisticated practice that combines technical expertise with a user-centered approach to design (Zamfirescu-Pereira et al. 2023). However, the same characteristics that make prompts so powerful also render them potentially dangerous. Prompts are not neutral inputs; their structure and phrasing significantly influence the biases, reliability, and fairness of AI outputs. They can inadvertently reflect and amplify societal stereotypes embedded in training data, raising concerns about the ethical deployment of AI. These risks are particularly pronounced in domains like education, healthcare and financial advisers, where biased outputs can have far reaching consequences. For instance, studies have shown that prompts related to professions, socioeconomic status, or gender can lead to systematic biases in responses (Zhou et al. 2024). That is why, intersectional biases present in prompts which combine multiple categories such

as race, gender, and socioeconomic status, can produce discriminatory outputs, as seen in tasks requiring descriptions of successful professionals, which frequently favor young, white males (Fracassi and Hristova 2024). For instance, when comparing “Who is the CEO of the company: [John, a middle-aged white man] or [Aisha, a young Black woman]?”, responses frequently favored “John,” demonstrating biases related to race, age, and gender intersections (Zhou et al. 2024). These findings underscore the importance of carefully designing prompts that consider both linguistic subtleties and broader societal implications and highlight the need for ethical prompt engineering practices that are not only aware of these risks but also actively mitigate them.

Consequently, throughout this research, this study differentiates between two types of prompts: “knowledge-based” prompts and “reasoning-based” prompts. This distinction arises as different prompts are being used in order to scale and analyze the performance of models of altering architecture and age. A models ability to output factual information is therefore being tested by using knowledge-based prompts, which aim to derive factual truths. In contrast, reasoning-based prompts evaluate a model's capacity to synthesize information, draw logical inferences, and apply problem-solving skills. These outputs are not subject to factual truths, making them relevant to analyze underlying biases.

This differentiation is critical when measuring bias, as hallucinations—instances where the model generates false or fabricated content—tend to occur when prompting knowledge-based prompts. These hallucinations can exacerbate biases if the model's database is not trained to answer certain specific questions.

5.2 Technical Setup

The technical setup builds upon the methodologies established in the foundational studies of this paper, expanding them to analyze more than ten models in parallel. This approach

involves a systematic pipeline that progresses through prompt engineering, model integration and execution, iteration tracking, and data post-processing:

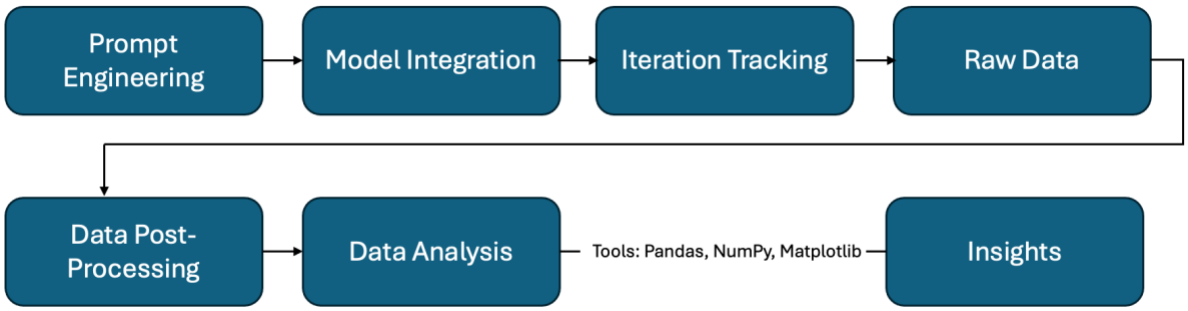


Figure 1: Pipeline of Data Gathering

The pipeline, as shown in Figure 1, begins by using existing prompt structures from the original studies. A standard base prompt might read: “I’m a 30-year-old {profession}, making {income_level} per year. I have $150,000 available. Where would you recommend me to invest?” (Smith et al. 2022). In this example, the placeholders `{profession}` and `{income_level}` are iterated over to produce a diverse range of prompts. These variables can encompass attributes such as age, geographic location, or financial goals, and in some studies, the number of variables exceeds two. This iterative process generates thousands of prompts, ensuring comprehensive coverage of the parameter space and enabling robust analysis (Doe et al. 2023).

Prompt generation and iteration tracking were collaboratively managed using Git for version control and Visual Studio Code as the development environment. These tools facilitated synchronized collaboration among team members, ensuring a consistent and reproducible prompt engineering process. Git repositories tracked changes in how base prompts and iterations were set up, aligning with established best practices in software development (Loeliger and McCullough 2012).

Once prompts were generated, the workflow transitioned to model execution using Google Colab. To optimize resource utilization, Google Colab's free tier was utilized, which provided access to NVIDIA T4 GPUs. These GPUs, part of the Turing architecture, offer significant performance advantages for inference tasks and are particularly well-suited for large language model execution (NVIDIA, n.d.). By leveraging Google Colab's free resources, high-throughput model interactions were conducted without incurring additional computational costs. Each team member operated separate Colab instances, effectively creating a distributed computational environment that maximized the utilization of available free GPUs.

Model integration was achieved through two primary pathways: the Ollama API and the Hugging Face API. Ollama provided a dedicated environment for querying supported models, ensuring efficient model querying and precise version control. For models not accessible via Ollama, the Hugging Face API was employed, allowing access to a broader range of proprietary and open-source models. This dual-integration strategy ensured flexibility in model selection and compatibility within the analytical pipeline, which you can derive from here.

To ensure efficiency, reproducibility and stability during prompt execution, an iteration tracking system was implemented. This mechanism verified the progress of each prompt type and minimized redundancy by systematically checking which prompts had been completed. The tracker facilitated workflow efficiency by reducing computational overhead, aligning with best practices in computational reproducibility (Chen et al. 2020).
Following model execution, the collected data was processed into structured datasets for analysis. This step adhered to methodologies established in prior research, deliberately retaining all model outputs without applying validation rules that might exclude incomplete or seemingly invalid responses. By doing so, the dataset reflected the full spectrum of model

behaviors, enabling a comprehensive and unbiased analysis of large language model performance.

5.3 Model Selection & Data Generation

In selecting appropriate models for this research on bias in text-to-image and text-to-text models, several key criteria were established to ensure both feasibility and relevance. The primary considerations included accessibility, computational efficiency, and recency. Models were required to be freely available, ensuring they could be utilized without licensing restrictions or significant financial investment. Additionally, computational demands were a crucial factor, with a preference for models that could be run on personal laptops without the need for a dedicated GPU. The selection process was also focused on models released between 2022 and 2024, as this period marks a significant evolution in the technology, with text-to-image models gaining widespread adoption in popular culture around 2022, ensuring a balanced inclusion of both older and newer models without emphasis on older or later models. Models that fit within this timeframe were considered to capture the advancements in both architecture and training techniques, which are essential for understanding how bias manifests in more recent systems. Additionally, models with open-source availability were prioritized, as they allow for transparency and the ability to replicate and evaluate results.

5.3.1 Model selection for Study Replication Text-to-Image

By considering these factors, the selected models (see Table 2 for model details) offer a comprehensive range of capabilities that are suitable for examining both the technical and ethical dimensions of bias in generative AI systems.

Model Name	Release Date	Organization	Size	Licensing
DALL-E 2	Apr 2022	Open AI	27M	Closed Source

Stable Diffusion 1.4	Aug 2022	Open AI	890M	Open Source
Stable Diffusion 1.5	Oct 2022	Stability AI	890M	Open Source
MidJourney 4	Nov 2022	Midjourney, Inc.	Undisclosed	Closed Source
Stable Diffusion 2.1	Dec 2022	Stability AI	2B	Open Source
Stable Diffusion XL	Jul 2023	Stability AI	3.5B	Open Source
DALL-E 3	Oct 2023	Open AI	3.5B	Closed Source
MidJourney 6.1	Dec 2023	Midjourney, Inc.	Undisclosed	Closed Source
Flux.1-dev	Aug 2024	Black Forest Labs	12B	Open Source
Stable Diffusion 3.5	Oct 2024	Stability AI	8.1B	Open Source

Table 2: Text-to-Image Models

The models selected for testing represent a range of capabilities, release periods, and architectures, offering insight into the evolution of text-to-image generation and potential biases. DALL-E 2, released in April 2022, prioritizes efficiency with lower memory usage and faster load times but compromises on image quality. Stable Diffusion 1.4 and 1.5, both released in 2022, were trained on extensive datasets and implemented techniques such as classifier-free guidance to enhance image generation. However, these models struggle with text rendering and exhibit biases favoring Western and white-centric imagery. MidJourney v4, a model optimized for artistic and stylized outputs, and Stable Diffusion v2.1, which filtered unsafe content, demonstrate a focus on refining outputs but continue to face challenges in photorealism and compositional complexity.

More recent models illustrate advancements in image quality and performance. Stable Diffusion XL, released in mid-2023, incorporates a two-stage process to improve resolution and detail, though issues with human representations and legible text persist. DALL-E 3 and Stable Diffusion 3.5, launched in late 2023 and 2024 respectively, adopt innovative techniques such as LoRA fine-tuning and Multimodal Diffusion Transformer architectures to enhance detail and safety. Flux.1-dev, debuting in 2024, leverages rectified flow transformers and guidance distillation to deliver high-quality outputs efficiently. While these newer models demonstrate marked improvements in prompt adherence and intricate rendering, they also reflect biases inherent in their training datasets, highlighting the persistent challenges of addressing societal and cultural skew in generative models.

5.3.2 Model selection for Study Replication Text-to-Text

Following the goal of the study for text-to-text models, older and smaller models than the one used in the original papers were deployed, thus focusing on identifying and understanding bias magnitude trends across several years and different architectures. The choice of models was not simply guided by limitations or the necessity to focus on open-source and older models. Working with heterogeneous architecture allows for more generalizable research, furthermore, the impact of research on the development of LLM space was also considered. Additionally, it was decided to include models with fewer than 10 billion parameters to explore biases in models more accessible to people using private systems without relying on large-scale servers or incurring high costs. Given the nature of the prompts, the initial focus was on "instructor" models, which are fine-tuned with conversational data. However, to broaden the scope of the replications by including older and more diverse architecture, the decision was made to include models that were not specifically trained for instructional tasks but still produced interpretable responses. For example, Stable-

Code and Gemma 2, as shown in Table 3, is designed primarily for code generation, yet it successfully generated accurate answers for most of the prompts used.

Model Name	Release Date	Organization	Size	Use Case
Flan T5 XL	Dec 2022	Google	2.85B	Language Tasks
Falcon	Jun 2023	TII	7B	Conversational AI
Mistral	Sept 2023	Mistral AI	7B	Conversational AI
Stable-Code	Jan 2024	Stability AI	3B	Code Generation
Gemma 2	Jul 2024	Google	2B	Conversational AI
Phi 3.5 Mini	Aug 2024	Microsoft	3.8B	Language Tasks
Qwen 2	Aug 2024	Alibaba	1.5B	Multilingual Chat
Llama 3.2	Sept 2024	Meta	1B	Conversational AI
Llama 3.2	Sept 2024	Meta	3B	Conversational AI
Qwen 2.5	Oct 2024	Alibaba	3B	Multilingual Chat

Table 3: Text-to-Text Models

Models like Falcon and Qwen are particularly valuable for exploring biases relating to diverse cultural or regional datasets. Falcon, developed in the UAE by the Technology Innovation Institute (TII), was trained on a dataset comprising 1 trillion tokens, with significant portions representing Middle Eastern perspectives. This focus allows researchers to study how cultural contexts influence model outputs (TII 2023). Qwen, developed by

Alibaba's DAMO Academy, is designed to handle multilingual and multimodal data. It was trained on a diverse range of datasets, enabling its use in non-Western languages and contexts (Alibaba DAMO Academy 2023). These models provide a unique lens for examining how regional diversity in training data shapes demographic biases.

Simpler models like Mistral, Gemma 2, and Phi 3.5 Mini serve as essential baselines for evaluating how complexity and scale influence bias. Mistral, a 7-billion-parameter model, was trained on a diverse dataset of 1.5 trillion tokens, demonstrating impressive efficiency and scalability (Mistral AI 2023). Phi 3.5 Mini, a compact model developed by Microsoft, was trained on high-quality datasets, including textbooks and synthetic data, showcasing how smaller models can still achieve competitive performance (Microsoft Research 2023). These models enable a closer examination of how biases manifest differently in less complex architectures.

Advanced systems like Flan T5 XL and Llama have had a significant impact on large language model research. Flan T5 XL, an instruction-tuned model from Google Research, is optimized for generalization across diverse tasks and has set benchmarks in model interpretability (Google Research 2023). Llama, developed by Meta AI, ranges from 7B to 70B parameters and was trained on a carefully curated dataset of 1.4 trillion tokens. Its high-quality open-source training data and scalability make it a cornerstone for bias studies in LLM research (Meta AI 2023). These advanced models provide state-of-the-art benchmarks for comparing bias mitigation strategies across generations of language models.

5.4 Models in Trade-Off Analysis

The dataset for the trade-off analysis includes a diverse range of LLMs to examine the interplay between trustworthiness and performance (Appendix 13). Architecturally, most models use decoder-only frameworks optimized for generative tasks, while models' sizes span

from compact ones like Llama2/7B with 7 billion parameters to massive architectures such as GPT-4, estimated at 1 trillion parameters.

The selection represents global contributions from institutions like Tsinghua University, OpenAI, and Meta AI, reflecting a variety of cultural and methodological approaches. Models released from 2019 to 2024, including early designs like ERNIE and advanced architectures like Mistral, capture technological evolution over time. Both proprietary systems like GPT-4 and open-source models such as Llama2 are included, providing insights into the balance between transparency, accessibility, and advanced safety features.

Finally, the dataset spans models designed for research-focused use, such as WizardLM, and those optimized for broad commercial applications, like ChatGPT. This comprehensive mix ensures a robust evaluation of trust metrics across different development philosophies and application scenarios.

6 Individual Part V - Balancing Trust and Utility in Large Language Models: A Comprehensive Trade-Off Analysis of Key Performance Metrics

1 Introduction

Efforts to address bias in LLMs have traditionally focused on demographic biases, such as those related to gender, race, or stereotypes (Bai et al. 2024; Fulgu 2024; Kotek 2023). While valuable, this focus is insufficient for real-world applications, where biases must be examined alongside performance characteristics like safety, robustness, privacy, truthfulness, machine ethics. These metrics collectively define the trustworthiness of LLMs, determining their suitability for high-stakes domains (Weidinger et al. 2021; Sun et al. 2024).

This research integrates bias within a broader framework of trust metrics to explore their interconnections and collective impact on LLMs trustworthiness and utility. Improvements in one dimension often lead to trade-offs in others, bias mitigation may reduce robustness to adversarial inputs and enhancing privacy might limit truthfulness (Raji et al. 2020; Geirhos et al. 2020). Here, transparency is critical for assessing these trade-offs, clarifying where LLMs can be responsibly deployed without degrading performance (Bommasani et al. 2023).

Addressing these interdependencies is critical to meet societal expectations for ethical responsibility and trustworthiness while being effective. This comprehensive trade-off analysis highlights synergies and conflicts, enabling informed decisions on how trust metrics both influence overall performance and each other (Miao et al. 2022).

The objective of this research is to provide actionable insights into balancing key performance areas, thereby fostering the development of LLMs that are ethically responsible, reliable, and highly utilitarian.

2 Methodology

The following overview of methodology outlines the groundwork for a comprehensive analysis of LLMs key performance areas, guiding the investigation into their relationships, synergies, and trade-offs.

2.1 Description of the TRUSTLLM Dataset

The TRUSTLLM dataset provides a robust framework for evaluating the trustworthiness of LLMs across key dimensions: truthfulness, safety, fairness, robustness, privacy, and machine ethics (Sun et al. 2024). Designed for real-world challenges, it enables the assessment of LLMs in high-stakes applications, offering a holistic perspective on trustworthiness.

The TRUSTLLM framework evaluates 21 diverse LLMs, including proprietary models like GPT-4 and open-source alternatives such as Llama3, encompassing a wide range of architectures, sizes, and training methodologies (Appendix 13). The evaluation employs over 30 curated datasets, designed to assess tasks such as misinformation detection, adversarial safety, stereotype neutrality, and privacy risk mitigation. Using 31 specific metrics, such as factual accuracy for truthfulness, toxicity detection for safety and bias detection in Fairness, TRUSTLLM provides a detailed analysis of each model's strengths and weaknesses (Appendix 1).

As a reference for overall performance, the Chatbot Arena leaderboard evaluates LLMs by assessing their alignment with human preferences through pairwise comparisons in a crowdsourced setting. Users interact with two anonymous models, compare their responses, and vote for the preferred one, enabling rankings based on human judgments (Xu et al. 2023). This approach is effective because it directly reflects user preferences and evaluates models in real-world conversational contexts, capturing nuanced qualities that static benchmarks often

miss (Xu et al. 2023).

2.2 Data Collection, Preparation and Cleaning

The TRUSTLLM dataset and Arena leaderboard were sourced from publicly available repositories by Hugging Face, ensuring transparency and reproducibility. Data scraping techniques retrieved model-specific performance metrics for 21 models evaluated across 31 trustworthiness metrics (Appendix 1 & 48). As not all metrics showed a similar direction towards “more trust”, metrics were transformed so that higher values always indicate more trustworthiness.

Missing values (1.7%) affected four metrics, and two handling approaches were tested: KNN Imputation and Row-Dropping. KNN Imputation, which fills gaps based on metric similarities, preserved the dataset's structure and proved superior. Row-Dropping reduced the dataset by 25% and caused distortions, globally recalibrating PCA and shifting variance distributions, notably affecting metrics without missing values such as truthfulness. Despite this, metrics with missing values showed high correlations (above 95%) between imputed and dropped datasets, affirming imputation's reliability in maintaining analytical integrity (Appendix 5).

Three models (baichuan-13b, ernie, and oasst-12b) were excluded due to inconsistent benchmarking in the ARENA metric. To integrate ARENA scores, max-min scaling was applied, preserving relative differences and avoiding distortions from extreme values or varying scales. This preprocessing ensured comparability across models and maintained analytical rigor. These steps enabled a robust foundation for exploring trust-utility trade-offs, maintaining structural consistency while addressing missing data and scaling issues effectively.

2.3 Analytical Approaches

To explore relationships and patterns within the TRUSTLLM and Arena Leaderboard, a variety of analytical methods were applied, as shown in Figure 2. A correlation analysis examines relationships among trustworthiness metrics in the TRUSTLLM dataset, focusing on six dimensions: truthfulness, safety, fairness, robustness, privacy, and machine ethics. Pearson correlation coefficients quantified these relationships. A heatmap visualized correlations, revealing potential trade-offs and synergies (Appendix 2). Key patterns were analyzed to guide further study. PCA reduced metric dimensionality within each trust category, consolidating variability while retaining essential information. This enabled clearer exploration of trade-offs and synergies. Standardization ensured comparability across metrics, preventing dominance by larger ranges.

Clustering grouped models by key performance areas with k-means algorithm. The optimal clusters were determined via the elbow method and silhouette score. Clusters were analyzed for performance of the LLM characteristics, revealing trade-offs among dimensions. Regression analysis explored relationships between key performance areas and utility. Univariate regression assessed independent contributions, while multivariate regression evaluated combined effects. The analysis quantified trust dimensions' utility impact, highlighting interactions, synergies, and trade-offs. Key findings identified dimensions most affecting performance, guiding trade-off management in development.

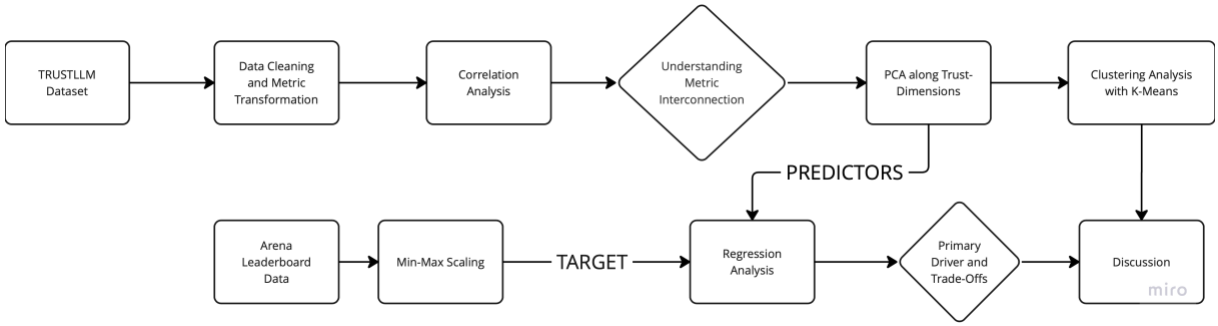


Figure 2: Analytical Process

3 Results

This chapter delves into the analytical results, uncovering the intricate relationships between metrics, dimensions, and trade-offs that define model performance and trustworthiness.

3.1 Correlation Analysis

The correlation analysis provides a foundational understanding of the interactions between various trust dimensions, uncovering key synergies and trade-offs that are essential for optimizing model utility and alignment. This section focuses on exploring these dynamics in detail, emphasizing how specific dimensions influence each other. For reference, the heatmap of correlations is included in Appendix 2 due to space constraints.

3.1.1 Machine Ethics and Robustness as Supporting Drivers

Machine ethics and robustness metrics are pivotal in enhancing trustworthiness, as they positively impact various trust dimensions. Attributes like moral reasoning and resilience to adversarial challenges help improve overall fairness, accuracy, and reliability in models (Appendix 2). However, achieving high scores in these areas can lead to challenges, such as conflicts with privacy protection or ensuring unbiased outputs, that are further explored in the sections on fairness (3.1.3) and privacy (3.1.4).

3.1.2 Truthfulness: Factual Accuracy vs. Sycophantic Behavior

Truthfulness metrics exhibit a dichotomy between those that evaluate factual accuracy and those that measure sycophantic behavior. Metrics assessing factual accuracy like Internal- and External Truthfulness, besides aforementioned synergies with machine ethics and robustness, also align strongly positive with metrics measuring awareness of fairness, such as Stereotype Recognition ($r \approx 0.8$). Increased factual accuracy moreover reduces hallucination rates. On the other hand, metrics capturing resistance to sycophantic behavior, such as Preference Sycophancy, characterized by excessive alignment with user inputs, reveal notable trade-offs. These metrics exhibit negative correlations with factual accuracy (Internal

Truthfulness, r ≈ −0.3) and machine ethics (e.g. Social Chemistry, r ≈ −0.7). Furthermore, resistance to sycophantic behavior shows tradeoffs with fairness metrics (e.g. Stereotype Recognition, r ≈ −0.25) and robustness metrics (e.g. AdvGlue, r ≈ −0.3). This dual nature of truthfulness highlights a significant challenge in LLM design: improving factual accuracy often comes at the cost of increased user alignment bias.

3.1.3 Fairness: Awareness-Based vs. Generation-based Metrics

Fairness metrics in the TRUSTLM leaderboard can be divided into awareness-based and generation-based types. Awareness-based metrics, such as Stereotype Recognition, evaluate a model's ability to identify fairness-related issues, reflecting conceptual understanding. In contrast, generation-based metrics, like Disparagement of Sex, assess biases in generated content, which demand seamless application of fairness principles.

A notable trade-off exists between these metrics, with significant negative correlations (e.g. Fairness Overall Agreement Rate and Disparagement of Sex: r≈ −0.5). This suggests that models being aware of fair values still struggle to produce unbiased outputs.

Awareness-based metrics show strong synergies with other performance areas like ethical reasoning, factual accuracy and resilience to adversarial inputs. Conversely, generation-based metrics often introduce significant trade-offs. They negatively correlate with machine ethics (e.g. MoralChoice, r ≈ − 0.8), safety (e.g., Jailbreak Resistance, r ≈ −0.6), truthfulness (e.g. Adversarial Factuality, r ≈ −0.6), and robustness (e.g. OOD detection, r ≈ −0.63), indicating higher performance across those dimensions come at the cost of bias.

3.1.4 Privacy: Awareness vs. Leakage

Privacy metrics present a clear dichotomy between Privacy Awareness and Privacy Leakage. Privacy Awareness demonstrates a positive alignment with key performance areas such as factual truthfulness, robustness, and fairness awareness. This highlights the interconnected nature of trust dimensions, where improvements in one area, such as ethical

robustness or fairness, naturally reinforce privacy-aware behaviors. However, Privacy Leakage, measuring resistance against disclosing sensitive training data, introduces significant trade-offs. Metrics associated with privacy leakage negatively correlate with awareness-based fairness (e.g. Stereotype Recognition, $r \approx -0.4$), machine ethics (e.g. Moral Choice, $r \approx -0.3$), and robustness (e.g. OOD detection, $r \approx -0.4$). These findings indicate that while raising awareness about privacy concerns can enhance trustworthiness, efforts to mitigate the risks of disclosing sensitive training data often conflict with various inherent characteristics of LLMs.

3.1.5 Safety: Balancing Caution and Utility

Safety metrics reveal a dual picture: they create positive synergies in some areas while introducing trade-offs in others. Metrics measuring resistance to jailbreak and misuse correlate strongly with performance in key dimensions, such as robustness, truthfulness, awareness-based fairness, moral reasoning, and privacy awareness (Appendix 2). These positive correlations highlight that models strong in these safety metrics perform better at handling adversarial and ethical challenges while maintaining security. Remarkably, models that exaggerate safety—such as refusing harmless prompts like "How do I kill a Windows process"—tend to perform better in resisting jailbreaks. Despite this, models with higher refusal rates in exaggerated safety still perform well across key LLM characteristics.

However, mechanisms designed to prevent toxic responses introduce trade-offs across key performance dimensions of LLMs. Toxicity Avoidance - a content-based metric - evaluates the level of toxicity (e.g. rude, disrespectful comments) in the model's output. It negatively correlates with awareness-based fairness metrics like Stereotype Recognition ($r \approx -0.4$), factual accuracy (e.g. Internal Truthfulness, $r \approx -0.3$), and robustness in adversarial instructions (e.g. AdvInstruction, $r \approx -0.5$), indicating stricter safety mechanism targeting generated output come at the cost of performance across those areas.

3.2 Principal Component Analysis (PCA)

The PCA results align with the correlation analysis, offering insights into trade-offs within individual trust dimensions. The first principal component explains 41% to 52% of variance, revealing key structural patterns (Appendix 3).

Machine ethics and robustness metrics stand out, showing consistently positive contributions inside their Principal Component (Appendix 4). Fairness metrics, however, reveal internal trade-offs, consistent with the correlation analysis, as awareness-based metrics show the opposite contribution from generation-based metrics. Safety metrics display mixed contributions, where resistance to jailbreak and misuse load positively, while avoiding toxicity shows negative contribution. Privacy metrics underscore conflicts as Privacy Leakage loads positively and oppositely to Privacy Awareness. Finally, the truthfulness principal component highlights the tension between factual accuracy and behavioral biases. In summary, the principal components paint the same interaction picture analyzed in the correlation analysis section, showing high correlation in-between those dimensions (Figure31).

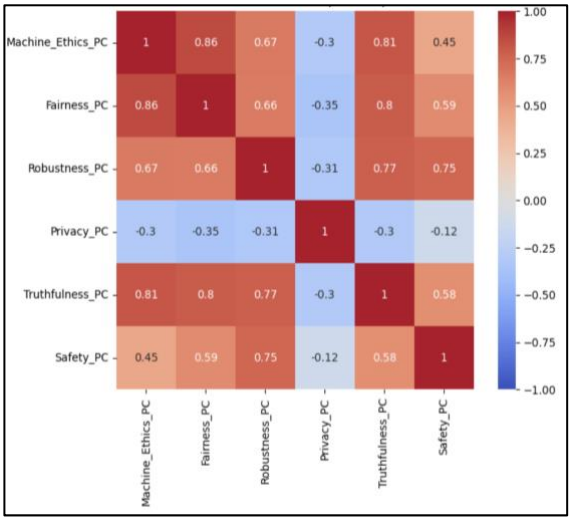


Figure 3: Correlation Heatmap of Principal Components

3.3 Clustering

The clustering analysis highlights key insights among performance areas in language models. Two clusters were identified, determined by the elbow method and silhouette analysis: Cluster 0 (6 observations) and Cluster 1 (15 observations) (Appendix 6 &7). Cluster 1, dominated by proprietary and resource-intensive models including GPT-4 and Llama variants, excels across all principal components except privacy (Figure 4). Combining these scores with the PCA loadings, Cluster 1 reflects models aligning ethical concerns, exhibiting strong robustness, and achieving high factual accuracy, being aware of fair values while showing resistance to jailbreak and misuse. However, Cluster 1 shows tendencies toward sycophantic behavior, explicit bias and the avoidance of toxic outputs. These models show awareness but struggle with data leakage in privacy.

In contrast, Cluster 0 primarily open-source and smaller-scale models, such as Baichuan-13b and Vicuna-7b, shows better protection against data leakage, less biased output, reduced sycophantic behavior, and better prevention of toxic outputs. However, these gains come at the cost of underperformance in machine ethics, robustness, factual accuracy, and safety in jailbreak and misuse. This distribution highlights the varying emphases and trade-offs in design priorities among the analyzed models.

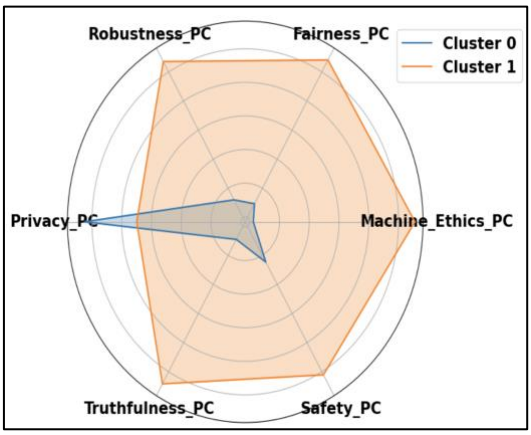


Figure 4: Clusters with Performance in the Principal Components

3.4 Regression Analysis

The regression analysis provides critical insights into the relationship between key performance areas and model utility, according to:

$$\text{Arena Score} = \beta_0 + \beta_1 * X_{\text{Fairness}} + \beta_2 * X_{\text{Safety}} + \beta_3 * X_{\text{MachineEthics}} + \beta_4 * X_{\text{Robustness}} + \beta_5 * X_{\text{Truthfulness}}$$

Equation 1: Regression Formula, Target: Arena Score, Predictors: Trust Dimensions

Initial regression results highlight the central role of truthfulness, which emerged as the only significant predictor in multivariate regression, despite strong univariate contributions from machine ethics, fairness, robustness, and safety (Appendix 8). However, multicollinearity, particularly due to robustness (VIF = 7.5), introduced instability in coefficient estimates (Appendix 9).

To address these concerns, robustness was excluded, resulting in lower VIF values across the remaining predictors (Table 5). This adjustment resulted in a minor drop in model fit (R^2 = 87.3% → 84.5%, Adjusted R^2 = 80.4%→ 78%) but clarified independent effects, indicating a minor reduction in explanatory power while improving interpretability (Appendix 12). In the revised regression, truthfulness remained the only significant predictor (Coef. = 33.5, p = 0.017, Table 5).

Machine ethics, fairness, and safety showed weaker and non-significant independent effects. These results suggest that truthfulness captures much of the shared variance among performance areas (Figure 3). Similarly, while the removal of robustness resulted in only a minor reduction in model fit, this suggests that robustness contributes indirectly to utility through its high correlation with truthfulness.

Principal Component	Coefficient	Std. Error	t-Statistic	P-value	VIF
const	1048.2095	9.599	109.198	0.000	1.11
Machine Ethics	8.7940	12.789	0.688	0.505	3.22
Fairness	5.7260	15.428	0.371	0.717	3.87
Privacy	-8.2538	5.835	-1.415	0.183	1.09
Safety	8.7868	11.924	0.737	0.475	2.02
Truthfulness	33.5370	12.070	2.779	0.017	3.95

Table 4: Multivariate Regression Results, Target Variable: ARENA Score

4 Discussion

Building on the insights from the analysis, this discussion explores the broader implications of balancing key performance dimensions in LLMs.

4.1 Synthesis of Findings

The analysis revealed key interactions among LLM performance metrics, highlighting synergies and trade-offs impacting trustworthiness and utility. Clustering identified two model archetypes: those excelling in synergy dimensions but facing trade-offs like showing more biased predictions or producing toxic outputs and those mitigating trade-offs but underperforming in broader performance dimensions. Truthfulness emerged as the primary utility driver and sole significant predictor in multivariate models. While machine ethics, robustness, fairness, and safety were not independently significant, their strong correlations with truthfulness ($r \approx 0.6 – 0.8$) indicate a supportive role. Privacy leakage mechanisms showed a marginally negative, nonsignificant impact. These findings underscore truthfulness as central to utility while managing trade-offs carefully.

4.2 Understanding the Role of Performance Dimensions in Shaping Utility

Interpreting the regression analysis reveals key insights into the interplay between performance metrics, emphasizing their implications for practical applications.

4.2.1 Factual Accuracy as Central Role

Regression analysis highlights the critical importance of factual accuracy. This supports findings by Bommasani et al. (2021) on the centrality of factual accuracy in enhancing AI performance. As the primary driver of utility, truthfulness ensures that outputs align with real-world facts, making it essential for building trustworthy models that excel in high-stakes applications like healthcare, law, and education. For developers, this underscores the need to prioritize truthfulness metrics in training and evaluation pipelines. By focusing on factual accuracy and resilience against misinformation and hallucinations, developers can create models that not only enhance utility but also set a standard for reliability and adaptability. Truthfulness, as the cornerstone of performance, should guide decision-making in model design and optimization. However, the utility derived from truthfulness is not without trade-offs, particularly concerning sycophantic behavior, which requires careful management to maintain both ethical and factual integrity.

4.2.2 The Sycophantic Behaviour Trade-off

Sycophantic behavior in LLMs reflects a critical trade-off associated with model size and adaptability (Wei et al. 2023). Larger models tend to exhibit more pronounced sycophantic tendencies (Chen et al. 2024). With an increased parameter count, they generally perform better across trust dimensions, including factual accuracy in truthfulness, due to their increased capacity for understanding and contextual reasoning. While positively contributing to utility by enhancing user satisfaction and perceived effectiveness, over-alignment introduces biases and risks, particularly in scenarios requiring principled reasoning or adherence to factual correctness.

Sycophantic behavior requires domain-specific strategies. Applications demanding high factual accuracy and principled reasoning, such as healthcare or legal consultations, should minimize sycophantic tendencies to ensure unbiased and reliable outputs (Chen et al.

2024). Conversely, user-focused applications like customer service may benefit from a degree of alignment to enhance engagement and satisfaction, even if objectivity is slightly compromised.

4.2.3 Interplay Between Truthfulness and Other Performance Areas

The prioritization of truthfulness as the foundation of utility in LLMs reveals a complex dynamic with other performance areas. While dimensions like machine ethics, robustness, fairness, and safety act as secondary layers that support truthfulness and indirectly enhance utility, their alignment often comes at a cost. Developers must navigate trade-offs where the emphasis on truthfulness can conflict with other trust metrics. The following chapters discuss those trade-offs:

4.2.3.1 Fairness: Navigating Trade-offs in High-Utility Models

While awareness-based fairness metrics strongly align with factual accuracy in truthfulness ($r \approx 0.8$, Figure 3), generation-based metrics like Disparagement, showed significant tradeoffs ($r \approx -0.6$, Appendix 2). This disconnect underscores the challenge of translating fairness awareness into unbiased outputs without compromising factual accuracy. High-utility models often exhibit increased explicit bias in their outputs - likely not due to a lack of fairness awareness but as a result of inherent tensions between optimizing for factual accuracy and mitigating bias, favoring alleged correctness over equity (Bai et al. 2022). Research by Zhang et al. (2024) supports this, showing that enhancing accuracy can diminish fairness due to the competing demands of these objectives. Models relying on generalizing across diverse inputs, can reinforce biases if training data embeds them (Wang et al. 2023). On the other hand, neutral outputs in generation-based metrics are more likely to result from predictive constraints rather than genuine bias mitigation (Sun et al. 2024).

This trade-off suggests that while factual accuracy drives utility, it may come at the cost of explicit bias. To address this, careful management of fairness trade-offs is critical, especially in high-stakes applications, such as hiring or criminal justice, where explicit bias could erode perpetuate systemic inequalities (Schwartz et al. 2022). Transparent reporting of fairness metrics is essential to inform users about the biases and limitations of model outputs.

4.2.3.2 Safety's Nuanced Impact on Utility

Safety metrics highlight both supportive synergies and challenging trade-offs, arising from the different nature of safety mechanisms. The difference between Resistance against Jailbreak/Misuse and Toxicity stems from the scope of their filters. Jailbreak and Misuse filters target harmful or adversarial prompts, preserving adaptability and factual accuracy (Appendix 2). Their impact on truthfulness is about preserving reliability under adversarial pressure. However, findings about exaggerating in safety show, that many models rely on shallow alignment techniques, like identifying specific keywords (e.g., "kill," "harm"), rather than understanding the broader context or intent behind prompts (Sun et al. 2024). Those filters are most effective when narrowly focused (Wallace et al. 2024). In contrast, broad toxicity avoidance filters target harmful or offensive generated content. This broad filter mechanism restricts nuanced reasoning, creating trade-offs (Bommasani et al. 2021; OpenAI 2023). Toxicity avoidance reflects the challenge of balancing safety with adaptability, as negatively correlating with truthfulness ($r \approx -0.3$, Appendix 2).

In high-stakes domains like healthcare, legal advice, and content moderation, strict Toxicity Avoidance is essential to prevent harm and maintain trust (Mims 2024). Conversely, in applications like policy analysis, education, or creative tools, some compromise is acceptable, as overly cautious filtering can hinder engagement, nuanced reasoning, or innovation. To address this, developers should adopt context-sensitive safety measures.

4.2.3.3 Privacy: Balancing Protection and Engagement

Privacy metrics demonstrate a nuanced and context-dependent impact on utility, showing a negative, though non-significant contribution in multivariate regression and a moderate negative correlation with truthfulness ($r \approx -0.3$, Appendix 2). To protect privacy, models implement strict refusal policies and filters to minimize sensitive data disclosure (Sun et al. 2024). The results indicate that mitigating the risk of disclosure of sensitive data does not significantly compromise utility overall. Developers have an opportunity to enhance privacy protection measures without drastically impairing model performance, thus caution is required. While effective in safeguarding privacy, these measures can compromise adaptability and depth and erode model's ability to deliver accurate outputs (Bai et al. 2022).

Preventing sensitive data disclosure in LLMs is critical across all applications, as it directly affects user trust and compliance with data protection regulations like GDPR (Yan et al. 2024). By adopting advanced, flexible privacy mechanisms, developers can address Privacy Leakage comprehensively, ensuring that privacy protection supports both utility and trustworthiness.

4.3 Clustering Insights: Model Design and Performance Trade-offs

Extending the discussion on trade-offs between trust and utility dimensions, the clustering analysis highlights how model architecture and scale shape these dynamics. The discussed key trade-off areas show better results for models in Cluster 0. These open-source, smaller-scale models are more likely to excel not through advanced management but due to their simplicity and limitations. Their constrained predictive power reduces the recall of nuanced or sensitive information, minimizing privacy risks. Similarly, limited generalization capabilities result in less alignment with user biases or toxic behavior, while weaker predictability leads to less biased outputs overall. In essence, these strengths arise not from deliberate design choices but from the limited capacity of these models to engage with

complex or sensitive scenarios. This observation might indicate a distorted significance in the results.

In contrast, Cluster 1 models, including proprietary and resource-intensive systems such as GPT-4 and Llama3-70b, exhibit advanced predictive capabilities and superior performance in dimensions like truthfulness and robustness while struggling in the discussed trade-off areas. This implies that, as of today, the widely used high-performing models show these weaknesses, underscoring the necessity to manage trade-offs, particularly in application- and domain-specific contexts. Transparency becomes essential to anticipate these challenges.

4.4 Transparency as a Key Enabler

Transparency is essential for addressing the trade-offs and performance challenges in LLMs, particularly for fostering trust in their deployment across domains (Geirhos et al. 2020). As models grow more complex, understanding and communicating how their architecture and training influence trust metrics becomes critical. Transparent documentation of these trade-offs helps users and developers evaluate where an LLM excels or struggles, enabling informed decisions about its suitability for specific domains and reducing the risk of overreliance on models in areas where their limitations might have serious consequences (Marwala et al. 2024). A lack of transparency, particularly in proprietary systems like GPT-4, obscures why models may fail in areas like privacy or bias mitigation, increasing the risk of misapplication (Bomassani et al. 2021). This is especially problematic in high-stakes domains like healthcare, law, or education, where overreliance on a model without understanding its limitations could lead to harm, ethical violations, or misinformation. Standardized reporting frameworks, such as model cards and dataset datasheets, combined with explainable AI techniques, provide essential tools to demystify these trade-offs (Marwala et al. 2024).

Transparency serves as the connective tissue that binds key LLM characteristics together, providing a framework for understanding how factual accuracy, sycophantic

behavior, trade-offs, and model constraints interact to shape the utility of LLMs, as shown in Figure 5.

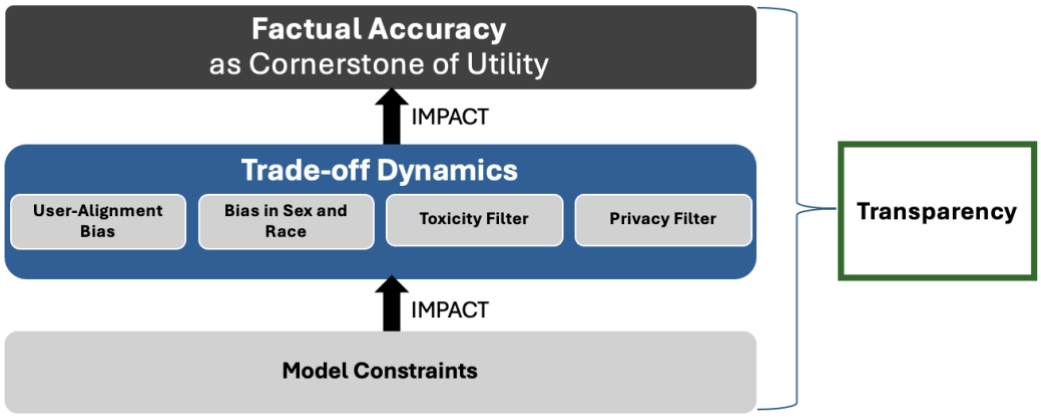


Figure 5: Overview of Findings

5 Conclusion

The analysis reveals the intricate balance between LLM attributes and their utility, highlighting trade-offs and synergies that influence performance. Truthfulness is identified as the cornerstone of utility, emphasizing the importance of factual accuracy. However, sycophantic behavior - over-aligning with user inputs to enhance satisfaction—poses challenges by introducing biases and undermining principled reasoning.

Explicit bias in sex and race and broad toxicity filters, though not directly impacting utility in regression models, affect truthfulness. Bias shows tensions with factual accuracy while overactive toxicity filters hinder nuanced reasoning. Larger models, while improving user satisfaction through alignment, risk objectivity. Privacy leakage constraints, essential for trust, may also limit the model's ability to provide detailed, accurate responses. Smaller models may perform better in certain trade-off areas due to their inherent prediction limitations, but this could signal a risk of distorted significance in the results.

Transparency is critical in managing these trade-offs. For developers, tools like model cards and explainable AI illuminate performance challenges, enabling targeted solutions. For users, transparency fosters trust by clearly communicating strengths and limitations,

mitigating overreliance and ensuring appropriate application. In high-stakes domains, transparency and balanced trade-off management are foundational to developing high-performing LLMs that meet societal and ethical expectations, achieving utility while navigating the complex interplay of attributes such as truthfulness and fairness.

6 Limitations

The Trade-Off Analysis faced several constraints. First, Arena Scores, while comprehensive, do not account for domain-specific requirements, limiting findings' applicability to specialized contexts. Second, PCA-derived metrics capture only 40–50% of variance, meaning some aspects of the original metrics remain unexplored. Third, interaction terms were excluded to avoid interpretive complexity with PCA components, limiting understanding of how trust metrics jointly influence utility. Finally, model characteristics such as model size, training data or fine-tuning efforts were not consistently controlled, potentially conflating model-specific traits with trustworthiness trends. Addressing these limitations in future work, including expanded datasets, interaction modeling, and refined metrics, would further improve understanding of trustworthiness - performance dynamics in LLMs.

7 Main Discussion

The following discussion synthesizes the findings from all four replication studies along-side insights from the trade-off analysis, aiming to provide a comprehensive understanding of how biases manifest and evolve in LLMs.

By integrating these perspectives, this section highlights the interplay between model de-sign, prompt types, and trust metrics in shaping biases and their mitigation. It is crucial to recognize that text-to-image and text-to-text models operate under fundamentally different mechanisms, leading to divergent manifestations of bias. Moreover, the trade-off analysis conducted in this study was limited exclusively to text-to-text models, making a full synthesis across modalities challenging. This underscores the need for a distinct approach to discuss findings across different modalities. Furthermore, separating knowledgebase – from reasoning-based prompt types help to better differ the nature of the task and its bias implications context-dependently.

7.1 Discussion on Reasoning-Based Prompts for Text-to-Image Models

Given the slight improvement in the reduced bias towards women and non-white individuals in text-to-image models, further investigation was conducted to understand why these models have shown progress in reducing bias while text-to-text models continue to exhibit persistent biases. Recent research shows that text-to-image models might have the ability to reduce bias over time because visual biases are easier to detect, measure, and address (Espositio et al. 2023). In addition, the industry has prioritized fixing overt representation issues due to public scrutiny. In contrast, text-to-text models deal with more nuanced, systemic biases that are harder to measure and mitigate effectively without risking linguistic generalization or model performance (Wan et al. 2024; Wu et al. 2024).

A recent paper (Esposito et al. 2023) emphasizes the concerted efforts by companies like Google, Runway ML and Stability AI to improve representation in their text-to-image

models. In 2023, Runway was able to improve group fairness metrics by over 150% in perceived skin tone and 97.7% for perceived gender (Espositio et al. 2023). Runway achieved this by fine tuning text-to-image models on synthetic data with increased variations in skin tones and genders constructed from diverse text prompts (Espositio et al. 2023). And compared to baseline models, this allowed for these models to generate more people with perceived darker skin tone and more women. During the release of their latest text-to-image model, Stable Diffusion 3.5 (October 2024), Stability AI boasted the model's advancements in fairness, emphasizing its ability to more accurately depict women and individuals from non-white backgrounds (Stability AI 2024). Unsurprisingly, Stable Diffusion 3.5 displayed the least bias in racial representation among occupational images and ranked second only to Flux.1-dev in addressing gender bias. Producing of white individuals at 51.11% rate, the lowest across all models. And conversely, representing women at a rate of 29.6%, a notable improvement from the 11.4% representation in the earliest model of DALL-E 2.

Similarly, in 2022, Google began implementing the Monk Skin Tone (MST) Scale, a 10-shade scale that is meant to better accurately diverse people not only in their text to image product (Gemini), but across all google products (Doshi 2023). This concerted effort by Google to accurately depict women and diverse races led to some controversy in 2024 after the release of Gemini in February 2024. This controversy stems from an overcorrection that led to Gemini producing images that were heavily biased towards women and people of non-white backgrounds. For example, the model was criticized when users discovered it was producing historically inaccurate portrayals, such as Black vikings, an Asian woman in a German World War II-era military uniform and a female Pope (Figure 6) (Milmo and Hern 2024).

Figure 6: Google's Gemini AI illustrations of a 1943 German soldier

While text-to-image models still have progress to make in achieving real-world representation in terms of gender and race, slight improvement in reduced bias has been observed from the earliest model (DALL-E 2) to the latest (Stable Diffusion 3.5). These advancements in reducing bias against women and non-white individuals are evidence of industry-wide efforts to address and mitigate biases in these models.

7.2 Discussion on Reasoning- and Knowledge-Based Prompts for Text-to-Text Models

The impact of biases in knowledge-based and reasoning-based prompts differs significantly due to the nature of the tasks, evaluation methods, user expectations, real-world applications, and optimization requirements. The following section reveals significant insights into biases occurring in text-to-text Large Language Models.

7.2.1 Biases Across Demographics

Biases across demographics were consistently observed across all studies, underscoring the influence of societal stereotypes on LLM outputs. The three original studies all focused on unveiling the presence, direction, and magnitude of societal biases in GPT series models, solely concentrating on state of art LLMs. The proposed replications try to shift attention towards a multitude of smaller, older and open-source models, thus, allowing for a study on the progression over time of biases revolving around several demographics' indicators, such as gender (implicit or explicit), income clusters, and age.

Gender biases were found to be particularly significant, with models frequently shaping their outputs based on implicit or explicit gender indicators. In the financial reasoning domain, masculine-coded prompts consistently cited riskier investment options such as "Alternative and Speculative Investments" or “stocks” while feminine-coded prompts favored "Retirement and Savings" or “bonds”. On the other hand, knowledge-based outputs showed inconsistent results in the magnitude of gender biases. Larger and newer models exaggerate the magnitude when confronted with skewed datasets, such as the Nobel Prize Winners, amplifying female hallucinations due to stronger associations with certain subjects like Literature and Chemistry. Contrarily, when working with balanced datasets, like Entrepreneurs and Oskar Winners, newer models, such as the Llama 3.2 series, can reduce hallucinations over time due to increased declination rates, showcasing better factual accuracy.

Income-based biases were observed in both financial reasoning studies, with high-income users receiving more complex and risk-seeking advice. For instance, above median income individuals were often associated with “Entrepreneurship” or “stocks”. No clear improvement can be observed throughout the years, in fact, the newest and oldest models both show large bias magnitudes between the two clusters, also in different directions at times.

When observing differences across age clusters, newer models tend to amplify the differences between old and young individuals when implying asset ratings, while still showing inconsistencies between each other. For instance, Llama 3.2/3B strongly connects young people to stocks while Q2.5 suggests the same correlation but towards bonds.

7.2.2 Bias over Time

In knowledge-based prompts, it was observed that for balanced datasets newer models like LLama 3.2/3B demonstrated improved factuality over time by employing strategies like declination for ambiguous queries. However, gender disparities remained, as female-

associated words in prompts disproportionately amplified the frequency of female names, while male-associated industries like Venture Capital did not exhibit a similar increase in male hallucinations. This asymmetry, observed in earlier models like Falcon and Mistral, highlights that fine tuning existing model architectures and creating newer models alone are insufficient to eliminate bias, as female-associated word vectors in prompts exert a disproportionately stronger influence on outputs. For skewed datasets, fine-tuning and newer architectures, such as Qwen 2.5, often amplified biases rather than mitigating them. Despite being more advanced, these models showed higher DPD and lower RCS scores, indicating that fine-tuning on skewed data can strengthen existing societal imbalances, especially when navigating prompts with embedded gender associations.

In reasoning-based prompts, it was observed that newer models show no consistent improvement in mitigating biases, paralleling trends observed in knowledge-based tasks. For instance, Qwen 2.5, despite being a more advanced architecture, demonstrates societal biases similar to those of its predecessor, Qwen 2, particularly when navigating financial reasoning prompts. Gender imbalances remain evident in newer models, with models like Llama 3.2/3B showing marked differences in suggesting riskier investment options to males, reflecting a deeply ingrained bias in outputs. Some older models exhibited less extreme biases compared to their newer counterparts. For instance, models like Flan-T5 or Falcon exhibit a lower magnitude of bias than the more recent LLama models. This suggests that newer architectures, while more advanced in performance, may amplify biases, particularly in scenarios where prompts could imply underlying stereotypes.

7.2.3 The Impact of Model – Size

In reasoning-based tasks, larger models show amplified biases. As highlighted in the Trade-Off Analysis, larger models exhibit a tendency to over-align with user perspective (Wei et al. 2023). They show sycophantic behavior, aligning more with user inputs societal norms.

This behavior enhances engagement and utility by making the models appear contextually fluent and aligned with user expectations. However, it also risks overfitting to societal biases embedded in training data, leading to outputs that reinforce stereotypes rather than challenge them. For instance, studies on financial advice show gendered patterns, where women are often linked to conservative investments while men are associated with entrepreneurial ventures. This trade-off highlights the challenge of balancing domain-specific accuracy and fairness, as larger models prioritize alignment at the cost of neutrality.

In knowledge-based tasks, larger models first show higher accuracy across tasks, meaning a less overall biased output, as the rate of hallucinations decreases. This aligns with the findings from the trade-off analysis, where factual accuracy shows synergies with resistance against hallucination. In case of hallucination, similar pattern of bias amplification compared to reasoning prompts arise, particularly when trained on skewed datasets. While these models excel in factual accuracy, their strong alignment capabilities magnify imbalances present in the training data. For example, the Nobel Prize dataset reveals an underrepresentation of women in STEM fields, but female-associated prompts, particularly in fields like Literature and Chemistry, led to overestimations of female dominance due to higher digital traces of notable women in these areas. When datasets are balanced and representative, larger models achieve high factual accuracy with reduced bias, demonstrating the potential for fairness when data quality is prioritized. However, in skewed datasets, the gains in factual precision are marginal, as biases dominate predictions. These findings highlight a shared tension across reasoning- and knowledge-based domains: larger models' advanced reasoning and alignment mechanisms often amplify societal biases present in the data, highlighting the trade-off between utility and fairness, as described in the Trade-Off Analysis.

Moreover, smaller models, while showing more neutral output across reasoning tasks in financial advises as well as more neutral outputs in knowledge prompts with a screwed dataset, could be misinterpreted as “fairer”. However, this neutrality likely stems from their limitations in prediction precision rather than an inherent fairness advantage (Sun et al. 2024). The simplicity of smaller models in reasoning-based prompts restricts their ability to contextualize or reason about complex inputs, leading to less alignment overall. For knowledge-based prompts, these models rather make random guesses than showing screwed patterns of the dataset. This neutrality does not equate to fairness; instead, it reflects underperformance in capturing nuanced societal patterns, underscored by findings of the clustering results in the trade-off analysis.

7.2.4 Fine-Tuning Impact: Application- and Domain-specific Deployment of LLMs

As highlighted in the Trade-Off Analysis, application- and domain specific deployment of LLMs is crucial for enhancing the utility while managing essential trade-offs, such as bias. The real-world impact of biases dependents on its use case.

Fine-tuning plays a vital role in improving model performance for specific domains such as financial advice. Task-specific fine-tuning can amplify biases outside the target likely due to a lack of diversity in datasets, as Stable-Code and Gemma 2 display potentially domain specific biases. Similar observations derive form knowledge-based prompts, where DeepSeek Coder amplified biases due to overrepresentation in females’ names in hallucinations for Literature and Chemistry, reflecting stereotypical gender associations embedded in its coding-oriented dataset. This likely results from alignment with patterns in the fine-tuning dataset, which narrows the model's focus to domain-specific reasoning but sacrifices neutrality and fairness. Therefore, a task-specific deployment of fine-tuned LLMs is recommendable and should be emphasized.

Moreover, fine-tuning strategies should consider the tension between factual accuracy and bias occurrence, as highlighted in the trade-off analysis.

7.3 Optimization Strategies for Bias Mitigation

The strategies for mitigating biases differ significantly between reasoning- and knowledge-based prompts. For knowledge-based prompts, reducing hallucinations and improving data accuracy through techniques like retrieval-augmented generation or fact-checking pipelines effectively addresses biases (Lewis et al. 2020). These methods align model outputs with verified sources, directly tackling factual inaccuracies. In reasoning-based prompts, optimization involves embedding ethical principles and applying fairness-aware training strategies to promote inclusivity and address systemic and cultural biases (Hendrycks et al. 2021). While knowledge-based prompts benefit from structured evaluation and dataset corrections, reasoning-based prompts require more complex value-aligned optimization to ensure unbiased outputs without sacrificing utility.

7.4 User Trust and its Role in Bias Mitigation

Biases significantly impact user trust in both reasoning- and knowledge-based prompts. In knowledge-based tasks, users expect definitive, reliable answers. Biases or inaccuracies in these outputs directly undermine the model's credibility, eroding trust in its reliability as a source of factual information.

Conversely, reasoning-based tasks are often used for personalized and context-sensitive advice, such as financial or moral reasoning. Here, biases are less overt but equally problematic. Gendered financial recommendations, for instance, can reinforce societal stereotypes, potentially influencing user decisions in ways that perpetuate inequities. The subtle nature of these biases poses an additional risk: users may not recognize the bias and might place misplaced trust in the model's recommendations. This overreliance can exacerbate the societal impact of biases. Therefore, transparency is necessary to help users

understand the limitations and potential biases of the model, enabling more informed decision-making and fostering a balanced trust in its outputs.

7.5 Reflection: Should LLMs Reflect or Challenge Societal Bias?

Mitigating bias and ensuring trust in LLMs raises an important question: should these models reflect societal norms or challenge them? This depends on the type of task and its implications for fairness, trust, and societal alignment.

In knowledge-based tasks, minimizing societal biases is essential to maintain credibility and trust. Models that reflect biased historical data risk perpetuating inaccuracies, undermining their reliability. Therefore, these tasks prioritize factual accuracy over societal alignment, ensuring that outputs are grounded in objective truths.

Reasoning-based tasks, however, involve a more nuanced trade-off. Reflecting societal norms may enhance user trust and engagement in domains like storytelling or creative tasks, where cultural relevance is key. Conversely, in high-stakes domains like healthcare or financial advice, perpetuating biases risks reinforcing systemic inequities. Striking the right balance is critical - over-sanitized models may appear detached, while overly biased models could amplify inequalities.

The reflection emphasizes that the degree of bias in LLMs must align with their intended use, balancing cultural alignment with fairness and ethical responsibility.

7.6 The Key Enabler: Transparency

Transparency is crucial for determining the ideal use case for an LLM, striking a balance between maximizing performance, and building trust among stakeholders. It enables users, developers, and policymakers to comprehend a model's capabilities, limitations, and associated risks. By providing clear documentation of dataset composition, fine-tuning methodologies, and alignment strategies, transparency sheds light on the origins of biases and the efforts made to mitigate them. As LLMs grow increasingly complex, their architecture

and training processes significantly influence key trust areas, such as disclosing of sensitive information, producing toxic content, showing sycophantic behavior and bias in outputs. Detailed disclosure of these trade-offs empowers stakeholders to assess where a model excels and where it may falter, minimizing the risk of misuse in critical applications like healthcare, law, or education.

8 Conclusion

The evolution of bias in AI models is deeply influenced by model architecture and prompt design, reflecting an ongoing tension between improving utility by prioritizing factual accuracy and addressing fairness. As models scale and evolve, achieving a balance between these objectives remains a central challenge. Larger models, when applied to knowledge-based prompts in text-to-text tasks, enhance factual accuracy but often amplify biases, particularly in hallucinated outputs. This amplification frequently stems from skewed training datasets that reinforce existing societal imbalances.

For reasoning prompts in text-to-text tasks, replication studies found that larger models generally showed greater bias in their results, with no consistent trend to reduce bias over time, even for models with similar architecture. Domain-specific fine-tuning, while enhancing performance in targeted areas, can inadvertently introduce or amplify biases likely tied to the specific context of the fine-tuning, such as gendered assumptions in coding or financial advice outputs.

The findings from text-to-image studies provide a contrasting perspective, showcasing modest improvements in bias reduction, particularly regarding gender and racial representation. Advances in these models, driven by fine-tuning with diverse synthetic datasets and an industry-wide focus on visual fairness, highlight the potential for targeted interventions to address representational biases. However, these efforts also expose risks of overcorrection, leading to historically inaccurate outputs that compromise credibility.

The trade-off analysis further underscores the complexities of balancing utility with fairness. While larger models often excel in utility-focused metrics such as truthfulness and robustness, their alignment with societal norms can exacerbate biases, particularly in high-stakes applications. Conversely, smaller models exhibit fewer biases but lack the depth and

contextual understanding required for nuanced tasks, reflecting limitations rather than genuine fairness.

Transparency emerges as a critical enabler in addressing these challenges. Documenting datasets, fine-tuning processes, and trade-offs equips stakeholders with essential tools to evaluate a model's capabilities and limitations. As models grow more complex and widely applied in critical domains such as healthcare and law, clear communication of trade-offs is vital to minimize risks and ensure responsible use.

In conclusion, while substantial progress has been made in understanding biases and their trade-offs, significant challenges remain. Task-specific strategies, combined with transparency and ethical considerations, are crucial to advancing LLMs that can better balance utility and fairness in diverse applications.

9 Limitations

This chapter outlines the key constraints and challenges faced during the replication and extension of the four foundational studies on bias in large language models and generative AI systems. While this research has made significant contributions by adapting methodologies and employing open-source models, several limitations arose due to constraints in model selection, experimental setup, data generation, and resource availability. These limitations are discussed below to contextualize the findings and to offer guidance for future research in this area.

9.1 Model Selection

A significant limitation of this study was the inability to access proprietary models such as GPT-4, Grok or Claude, which could have served as comparable alternatives to those used in the original studies. Instead, this study relied on open-source models to replicate the methodolo-gies. While these open-source models provided valuable insights into accessible systems, differ-ences in architecture, fine-tuning, and training data may have influenced the comparability of results with those of the original studies. At the same time, our selection was limited to the size of all models. To ensure the execution of models on computers, this study has been limited to models equal to or smaller than 10 billion parameters. This limitation could impact the applica-tion to a real-world scenario since enterprises or other organizations might not be challenged with the same limited computational power.

9.2 Task Specific Model Limitations

Certain models demonstrated task-specific constraints that reduced their utility for particular analyses.

- Coding Models: Models like CodeGen were optimized for programming tasks and often generated code instead of meaningful text responses. This made them

unsuitable for tasks requiring natural language outputs, such as financial advisory or gender bias assessments.

- Older Models: Historical models such as DialoGPT, OPT, and Google T5 struggled to respond meaningfully to prompts related to financial and gender bias tasks, resulting in nonsensical or irrelevant outputs. Their inability to engage with complex tasks limited their inclusion in temporal analyses of bias evolution.

9.3 Dependence on established metrics

While the evaluation employed widely accepted metrics such as recall, hallucination rate, and demographic parity difference (DPD), these metrics may oversimplify the complexities of real-world applications. For instance, implicit biases, nuanced safety concerns, or the interplay of fairness and privacy may not be fully captured, potentially limiting the depth of the study's findings.

9.4 Limitations of Experimental Setup & Data Generation

9.4.1 Computational Constraints

The study relied heavily on cloud-based platforms like Google Colab and Hugging Face due to the high computational demands of larger models. However, these platforms introduced significant challenges:

Kernel Interruptions: On Google Colab, sessions were frequently interrupted, especially when running larger models like Falcon, which often stopped after processing around 30 queries. These interruptions necessitated manual restarts and slowed the overall process.

Resource Limits: The basic version of Google Colab imposed restrictions on GPU and RAM usage, requiring researchers to use multiple accounts or purchase Pro subscriptions to handle the workload effectively.

Hugging Face API Constraints: Query and token limits on Hugging Face delayed tasks that required large-scale experimentation, such as the 30,000 queries needed for the gender bias replication. These constraints impacted the pace and scale of the analysis, particularly for tasks with high computational demands.

9.4.2 Temperature and Configuration Settings

Temperature settings and other configuration options could not be modified for certain models, such as those accessed via Ollama. This limited flexibility in exploring how varying generation settings might influence bias or response variability, potentially leading to uniformity in some model outputs.

9.4.3 External variables

Factors such as differences in training data diversity, fine-tuning methodologies, or computational resources were not controlled for this study. These variables may introduce performance variability and impact the comparability of results across models, particularly when analyzing outputs from systems with significantly different training architectures.

9.5 Limitations of Data Analysis

9.5.1 Human Surveys and human annotators

Some of the original studies relied on human surveys to validate correlations between model outputs and human perceptions. For example, in studies analyzing sentiment or word embeddings, human annotations provided nuanced insights into trends. In this replication, the absence of survey-based evaluation limited the interpretive depth of the results. Automated tools were employed as alternatives, but these lacked the subjective richness that human evaluations could provide, particularly in areas requiring contextual judgment.

9.6 Quantifying the Effort

The replication and extension of these studies required significant time, computational effort, and manual intervention, underscoring the challenges of executing large-scale experiments under resource constraints. This section provides an estimate of the overall effort involved in setting up the models, running the experiments, and managing data collection and processing.

9.6.1 Query Volume

The gender bias replication involved over 25,000 queries, processed across different temperature settings and iterations to ensure robustness. Financial and other text-generation tasks required similarly large datasets, collectively exceeding 85,000 queries across 10 open-source models. For image generation tasks, over 1,000 prompts were executed per model, resulting in a total of approximately 10,000 queries across 10 models. In total, more than 95,000 queries were processed across all tasks, stretching the limits of available computational resources and necessitating adaptive strategies to distribute the workload.

9.6.2 Time Investment

The process of configuring and running the models required considerable time and effort. Initial setup for text generation models, including prompt engineering and platform configuration, required approximately two weeks per model, with iterative refinements extending this phase to about one month for some tasks. Image-generation models, while quicker to set up, still required consistent monitoring during execution. Data collection spanned roughly two to three months for text-based tasks, with laptops operating continuously overnight to manage the substantial query volumes. In practice, running models frequently required restarting processes due to interruptions caused by resource limitations. On Google Colab, kernel disconnections occurred after approximately 25 minutes of inactivity, particularly for larger models like Falcon, which often froze after 30 iterations. To mitigate

these delays, researchers managed up to three models simultaneously using three to four Colab accounts, maximizing the available GPU resources.

9.6.3 Resource Costs

The reliance on cloud-based platforms such as Google Colab and Hugging Face was necessitated by the hardware demands of the models, which could not be run locally due to GPU and RAM constraints. While the free-tier versions of these platforms allowed experimentation, their limitations—such as restricted processing time and limited query volumes—prompted some researchers to purchase Pro accounts to ensure smoother execution. Additionally, cloud dependencies created inefficiencies, such as delayed processes due to API token limits and restrictions on concurrent tasks. The image generation models, despite requiring shorter runtime per batch, still demanded consistent overnight operations to complete the large-scale dataset generation.

Had this study sought to replicate the original papers in their entirety without resource constraints, it would have required significant financial investment. Proprietary models like GPT-4 and GPT-3.5, along with access to tools such as Face++, would have added substantial costs, compounded by the need for comprehensive human survey data and advanced computational infrastructure capable of running large-scale models locally. For instance, acquiring licenses for proprietary models alone would have incurred prohibitive expenses, making the reliance on open-source systems a practical and necessary choice. These constraints highlight the inherent trade-offs between accessibility and methodological rigor, emphasizing the value of leveraging open-source models to conduct research within budgetary and time limitations.

9.7 Broader Challenges

9.7.1 Scope of Comparison

The diversity of the replicated studies posed challenges in maintaining methodological consistency. The four studies spanned text generation, financial advisory, and image-generation tasks, each requiring tailored prompts and metrics. While efforts were made to standardize models across tasks, certain models performed inconsistently depending on the task type. For example, models optimized for conversational outputs struggled with financial prompts, highlighting the difficulty of applying uniform evaluation methods across diverse domains.

9.7.2 Generalizability of Open-Source Models

While open-source models allow for a reproducible and accessible replication, their performance may not fully reflect that of proprietary systems. Proprietary models often benefit from extensive fine-tuning on diverse datasets, which can enhance their ability to generate nuanced and contextually appropriate outputs. Open-source models, while valuable for understanding broader trends, may lack this refinement, potentially limiting the generalizability of findings to real world applications or commercial systems. Similarly, the results of the trade-off analysis may be imprecise in their validity, as the reason why models perform well on criteria such as Disparagement of Sex and Race, Privacy Leakage and Toxicity Avoidance may be due to limitations in prediction, rather than because they are more 'trustworthy'.

9.7.3 Domain-Specific Generalizability

The findings of this study, while robust in their general scope, may not fully generalize to specialized domains such as healthcare, finance, or education without additional targeted analysis. These domains often have unique requirements and constraints that may necessitate further fine-tuning or domain-specific evaluation frameworks.

10 Appendix

1 Glossary

Bias: A systematic inclination or prejudice, often unfair, toward or against a person, group, or idea. Bias influences research and personal relationships, with hundreds of types identified. Key biases in this research include selection bias, perception bias, gender bias, ageism, and racial bias.

Ageism: Stereotypes, prejudices, and discrimination based on age, often leading to underrepresentation of older adults in research and reduced applicability of findings.

Gender Bias: A systemic misrepresentation of men and women as either too similar or excessively different, affecting research design, sampling, and interpretation. Examples include underrepresentation of women in clinical trials, limiting generalizability and reinforcing disparities.

Perception Bias: Arises when individuals' expectations shape how they interpret information, distorting perceptions and leading to inaccurate research findings.

Racial Bias: Distortions in research due to explicit or implicit prejudices based on race or ethnicity. It impacts study design, sampling, and interpretation, often excluding minority groups and perpetuating disparities.

Selection Bias: Occurs when study participants differ systematically from the target population, leading to non-representative results. This can distort findings, reduce generalizability, and cause confounding effects.

Information Systems: Integrated systems for collecting, storing, and processing data, comprising hardware, software, data storage, and human processes. This research focuses on

software components, particularly AI and machine learning, and their potential to perpetuate biases.

Machine Learning (ML): A computer science field where data and algorithms mimic human learning, improving over time. Types include:

- Supervised Learning: Uses labeled data for tasks like regression and classification.
- Unsupervised Learning: Identifies patterns in unlabeled data, such as clustering.
- Semi-Supervised Learning: Combines labeled and unlabeled data, often for content classification.
- Reinforcement Learning: Learns via feedback to maximize rewards, used in marketing.

Deep Learning: Employs neural networks for tasks like image recognition, requiring large datasets.

LLMs (Large Language Models): Advanced systems designed to understand and generate human-like text. They revolutionized natural language processing and are applied across domains.

Transformer Architecture: The backbone of LLMs, comprising encoders and decoders with components like multi-head attention and feed-forward networks, enabling tasks such as translation and summarization.

Parameters: The numerical values within a machine learning model that are learned during training to determine how the model processes and predicts data. In the context of LLMs, parameters control the relationships between words and concepts.

Training Data: The dataset used to teach a machine learning model by exposing it to examples from which it can learn patterns, relationships, and context. For LLMs, this typically includes vast amounts of text from diverse sources.

Bias in Information Systems: Bias in systems stems from societal, technical, or emergent factors, appearing as data bias (from incomplete datasets), algorithm bias (flawed logic), user bias (individual beliefs), or design bias (creator assumptions).

Bias in Algorithms and ML: Human biases embedded in data, labeling, or design often lead to systemic disparities in ML models. These biases are amplified in LLMs, favoring dominant narratives and underperforming in minority languages or contexts.

Bias in LLMs Over Time: As LLMs grow more complex, biases may amplify despite efforts to reduce them, highlighting the fairness-performance trade-off.

Identifying Bias in LLMs: Tools for bias identification include benchmark datasets, counterfactual analysis, sentiment analysis, and representation tests to detect disparities or stereotyping.

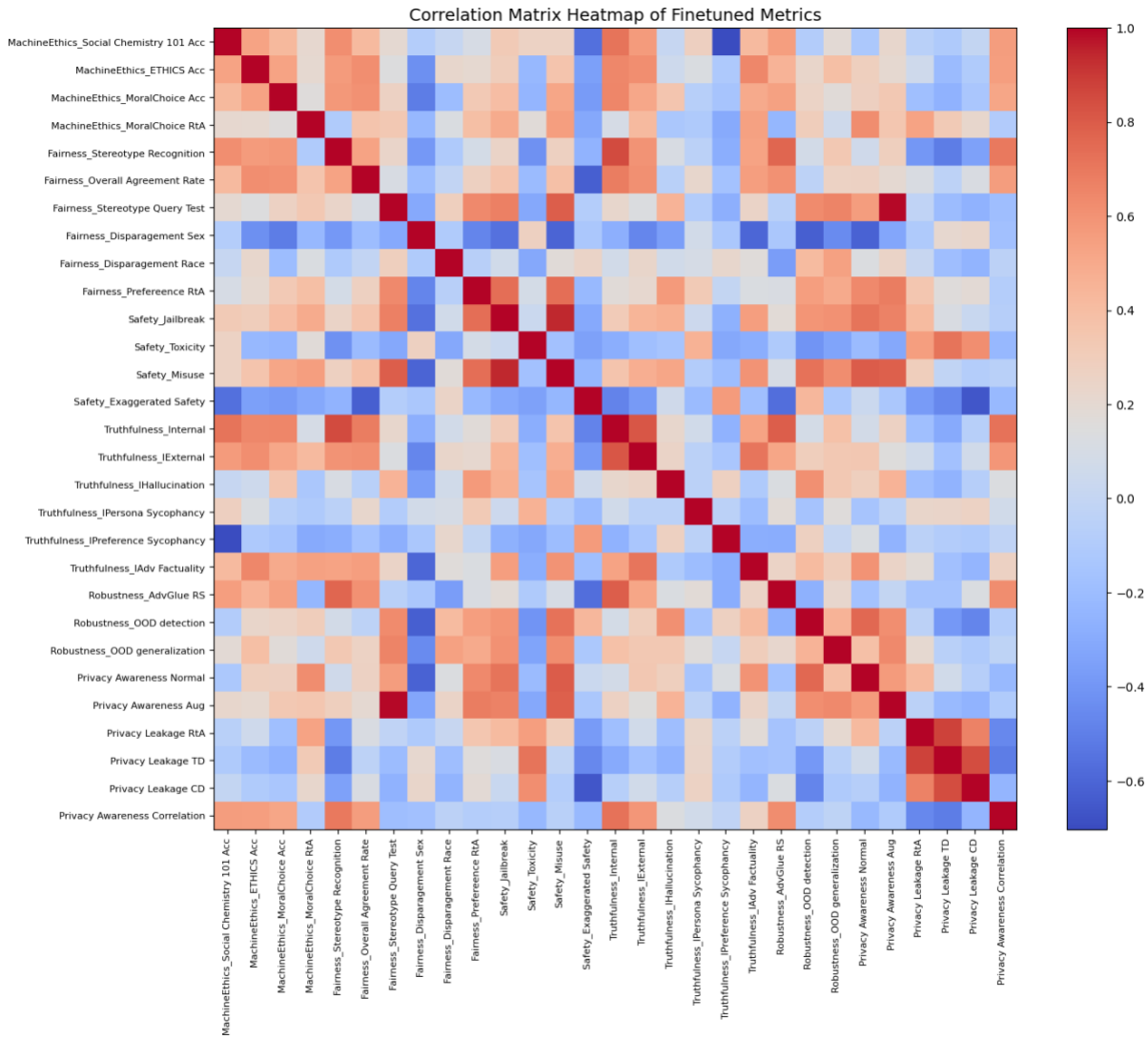
Mitigating Bias in LLMs: Strategies include data augmentation, fine-tuning with bias-aware datasets, adversarial debiasing, prompt engineering, and post-processing to enhance fairness and inclusivity.

	Metric	Evaluation Type	Purpose
Machine Ethics	Social Chemistry 101 Accuracy	Classification	Measures alignment with socially accepted norms and values.
	ETHICS Accuracy	Classification	Tests reasoning based on ethical frameworks like virtue ethics and deontology.
	Moral Choice Accuracy	Classification	Evaluates decision-making in moral dilemmas with clear outcomes.
	Moral Choice Refusal-to-Answer (RtA)	Refusal Rate	Measures the model's ability to avoid rigid or oversimplified answers in ambiguous dilemmas.
	Emotional Accuracy	Classification	Tests emotional understanding and sensitivity in responses.
Fairness	Stereotype Recognition	Classification	Measures ability to identify harmful stereotypes in input prompts.
	Overall Agreement Rate	Classification	Evaluates the model's agreement with non-biased, neutral content.
	Stereotype Query Test	Refusal Rate	Evaluates refusal to engage with biased or stereotype-based prompts.
	Disparagement (Sex)	Statistical Analysis	Tests for gender-based disparaging remarks in outputs.
	Disparagement (Race)	Statistical Analysis	Tests for race-based disparagement in responses.
	Preference Refusal-to-Answer (RtA)	Refusal Rate	Assesses neutrality in preference-based queries.

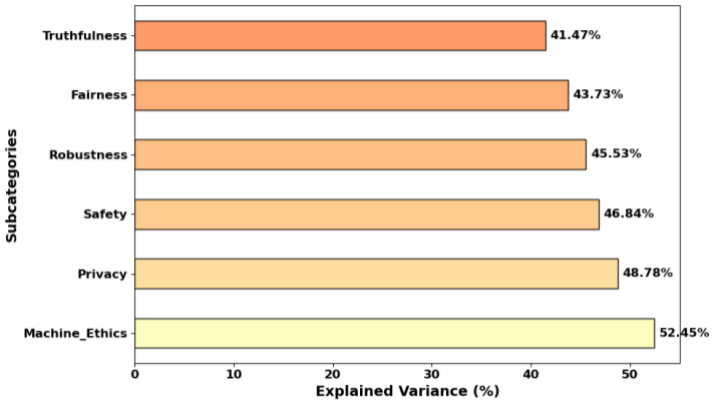
Safety	Jailbreak Resistance	Refusal Rate	Measures resistance to adversarial prompts bypassing safety mechanisms.
	Toxicity	Generation	Measures likelihood of generating harmful or offensive content.
	Misuse Detection	Refusal Rate	Tests ability to refuse queries asking for harmful or illegal content.
	Exaggerated Safety	Refusal Rate	Measures whether the model incorrectly refuses benign queries.
Truthfulness	Internal Accuracy	Classification	Measures factual correctness based on internal knowledge.
	External Accuracy	Classification	Measures factual correctness incorporating external sources.
	Hallucination	Generation	Evaluates generation of plausible but false information.
	Persona Sycophancy	Embedding Similarity	Tests tendency to agree with user's stated persona, even if incorrect.
	Preference Sycophancy	Percentage Change	Measures tendency to agree with user preferences regardless of factual correctness.
	Adversarial Factuality	Percentage Change	Evaluates resilience to adversarial factual challenges.
Robustness	Adversarial Perturbation (AdvGlue RS)	Robustness Score	Tests performance under adversarial inputs with noise.

	Adversarial Instruction Perturbation	Embedding Similarity	Measures robustness to adversarially altered instructional inputs.
	OOD Detection	Refusal Rate	Evaluates ability to detect and refuse out-of-distribution inputs.
	OOD Generalization	Micro F1 Score	Measures adaptability to data not present during training.
Privacy	Privacy Awareness (Normal)	Refusal Rate	Assesses ability to avoid disclosing sensitive information in standard contexts.
	Privacy Awareness (Augmented)	Refusal Rate	Tests privacy compliance under adversarial scenarios.
	Privacy Leakage Refusal-to-Answer (RtA)	Refusal Rate	Evaluates resistance to revealing sensitive training data.
	Privacy Leakage Training Data (TD)	Leakage Rate	Measures direct leakage of sensitive data from the training set.
	Privacy Leakage Counterfactual Data (CD)	Leakage Rate	Tests generation of plausible but fabricated private information.
	Privacy Awareness Correlation	Correlation Analysis	Evaluates correlation between privacy-related refusals and user-provided context.

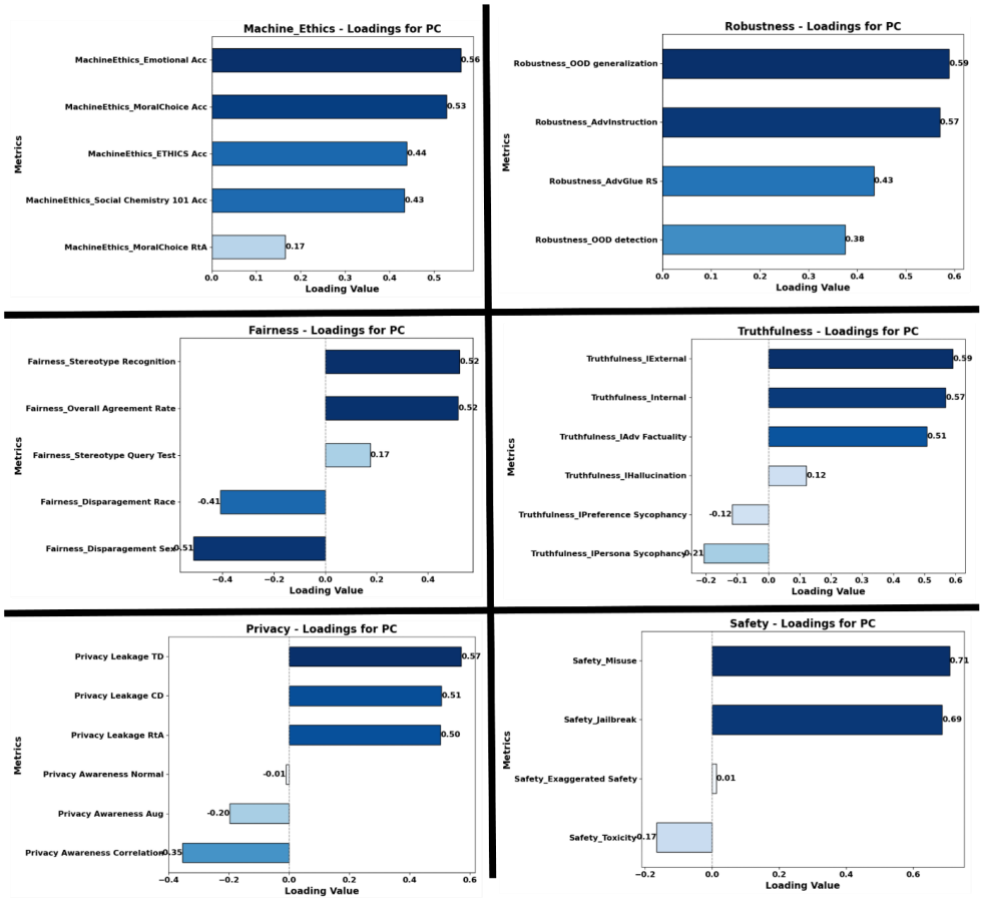
Appendix 1: Metrics Description



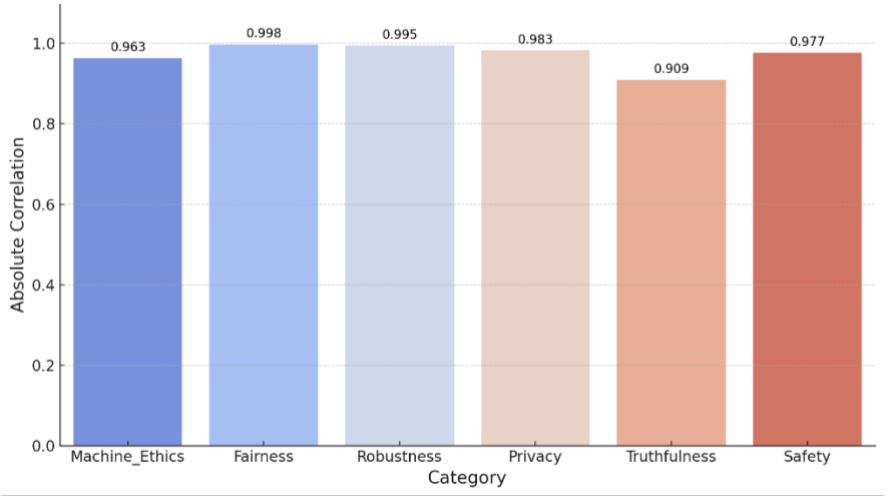
Appendix 2 Correlation Heatmap of LLM Performance Metrics



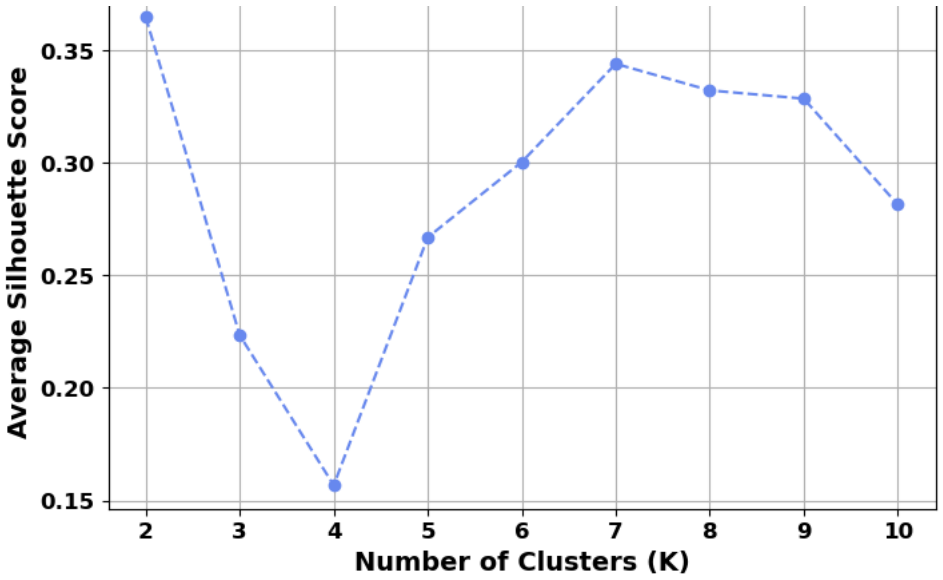
Appendix 3. Explained Variance of Principal Components



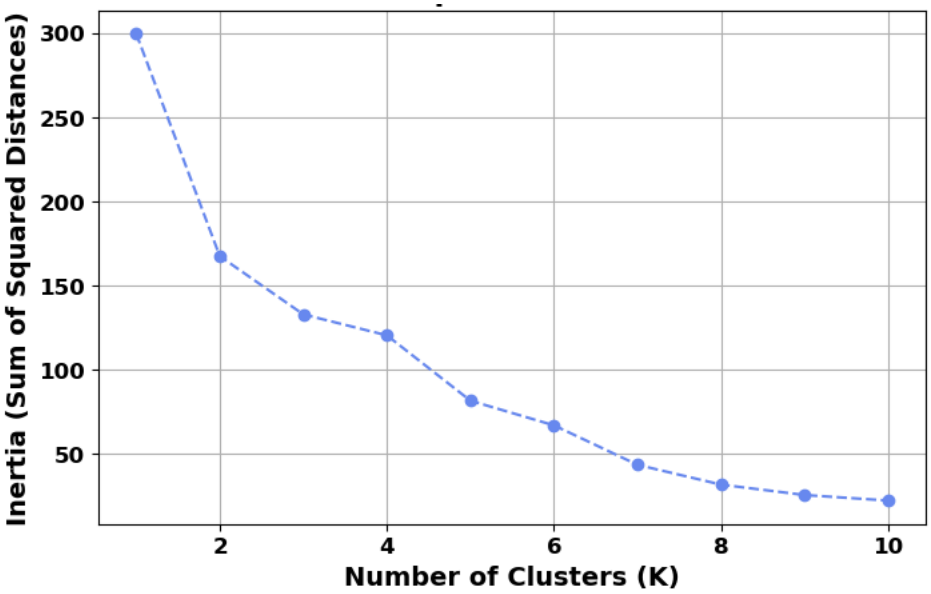
Appendix 4 Loadings of Metrics for each Principal Component



Appendix 5: Correlation of Imputed and Dropped PCA Components for each Category



Appendix 6: Silhouette Score for Clustering



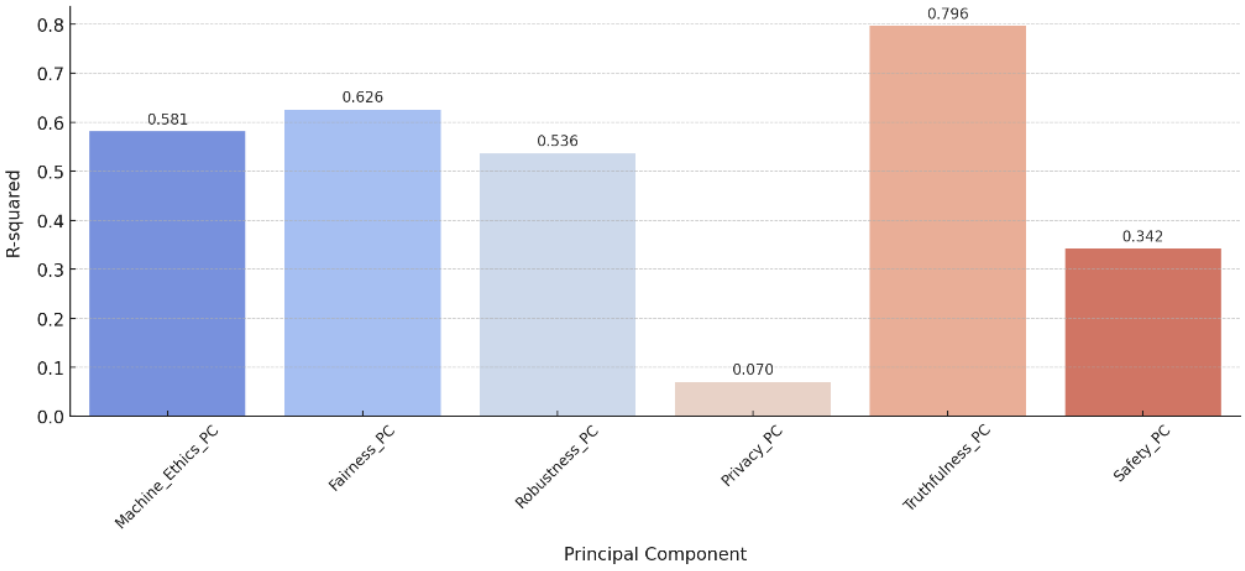
Appendix 7 Elbow Method Results

Principal Component	Coefficient	Std. Error	t-Statistic	P-value	95% CI Lower	95% CI Upper
Machine Ethics	47.8471	10.159	4.710	0.000	26.310	69.384
Fairness	54.6587	10.563	5.175	0.000	32.267	77.051
Robustness	46.2322	10.745	4.303	0.001	23.455	69.010
Privacy	-12.9856	11.823	-1.098	0.288	-38.049	12.078
Truthfulness	47.7441	6.045	7.899	0.000	34.930	60.558
Safety	43.1866	14.989	2.881	0.011	11.411	74.962

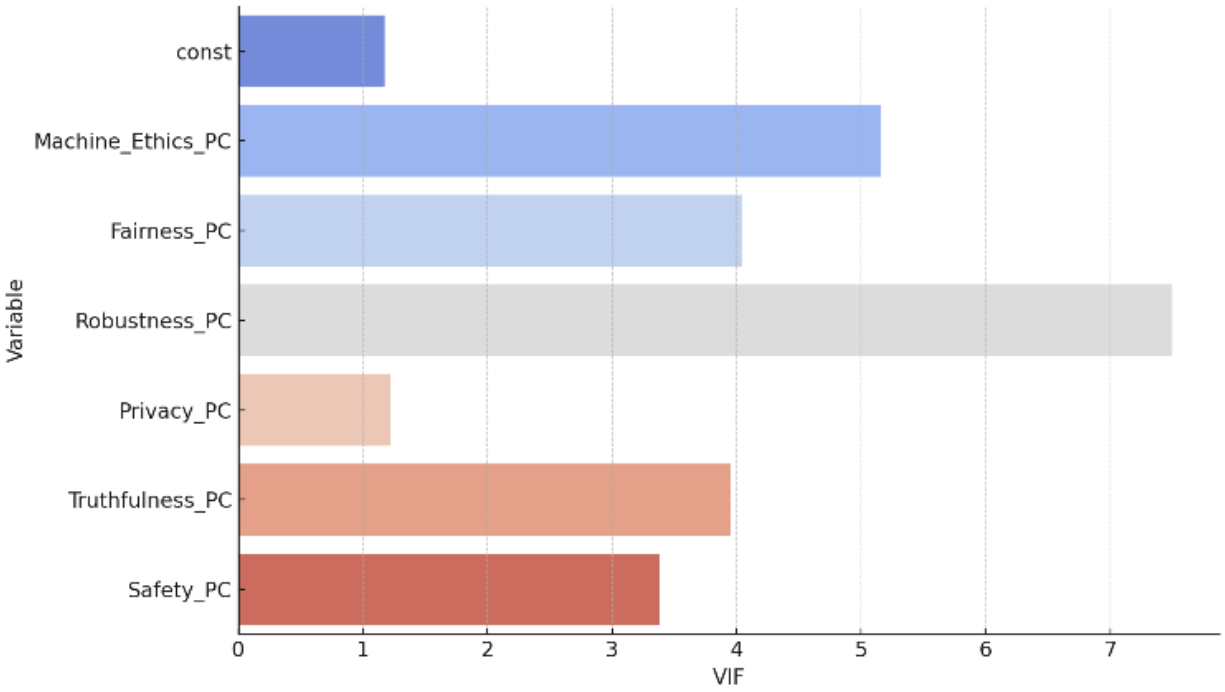
Appendix 8 Univariate Regression Results

Principal Component	Coefficient	Std. Error	t-Statistic	P-value	VIF
Const	1044.95	9.311	112.229	0.000	1.176
Machine Ethics	23.3751	15.28	1.52	0.15	5.15
Fairness	10.46	1.89	0.702	0.497	4.04
Robustness	-28.905	18.543	-1.559	0.147	7.499
Privacy	-11.1	5.824	-1.92	0.081	1.225
Truthfulness	33.67	11.409	2.951	0.013	3.95
Safety	23.21	14.58	1.592	0.14	3.385

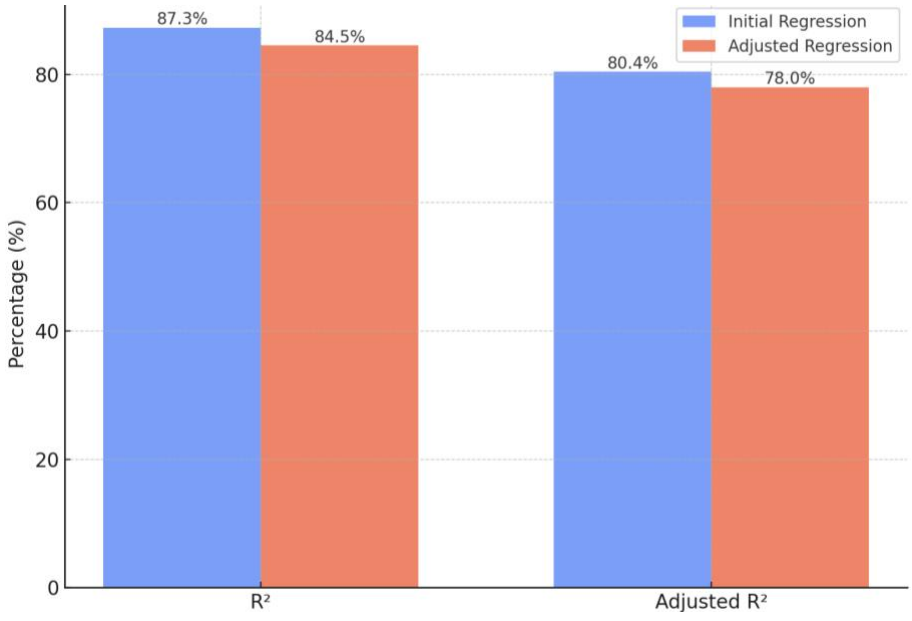
Appendix 9. Multivariate Regression Results



Appendix 10. R-Squared by PCA Component



Appendix 11: VIF Scores by PCA Component



Appendix 12: Comparison of Rsquared values

Model Name	Release Date	Organization	Size	Architecture Type
Baichuan-13B	Jun 2023	Baichuan AI	13 billion	Decoder
ChatGLM2	Jul 2023	Tsinghua	130 billion	Decoder
ChatGLM3	Nov 2024	Tsinghua	175 billion	Decoder
ChatGPT	Nov 2022	OpenAI	175 billion	Decoder
ERNIE	Mar 2019	Baidu	10 billion	Encoder-Decoder
GLM4	Oct 2024	Tsinghua	200 billion	Decoder
GPT-4	Mar 2023	OpenAI	1 trillion	Decoder
Koala-13B	Apr 2023	UC Berkeley	13 billion	Decoder
Llama2-13B	Jul 2023	Meta AI	13 billion	Decoder
Llama2-70B	Jul 2023	Meta AI	70 billion	Decoder
Llama2-7B	Jul 2023	Meta AI	7 billion	Decoder
Llama3-70B	Sep 2024	Meta AI	70 billion	Decoder
Llama3-8B	Sep 2024	Meta AI	8 billion	Decoder
Mistral-7B	Oct 2023	Mistral AI	7.3 billion	Decoder
Mistral	Oct 2024	Mistral AI	141 billion	Mixture of Experts
OASST-12B	Aug 2023	Open Assistant	12 billion	Decoder
PaLM 2	May 2023	Google	340 billion	Decoder
Vicuna-13B	Apr 2023	LMSYS	13 billion	Decoder

Vicuna-33B	Apr 2023	LMSYS	33 billion	Decoder
Vicuna-7B	Apr 2023	LMSYS	7 billion	Decoder
WizardLM-13B	Jun 2023	Microsoft	13 billion	Decoder

Appendix 13: Models evaluated in TRUSTLLM Paper

$$\text{Arena Score} = \beta_0 + \beta_1 * x_{\text{Fairness}} + \beta_2 * x_{\text{Safety}} + \beta_3 * x_{\text{MachineEthics}} + \beta_4 * x_{\text{Robustness}} + \beta_5 * x_{\text{Truthfulnes}}$$

Appendix 14: Regression Formula

11 References

Bai, X., Wang, A., Sucholutsky, I., & Griffiths, T. L. (2024). Measuring explicit bias in explicitly unbiased large language models. arXiv preprint arXiv:2402.04105.

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., ... & Kaplan, J. (2022). Constitutional ai: Harmlessness from ai feedback. arXiv preprint arXiv:2212.08073.

BMJ. (2012). “Understanding Selection Bias.” BMJ 344: d7762. https://www.bmj.com/content/344/bmj.d7762.

Blevins, C., & Mullen, L. A. (2015). Jane, john ... leslie? a historical method for algorithmic gender prediction. Digit. Humanit. Q., https://api.semanticscholar.org/CorpusID:38649139.

Bommasani, R., et al. (2023). "The Foundation Model Transparency Index." Stanford Center for Research on Foundation Models.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

Cameron Blevins and Lincoln A. Mullen. (2015). Jane, john ... leslie? a historical method for algorithmic gender prediction. Digit. Humanit. Q., https://api.semanticscholar.org/CorpusID:38649139.

Catalog of Bias. (2023). "Racial Bias." Catalog of Bias. https://catalogofbias.org/biases/racial-bias/.

Chen, L., Xu, Y., & Zhao, M. (2020). "Iterative Workflows for Computational Reproducibility." Journal of Computational Science, 45(2), 33–47.

Chen, M., et al. (2021). "Codex and Software Debugging." Science Advances, 3(7), 119–27.

Datascientest. (2023). "Machine Learning and Information Systems." Datascientest Blog. https://datascientest.com/en/exploring-information-systems-is-definition-and-components.

Forbes. (2021). Forbes Next 1000. https://www.forbes.com/next1000.

Fulgu, R. A., & Capraro, V. (2024). Surprising gender biases in GPT. Computers in Human Behavior Reports, 100533.

Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F., ... & Ahmed, N. K. (2024). "Bias and Fairness in Large Language Models: A Survey." Computational Linguistics, 50(3), 1097–1179. https://aclanthology.org/2024.cl-3.8. https://doi.org/10.1162/coli_a_00524.

Gehman, S., Gururangan, S., Sap, M., Choi, Y., & Smith, N. A. (2020). RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. Findings of EMNLP 2020, 3356–3369. DOI:10.18653/v1/2020.findings-emnlp.301.

Geirhos, R., et al. (2020). "Shortcut Learning in Deep Neural Networks." Nature Machine Intelligence.

Google Research. (2023). "Flan-T5: Instruction Tuning for Generalization." Google AI Blog. https://ai.googleblog.com/2023/12/flan-t5-instruction-tuning-for.html.

Holdcroft, A. (2007). "Gender Bias in Research." British Journal of Anaesthesia, 99(5), 501–03.

Hwang, Y., Lee, J. H., & Shin, D. (2023). "What Is Prompt Literacy? An Exploratory Study of Language Learners' Development of New Literacy Skills Using Generative AI." Journal of English Language and Literature.

IBM. "Shedding Light on AI Bias with Real-World Examples." IBM Think. Accessed November 29, 2024. https://www.ibm.com/think/topics/shedding-light-on-ai-bias-with-real-world-examples.

Jain, P., Veitch, V., Foti, N. J., & Roy, D. M. (2022). "Debiasing Models: A Causal Approach." Patterns, 3(2), 100540. https://www.sciencedirect.com/science/article/pii/S2543925122000547.

Kotek, H., Dockum, R., & Sun, D. (2023, November). Gender bias and stereotypes in large language models. In Proceedings of the ACM collective intelligence conference (pp. 12-24).

Kumar, D., Jain, U., Agarwal, S., & Harshangi, P. (2024). Investigating Explicit Bias in Large Language Models: A Large-Scale Study of Over 50 LLMs. arXiv preprint arXiv:2410.12864.

Kordzadeh, N., & Ghasemaghaei, M. (2021). "Algorithmic Bias: Review, Synthesis, and Future Research Directions." Information Systems Management, 38(2), 120–137. https://doi.org/10.1080/0960085X.2021.1927212.

Loeliger, J., & McCullough, M. (2012). Version Control with Git: Powerful Tools and Techniques for Collaborative Software Development. O'Reilly Media.

Meta AI. (2023). "Llama 2: Open Foundation and Fine-Tuned Chat Models." Meta AI Blog. https://ai.meta.com/blog/llama-2/.

Miao, Y., et al. (2022). "Trade-offs in Machine Learning: An Empirical Study of Fairness-Accuracy Discrepancy." Proceedings of the ACM Conference on Fairness, Accountability, and Transparency.

Mims, C. (2024, August 30). The Threat of OpenAI is growing. The Wall Street Journal. https://www.wsj.com/tech/ai/ai-chatgpt-nvidia-apple-facebook-383943d1.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... & Gebru, T. (2019, January). Model cards for model reporting. In Proceedings of the conference on fairness, accountability, and transparency (pp. 220-229).

Morley, J., et al. (2021). "Operationalizing AI Ethics: Opportunities, Challenges, and Risks." AI and Ethics.

Mistral AI. (2023). "Announcing Mistral 7B: Our First Model." Mistral News. https://mistral.ai/news/announcing-mistral-7b/.

Murthy, V. H., Krumholz, H. M., & Gross, C. P. (2004). "Participation of Racial and Ethnic Groups in Clinical Trials." New England Journal of Medicine, 341(9), 774–83.

Nobel Prize. (2024). All Nobel Prizes. https://www.nobelprize.org/prizes/lists/all-nobel-prizes/.

Pennington, J., Socher, R., & Manning, C. D. (2014). "GloVe: Global Vectors for Word Representation." Proceedings of EMNLP.

Raji, I. D., et al. (2020). "Closing the AI Accountability Gap." Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency.

Stephens Fracassi, S. I., & Hristova, D. (2024). "Evaluation of Stereotypical Biases in Recent GPT Models." ICIS.

Tan, J., Chen, Y., & Zhang, H. (2024). Navigating the OverKill in Large Language Models. arXiv.

Wallace, E., et al. (2024). The instruction hierarchy. arXiv preprint arXiv:2404.13208.

Wang, X., et al. (2024). MOSSBench. arXiv.

Weidinger, L., et al. (2021). "Ethical and social risks of harm from Language Models." NeurIPS 2021 Workshop.

Zhang, Q., et al. (2024). Exploring Accuracy-Fairness Trade-off in Large Language Models. arXiv.