UNIVERSITY OF TURKU
Turku School of Economics

Mapping the Impact of Biases in Large Language Model Chatbots on User Satisfaction

Master's thesis
International Business

Author(s):
Olayinka Vicente

Supervisor(s):
D.Sc. Majid Aleem
D.Sc. Elina Pelto

20.05.2025
Turku

The originality of this thesis has been checked in accordance with the University of Turku quality assurance system using the Turnitin Originality Check service. I hereby confirm that I have utilized Artificial Intelligence (AI) assistance; please refer to the Appendix titled Artificial Intelligence (AI) Assistance Declaration for details.

Master's Thesis

Subject: International Business
Author: Olayinka Vicente
Title: Mapping the Impact of Bias in Large Language Model Chatbots on User Satisfaction
Supervisor(s): D.Sc. Majid Aleem, D.Sc. Elina Pelto
Number of pages: 74 pages + Appendices 4 pages
Date: 20.05.2025

This thesis explores how biases in Large Language Model (LLM) chatbots have evolved over time and how they affect user satisfaction. Through a structured literature review of 89 peer-reviewed articles, the study maps the historical trajectory of chatbot technologies—from rule-based systems to advanced LLMs—and reviews the persistence and emergence of different bias types based on their sources. It also explores user trust, experience, and perceptions in relation to biased interactions, while assessing the evolution and limitations of various mitigation strategies.

The findings revealed that biases in chatbots often persists across generations and have grown more complex, with recent emergent behaviours linked to training data, algorithmic design, and human interaction. Additionally, despite the improvements in mitigation techniques, there are still some inconsistencies and ethical gaps that remain. This study contributes to AI fairness and user experience research by proposing an evolution-informed understanding of bias in chatbots, and offering recommendations for research in ethically aligned, user-sensitive chatbot development.

Key words: Large Language Models (LLMs), Chatbot Bias, User Satisfaction, User Trust, AI Ethics, Human-AI Interaction, Algorithmic Fairness, Bias Mitigation, Conversational Agents.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

1 Introduction

1.1 Background

Caldarini et al. (2022, 1) defined chatbots as intelligent conversational computer programs that are designed to mimic natural human conversations to enable automated online guidance and support. These chatbots have contributed to the service-oriented industries by improving the way they interact with their customers, i.e. customer service (Lee & Chan 2024, 88). Over the years, chatbots have transformed from simple constrained systems to more advanced models, ultimately culminating in the development of Large Language Model (LLM)-powered chatbots (Dam et al. 2024, 1–2). Considering these advancements, LLM chatbots like OpenAI's ChatGPT, Meta's LLaMA and Google's Gemini, now offer human-like interactions by adapting to users' needs and responding to complex queries in real time (Meduri 2024, 722). Thus, compared to their predecessors, these chatbots do not depend on scripted responses, instead they generate their own replies based on pattern recognition from pre-trained data that is usually conversational, adaptive and contextually relevant (Zhou 2024, 86).

As service-oriented businesses continue to implement AI-chatbots as the initial point of contact as well as in other customer interactions, user satisfaction has become a key construct that affects operational efficiency gains, brand perceptions, perceived service quality and customer retention rates (Nicolescu et al. 2022, 19; Meduri 2024, 722–723; Wut et al. 2024, 9; Al-Shafei 2025, 412). Therefore, high levels of satisfaction can lead to positive outcomes such as repeat usage and operational cost savings while dissatisfaction would cause more negative outcomes such as of reputational damage or erosion of customer base (Nicolescu et al. 2022, 19-20).

However, despite these significant advances, concerns regarding bias in LLM-powered chatbots have also intensified (Chan & Wong 2024, 1; Zhou 2024, 86). Bias in AI itself is not new, with studies having already revealed that chatbots could inherently offer stereotypical or prejudiced responses related to ethnicity, race, and/or gender (Chan & Wong 2024, 2). These biases stem from underlying training data, fine-tuning mechanisms and reinforcement learning techniques (Zhou 2024, 89). So far, various studies have identified a few other ways bias is able to manifest in customer service chatbots. For instance, there is rule-based bias which was mostly seen in older chatbots (such as ELIZA)

and was likely exhibited through their "poor understanding of context and language" (Xue et al. 2023, 6) due to being directly encoded with specific rules and patterns by their developers. Another more common example found in modern LLMs would be data-driven bias, where they "inherit and amplify biases present in their training data" (Chan & Wong 2024, 2). There is also bias related to sentiment and response where certain phrases, or customer demographics could receive more positive or negative responses (Huang et al. 2019, 7).

Nevertheless, as these systems have become more human-like, user expectations have evolved, reflecting an increasing demand for trustworthy and reliable interactions (Følstad & Brandtzæg 2017; Brandtzaeg & Følstad 2018; Lee & Chan 2024, 89). However, the presence of such biases directly affects overall user trust and satisfaction (Xue et al. 2023, 8; Lu 2024, 823), making it vital to address transparency, bias and fairness to maintain user confidence and ethical integrity (Dam et al. 2024, 15).

1.2 Research Gaps

Although there are numerous comprehensive studies on bias in LLMs and AI-powered chatbots, such as (Xue et al. 2023; Dam et al. 2024; Das & Sakib 2024; Chan & Wong 2024), a majority of them focus on identifying, measuring, and mitigating biases at specific points in time as opposed to examining their progression. This point-in-time approach is unable to fully capture how biases evolve when AI models are updated, thus potentially leading to overlooked patterns and possibly ineffective long-term mitigation strategies (Eticas 2025). In parallel to this, various studies have explored the evolution of LLM chatbots, such as those by (Caldarini et al. 2022; Xue et al. 2023; Akhtar 2024; Wang et al. 2024; Dam et al. 2024; Naik et al. 2024), focusing on areas such as conversational capabilities, practical applications and overall performance.

While the above authors also acknowledge the presence of different types of biases (like algorithmic and sentiment bias) and propose potential mitigation strategies to address them (such as diversifying training data and using counter-stereotypic imagining), it is unclear how persistent or emerging biases may have explicitly influenced users' trust, engagement, or perceived fairness over time. A major reason for this uncertainty could stem from how biases can be subtle in nature, thus going undetected by users. (Holroyd 2012, 292). It therefore becomes difficult to quantify the degree in which it occurs and possibly how it

affects user experience and satisfaction, especially with current models sometimes struggling to understand nuanced user intents (Holroyd 2012, 292; Howard & Borenstein 2018, 1527; Lu 2024, 3). Furthermore, recent research has shown transparency and user satisfaction play key roles in establishing the credibility of a chatbot (Chan & Wong 2024, 6; Lee & Chan 2024, 93). Therefore, the presence of bias can adversely affect users' satisfaction and perceptions of chatbot credibility, which would essentially influence customer trust (Lee & Chan 2024, 89). A few other studies also indicate that since chatbots can be integrated across different channels in different industries, they are usually required to adapt their responses, taking contextual situations, linguistic and/or regional inputs into account (Ekechi et al. 2024, 1264).

However, various alignment and adaptability processes could lead to inconsistencies in how bias manifests and thus how users are affected (Ryan et al. 2024, 8). Therefore, bias is not experienced uniformly as chatbot behaviour could systematically favour certain patterns over others (Xue et al. 2023, 11). Some bias mitigation techniques and/or strategies have evolved but are also inconsistent, where some proposed models still inadvertently demonstrate sentiment bias, reinforcement learning alignment issues and response filtering (Huang et al. 2019; Dai et al. 2024; Ryan et al. 2024).

That said, the current body of research on chatbot bias presents three key limitations when looking at the context of this thesis. The first is in relation to how most of these studies, such as (Talboy & Fuller 2023; Zhou 2024; Guo et al. 2024), examine bias at a single point in technological development, with little to no connections between findings across different chatbot generations. This hinders researchers from understanding how biases persist, evolve or even emerge across the different generations, which could in turn allow harmful patterns to go unaddressed. The second revolves around the explicitly minimal analysis tracking on how bias has evolved from the rule-based systems into modern LLMs, despite the extensive research into current LLM systems (Eticas 2025). The historical gap here limits researchers' understanding of which biases are inherent to the model's foundations, and which are introduced through newer techniques.

Then third, already existing relevant insights are dispersed, scattered across human-computer interaction (HCI), business, health, AI ethics, amongst other fields. Considering this, it reflects how some studies on algorithmic bias in AI operate in disciplinary silos, without engaging with broader perspectives (Blodgett et al. 2020, 5461–5462). Therefore,

in the absence of interdisciplinary collaboration, some of the proposed mitigation strategies likely end up being misaligned with applications outside a specific disciplinary silo (Blodgett et al. 2020, 5462).

1.3 Purpose of the Study

Therefore, the aim of this study is to review and map the historical trajectory of bias in LLM-powered chatbots, and the impacts it may have on user satisfaction. To address the above gaps, a systematic literature review will be conducted to achieve the below:

1. Analyse the historical trajectory of biases from early AI-powered chatbots to modern LLM-powered chatbots, thus mapping and identifying which biases may have persisted, evolved and likely emerged in recent years.

2. Examine the evolution of user responses and/or reactions to biased interactions, by tracking potential changes in user trust, satisfaction, and engagement across different chatbot generations.

3. Assess the development of proposed mitigation strategies by evaluating how these approaches have evolved and where inconsistencies and limitations remain.

4. Then synthesize potential insights to develop a comprehensive understanding of bias evolution and propose targeted recommendations for future studies.

By investigating bias as an evolving issue, this study offers contributions to AI fairness research and user experience design, therefore informing potential strategies for AI developers and businesses to improve chatbot credibility. The findings from this study are expected to facilitate the development of more effective bias mitigation strategies based on historical patterns and trends rather than just point-in-time analyses. Building upon the identified research gaps and the established purpose of this study through the highlighted objectives, the following chapter will outline the methodological approach designed to systematically address these limitations and advance understanding of bias evolution in AI-powered conversational systems.

2 Research Design

2.1 Research Approach

In line with the earlier highlighted research gaps and objectives of this thesis, a systematic literature review was conducted in accordance to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines (Lame 2019; Page et al. 2021) as well as Egger et al.'s (2022, 20–21) proposed steps to conducting a systematic review. The PRISMA framework provided a structured approach to identify, screen and report eligible studies which minimized reporting bias and improved the overall consistency of study inclusion across the used databases (Page et al. 2021, 2). Egger et al.'s (2022) proposed steps further complemented this by offering detailed techniques to assess potential biases, manage heterogeneity and synthesize data. Integrating these two guidelines ensured transparency and reproducibility in evaluating and synthesizing existing research on the evolving nature of bias in LLM chatbots, while also enhancing the overall study selection process.

The review process began by clearly defining the research objectives to guide the literature search. Then during the identification stage (Page et al. 2021, 8), the inclusion and exclusion criteria was established and structured around the PICO framework (participants, interventions, comparators and outcomes). This framework was especially useful in fostering the selection of relevant studies as it provided a systematic lens to frame the research focus on bias in chatbot systems and its impacts on user experiences. That said, the "*participants*" in this case were seen as users interacting with LLM chatbots; "*interventions*" were the existing bias mitigation techniques; "*comparators*" made it possible to differentiate between the older and newer chatbot generations; and "*outcomes*" included user perception, satisfaction, trust and engagement.

The inclusion criteria focused on studies that offered overviews of chatbots, analysed potential biases and their mitigation strategies in AI-powered chatbots and other LLM-based models and examined user satisfaction, trust, and/or engagement in chatbot interactions. On the other hand, blog posts and/or opinion pieces lacking empirical evidence and research that did not focus on AI biases and mitigation strategies as well as AI related user satisfaction was overlooked.

The next phase involved screening and assessing the eligibility of relevant studies across various databases using the PRISMA abstract checklist (Page et al. 2021, 7). These checklists ensured a consistent screening process by providing and maintaining the evaluation criteria. They can be found under Appendix 2 and 3 respectively. Following this, Blodgett et al.'s (2020, 5460) heterogeneity checks were applied to track as well as evaluate the quality and potential biases of the selected studies.

Once the articles were selected, they were thematically analysed to identify trends across the different chatbot generations. This approach permitted the findings to be categorized based on persistent and emerging chatbot biases; their corresponding proposed mitigation approaches as well as user perceptions of them overtime. Such categorization was possible as a result of combining the PRISMA and Egger et al. (2022) protocols, which facilitated the inclusion procedures while improving overall traceability across the different periods of chatbot development. Finally, the implications of these findings were interpreted in the context of AI fairness, user satisfaction and responsible chatbot deployment, thus giving way to future research recommendations.

To therefore systematically track and document the processes described in the research approach, a flowchart was adapted from Page et al.'s (2021, 8) work and is listed under Figure 1. The figure provides a visual representation of the systematic review process which is eventually used to create the final selection of articles (see Table 2).

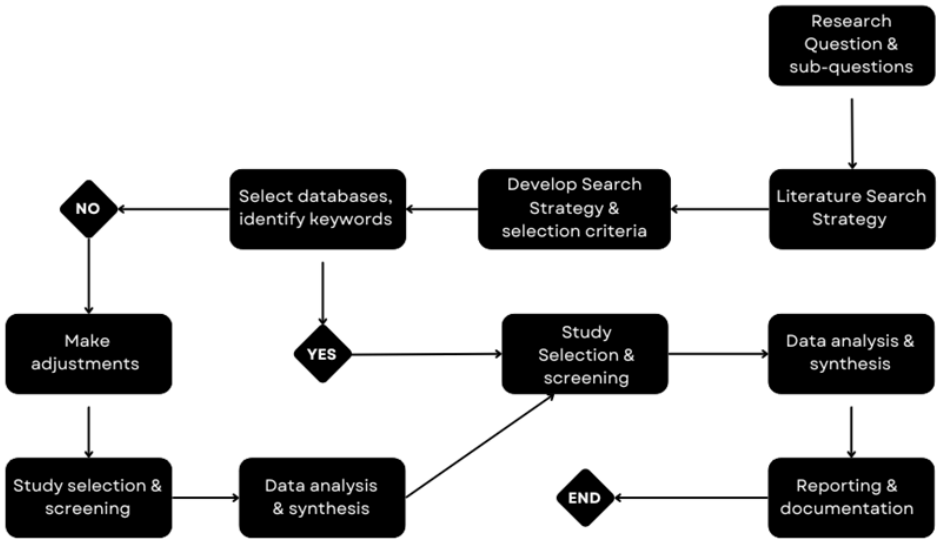

Figure 1. Flowchart depicting the entire Systematic Review Process in this thesis

Taking the research objectives of this thesis into account, this approach ensured that the final selected studies met the inclusion criteria. The iterative nature of the process displayed in Figure 1, especially the feedback loop used for adjustments and refining the review protocol, further strengthened the reliability and responsiveness of the final literature corpus. Additionally, the loop permitted real-time refinements to search terms, and the inclusion criteria based on the emerging insights and preliminary results. Revisiting and adjusting the strategy allowed the process to become more targeted and relevant, thus reducing the risk of potentially overlooking critical studies while adapting to the evolving terminology and research scope within the field. This resulted in an enhanced approach allowing for overall methodological rigor while also ensuring the final dataset was both comprehensive and contextually aligned with the thesis objectives**.** It also served as a critical foundation for the subsequent data analysis and bias evolution mapping, enabling a nuanced understanding of how bias in LLM chatbots has changed over time and affected user experience.

2.2 Data Collection

2.2.1 Literature Search Strategy

The literature search strategy was aimed to identify relevant studies that address the historical trajectory of bias in AI-powered chatbots and their impact on overall user experience. It involved using the key concepts central to this thesis, (i.e, AI bias, user satisfaction, chatbot bias and bias mitigation), which initially yielded numerous results. As a result, pilot searches were conducted early in the process to narrow down and refine the selection of available articles that would account for terminological shifts and developmental stages of chatbot technology. These searches were essential in identifying missing relevant literature and overly generic terms as well as revealing inconsistent terminology across various disciplines.

Therefore, based on the outcomes, the initial keywords were expanded and refined to include other related terms and synonyms. For instance, “chatbot bias” was initially framed more as “chatbot limitations” in literature concerning early chatbot generations, directly relating to the system’s design constraints, instead of the modern-day socio-technical bias that we now recognise (Xue et al. 2023, 6). So, including word variations like chatbot constraints and chatbot limitations allowed the search to capture foundational work that

could have been overlooked using modern terminology alone. Additionally, these keywords and concepts were combined using the Boolean truth operators “AND”, “OR”, and “NOT” in order to build search queries that controlled potential irrelevant results (Sullivan 2004, 697). Table 1 illustrates the major search terms used in this thesis along with examples of how they were combined using the Boolean operators.

Table 1. Search Terms and Boolean Combinations to Identify Relevant Articles

Major Search Terms	Examples of Boolean Combination Used
AI Technologies Large Language Models AI Chatbots	“Chatbots” AND “Chatbot Limitations” AND “User Perception”
LLM Chatbots Chatbots Chatbot History	“AI chatbots” AND (“AI Bias” OR “Chatbot Bias”)
Conversational Agents Bias	“Bias” AND “Human-Machine Interactions”
AI Bias Chatbot Bias Chatbot Limitations	“Large Language Models” AND “Chatbot Bias” AND “User Satisfaction”
Dialogue Systems Dialogue System Limitations	“User Experience” AND “Chatbot Bias”
Chatbot Constraints User Perception User Experience	“User Perception” AND “AI Chatbots” AND “Transparency”
User Satisfaction Human-Machine Interactions Chatbot User Perception Chatbot User Satisfaction AI Fairness AI Transparency	“AI Chatbots” OR “LLM Chatbots” AND (“AI Bias” OR “Chatbot Bias”) AND (“User Satisfaction” OR “User Experience”)

The combinations listed in Table 1 emerged from iterative refinements during the pilot searches and also reflect the evolving terminology and complexity of literature on chatbot bias and user experience. Including additional terms like AI Transparency, AI Fairness, and Human-Machine Interactions reflected this evolving terminology across different fields (e.g. AI ethics and business) and literature complexity. This makes the process consistent with the thesis objective of mapping the evolution of bias in LLM chatbots and its impact on user satisfaction.

2.2.2 Literature Search Process

Following the process illustrated in Figure 1, the literature search strategy was then implemented to collect articles published between January 2014 and March 2025 across three databases, namely Scopus, Google Scholar and Volter. These databases allowed access to a sizeable assortment of publications and their original sources while also supporting the keyword-based queries shown in Table 1. Scopus is a comprehensive abstract and citation database, featuring peer-reviewed scientific and technical literature including journals, books and conference proceedings. Volter on the other hand is a database owned and operated by the University of Turku that offers access to e-books, books, and academic journals. Then Google Scholar is a free search engine that indexes scholarly literature. The use of these databases was done to minimize publication bias and capture a wide range of both peer-reviewed and emerging literature relevant to this thesis.

The first phase of the search yielded a total of 434 articles, which were then screened and assessed against the predetermined inclusion and exclusion criteria highlighted in section 2.2.3. When considering the ideas presented by Hiebl (2023), the chosen sample size reflected a pragmatic approach to balance time constraints with academic rigor and feasibility with a sufficient breadth to address the research objectives of this thesis. After the screening and eligibility assessment, 89 articles were retained for the final literature corpus. Table 2 presents the sources of the final article selection, broken down based on publication outlet

Table 2. Final Literature Selection Sources (n = 89)

Database Names and Sources	Number of Articles
Peer-reviewed journals (e.g. MDPI, Elsevier, IEEE Xplore, ACM DL, Taylor & Francis, etc)	48
Conference Proceedings and workshops	12
Academic repositories (e.g. arXiv, osf.io, etc)	19
University or Institutional repositories (e.g. UTUPUB, Iowa State University Digital Press etc)	4
Open-access platforms & trade/professional outlets (e.g. AI Magazine, Royal Society Open Science, The Regional Tribune)	6

While most of the selected works were retrieved across the three core databases, the final list includes articles and publications from peer-reviewed journals, professional associations and workshops, open-access platforms and institutional repositories. Listing the source titles of these articles ensures full transparency while illustrating the diverse landscape represented in the review—ranging from computer science to business and ethics. This breakdown also reflects the nature of the topic, by emphasizing the need to capture different perspectives on bias, transparency, and user satisfaction in AI-powered chatbot systems.

2.2.3 Inclusion and Exclusion Criteria

A clear set of inclusion and exclusion criteria was developed based on the research objectives of this thesis to ensure the selection of the most relevant studies. The criteria were established prior to the screening phase to maintain consistency throughout the review process. As seen in figure 2, the studies selected for this thesis included articles with direct empirical research on chatbots, their potential biases and mitigation techniques as well as how they affect user satisfaction.

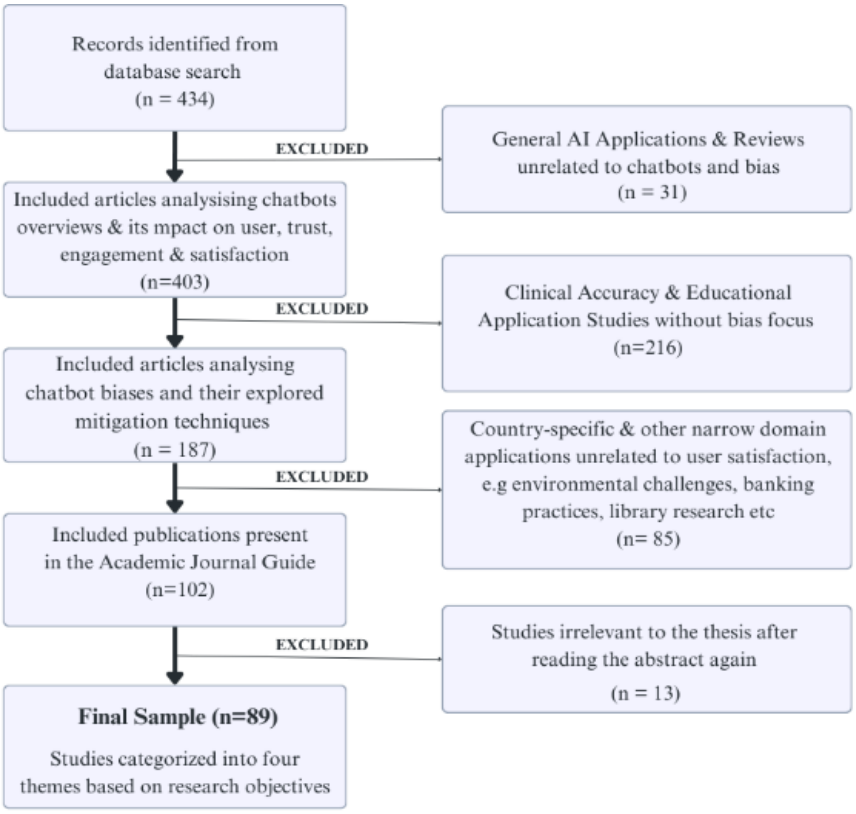

Figure 2. PRISMA Flowchart (modified from Page et al. 2021, 8)

On the other hand, articles that explored more general AI discussions—including user adoption (in relation to educational applications) and technical implementation (in relation to clinical accuracy)—without bias analysis were excluded. Studies that were country-specific and/or had narrow domain applications that did not address broader bias identification or mitigation techniques were also excluded. Although they are valuable in their own right, narrow studies often lack generalizability, and some of the discovered insights may not have been transferable given the thesis's context. Therefore, excluding these studies would focus the review on integrative findings, which would ensure methodological focus while enhancing the depth of the analysis.

The decision for this, aligns with standard practices in systematic literature reviews discussed by Egger et al. (2022, 20–22) where analytical depth and quality of insights would outweigh the breadth of coverage. In other words, considering the obvious increase in literature around the research area, keeping a manageable number of carefully vetted studies would ensure key themes (i.e. the research objectives for instance) are comprehensively understood. All in all, this methodical process is the foundation from which a comprehensive thematic analysis is conducted.

2.3 Thematic and Data Analysis

This section presents the strategy used to explore how biases in LLM-based chatbots affect user satisfaction, and it was done in two phases. The first consisting of creating a generational Evolution Mapping Strategy to contextualize the progression of chatbot technologies over time and identify potential corresponding shifts in bias-related issues. While the second is the Thematic Analysis which is carried out to categorise the insights drawn from the selected literature, culminating in a set number of themes. Therefore, the three subsections that follow outline the evolution mapping strategy, the rationale for the thematic organisation, and then how the discovered thematic overlaps were interpreted.

2.3.1 Evolution Mapping Strategy

Based on the concept of a systematic mapping study discussed by Petersen et al. (2008, 2) and Salama et al. (2017, 252), this research tracked and examined how bias in LLM-chatbots have evolved. That said, the aim behind this section is to categorize chatbot models into generational phases that would aid in identifying potential trends in how biases

may have evolved as well as their impact on user satisfaction. Using the historical perspectives explored by Caldarini et al. (2022, 2–3), Xue et al. (2023, 3–4), Naik et al. (2024, 6–7) and Al-Amin et al. (2024, 5–16) lead to the identification of four different generational phases.

Figure 3 provides a visual of the proposed generational phases, from the early 20th century till today's modern language models. It is important to note that the figure does not encompass every chatbot that has been developed due to the decentralized and rapidly evolving nature of chatbot development (Greyling 2024). However, looking past historical contexts, these phases were intentionally identified to serve as a framework for comparing bias types and mitigation strategies aligned with each technological paradigm. Grouping chatbot development into these phases facilitated the identification of clearer insights into how and why certain biases persist, diminish or emerge over time.

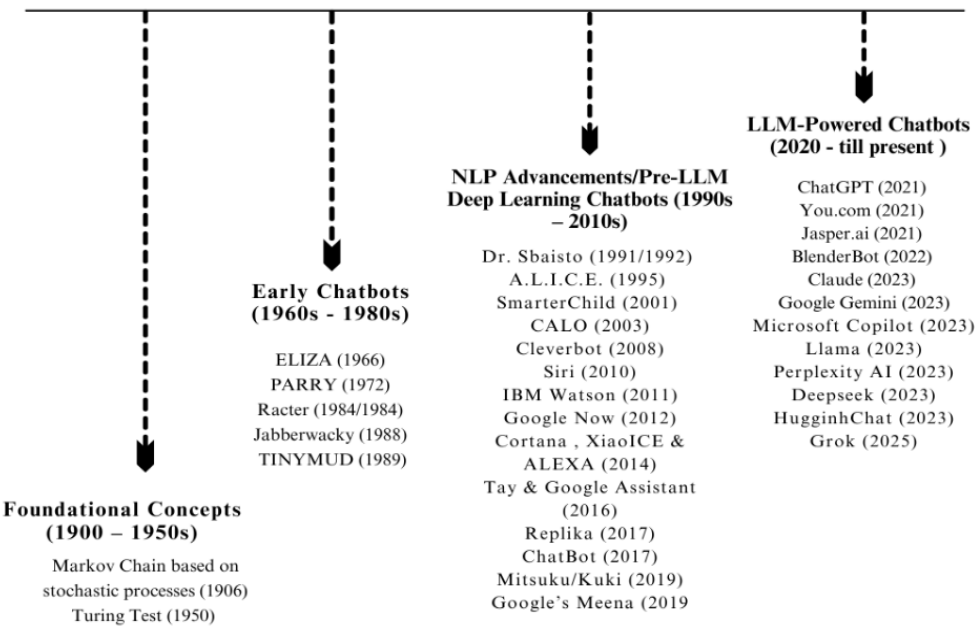

Figure 3. Chatbot Timeline

The first phase explores the *foundational concepts of chatbots* from *1900* till the *mid-1950s*, which were simple statistical models that served as the earliest forms of what would eventually become chatbots (Al-Amin et al. 2024, 5). The described models of this period are Alan Turing's Imitation Game and the Markov Chain (Xue et al. 2023, 3; Al-Amin et

al. 2024, 5–6). The second phase starts in the *early 1960s* and continues till the *late 1980s*. They are otherwise referred to as the *early chatbots*, which had a limited understanding of complex languages, and could only communicate through predetermined keyword matching programs created by their programmers (Caldarini et al. 2022, 2; Xue et al. 2023, 3; Naik et al. 2024, 6; Al-Amin et al. 2024, 6–8). Common examples of chatbots at this stage are ELIZA and ALICE/A.L.I.C.E.

The third phase is centred around *Natural Language Processing (NLP) Advancements* and *Pre-LLM Deep Learning Chatbots***,** starting in the *early 1990s* and ending in the later from possible interactions and using context matching to generate appropriate responses (Xue et al. 2023, 4). Finally, the fourth and current phase is focused on the *LLM-Powered Chatbots* that began being released in *2020*. These chatbots are pre-trained on large datasets and are capable of providing rich responses across a wide range of areas (Xue et al. 2023, 4; Naik et al. 2024, 7). One prime example would be OpenAI's ChatGPT.

Categorizing chatbot development into different generational phases created a bias evolution map (see figure 3), that supports the research objectives by tracing how chatbot design has shifted from rule-based logic to probabilistic and generative models (Schwartz et al. 2022, 20-22). It also supports how the shift has affected the nature of bias, bias mitigation strategies and the challenges of building and maintaining user satisfaction (Schwartz et al. 2022, 1-3). have handled or failed to handle bias over time. The historical overview of chatbots and its relevance to bias analysis is explored more in depth under Section 3.1.1.

2.3.2 Rationale for Thematic Organization

While the evolution mapping provides a way to understand how biases in chatbot systems have emerged and transformed across different generational phases, it is equally important to examine the content and focus of existing scholarly discourse. To that end, a thematic analysis was conducted to synthesise the literature and uncover the dominant patterns and theoretical frameworks associated with bias in LLM-powered chatbots. Thematic analysis is a rather prevalent qualitative research method used to identify, analyse, and report themes (or patterns) within data sets (Braun & Clarke 2012, 58). This study employed the use of this analysis in hand with the PRISMA guidelines to categorise and synthesise insights from the final selection of literature. As a result, the final literature corpus consisted of 89

articles, which aligned with four dominant themes. Each of the themes themselves were based on the combination of the research objectives and the analysis of the selected articles. This therefore ensured a logical flow that supported the structure of the literature review conducted in this thesis.

The development of these themes was guided by the core objective of this thesis, to map the historical trajectory of bias in LLM-powered chatbots, and how these biases have impacted user satisfaction. During the thematic analysis, a set of recurring focal points arose across multiple articles and aligned with the aim of the thesis. Grouping these findings into four themes enable a structured analysis of the topic while, also capturing the lifecycle of bias in AI chatbots and ensuring that both human-centered and technical dimensions are addressed.

The first theme, "*Bias Identification and Evaluation*" was focused on articles that detected, measured and analysed bias in AI systems, specifically in LLMs and AI/LLM chatbots. Additionally, these articles either explored the historical progression of chatbots and their potential limitations and/or biases found, or dove straight into categorizing different types of bias. The articles that distinguished the different types of bias also provided ways to audit and benchmark the prevalence of bias across various models. Building on this, the second theme, "*Bias Mitigation Techniques*" went into the various approaches that have been or that are currently being explored to mitigate and/or reduce bias in AI and LLM chatbots. The techniques that were covered here range from adversarial learning to counterfactual evaluation and quantum-inspired approaches.

Then the third theme, "*User Trust, Satisfaction, and Experience*" was centred around how biases may or may not have affected user experience overtime and how user trust and satisfaction could potentially be improved in chatbot interactions. The articles under this theme examined how user psychology, their interaction patters and the algorithm design principles may additionally influence their acceptance of AI chatbots. Finally, the fourth theme, "*Ethical and Theoretical Frameworks*", consisted of articles that discussed the evolution of ethical considerations in AI chatbots development, and the role of conceptual frameworks in guiding responsible design, communication and bias mitigation across the chatbot lifecycle. Table 3 is a summary of the themes that are identified in this thesis.

Table 3. Summary of Identified Key Themes

Key Themes	Summary
Theme 1: Bias Identification and Evaluation	Focused on articles that detected, measured and analysed bias in AI systems, specifically in LLMs and AI/LLM chatbots
Theme 2: Bias Mitigation Techniques	Focused on various approaches that have been or that are currently being explored to mitigate and/or reduce bias in AI and LLM chatbots
Theme 3: User Trust, Satisfaction, and Experience	Focused on how biases may or may not have affected user experience overtime and how user trust and satisfaction could potentially be improved in chatbot interactions
Theme 4: Ethical and Theoretical Frameworks	Focused on examining ethical frameworks and their role in shaping user trust, satisfaction, and bias mitigation

The frameworks in these articles provided theoretical foundations on understanding the evolution and implications of biased language models.

Although these articles were originally intended to be classified into a single, more dominant theme, several of them appeared to overlap with at least one second theme. In line with the ideas presented by Beattie et al. (2022), Alhajjar & Bradley (2022), Navigli et al. (2023), Xue et al. (2023), and Aninze (2024), bias identification often leads to varied mitigation strategies (depending on the context), which in turn could create ethical dilemmas when considering how users may be affected and/or react. Consequently, those with more than one theme were noted to have combined focuses on multiple aspects of bias, their potential mitigation strategies, user experience and perceptions, as well as overall AI ethical considerations.

2.3.3 Interpreting Thematic Overlaps

To identify and organise the selected articles into their potentially corresponding themes, they were each manually reviewed considering their titles, abstracts, and findings. As previously mentioned, some articles clearly aligned with a single theme, while the vast majority exhibited overlaps. Below, table 4 highlights the number of articles that correlate

with a particular theme or combination of themes, thus capturing the extent to which the selected articles intersect across the four themes.

Table 4. Article Distribution Across Themes and Their Overlaps

Themes and Their Overlaps	Number of Articles per theme/overlapping themes
3	31
4	5
1,2, 3 & 4	6
1&2	11
1&3	5
1&4	10
1,2 & 4	3
2&4	4
3&4	13
2,3&4	1

The third theme "*User Trust, Satisfaction and Experience*" was notably the most standalone theme with 31 articles focusing solely on how chatbots (biased or not) affect users when they interact. However as seen in the table, the general consensus of the literature is that it engaged with multiple themes simultaneously. The other more common overlaps shown in the table are Themes 3 and 4, Themes 1 and 2 and Themes 1 and 4, further supporting the notion on how current research in this domain can span multiple thematic areas as studies increasingly (inherently or not) engage with overlapping domains rather than isolated themes.

For instance, there is a notable connection between user satisfaction and ethical considerations (themes 3 and 4) where research conducted by Xie et al. (2024) implies that these two domains are intertwined. Users are more likely to be satisfied with chatbots that are transparent in regard to their capabilities and/or limitations, respect and protect their, and avoid misleading or even manipulative responses. Additionally when considering the themes related to bias identification (theme 1), mitigation techniques (theme 2) and ethical considerations (theme 4), there is another overlap here because recognizing bias in AI systems is considered to be the first crucial step towards mitigating it (Aninze 2024, 45).

Overall, the literature surrounding these chatbots are multifaceted but interdependent in nature. The themes highlighted in this section have the tendency to converge one way or the other, therefore reflecting the complexity of building and maintaining fair, user-sensitive systems.

2.4 Evaluation of Study

The thematic analysis discussed above has offered a foundation to evaluate the quality and contribution of the selected studies. This section evaluates the systematic literature review on the evolution of bias in LLM-chatbots and its impacts on user satisfaction within the broader framework of the evaluation research. The analysis draws on insights from the Evolution Mapping Strategy and the Thematic analysis in order to reflect on the strengths, encountered limitations, and likely implications of the study's findings. Therefore, this process illustrates a form of disciplined and systematic inquiry that assesses an object, program, practice, activity or system to provide actionable information for decision-making (Kellaghan 2010, 150).

The Critical Appraisal Skills Programme (CASP) Framework is used to guide this process due to its widespread adaptability and acceptance for novice researchers conducting qualitative evidence synthesis (Long et al. 2020, 33). The framework provides a structured but flexible set of questions that align well with the goal of this thesis, permitting evaluation of methodological rigor, research transparency and overall relevance. As a result, five core areas are examined based on the original ten core questions posed as part of the framework. These areas are research objective clarity, methodology appropriateness, data collection and analysis, results and interpretation, and relevance of study.

The research objective clarity is clearly seen where the major aim behind this thesis is to examine the evolution of bias in LLM-chatbots and how this evolution may influence user satisfaction. This involves not only assessing how well the literature review met its aims (especially from an ethically methodological standpoint) but also how it may inform future research, responsible AI development, and policy design related to chatbot systems. This aligns with the CASP's focus on transparent and well-justified research aims. However, the type of bias that is being examined is not explicitly predefined in the research objectives which could cause confusion in terms of what the study is actually trying to uncover, or how the notion of bias itself is being scoped. Nevertheless, they do eventually emerge

inductively through the thematic synthesis of literature which in a way reflects the complex and ever-changing nature of bias in LLM systems, not just chatbots.

The objective clarity then connects to the methodological appropriateness of the thesis. In this case, a key segment to begin with would be the representational fairness in the literature selection process, a key concern when conducting a systematic literature review. Although systematic inclusion and exclusion criteria were used in selecting the final literature corpus, it is important to acknowledge that the selected literature may reflect some varying levels of representational and/or publication bias. The 89 articles are primarily dominated by Western-sourced studies and the English language. This gives room for the possibility of publication bias where research from non-English-speaking contexts or maybe even underrepresented geographic regions are practically invisible in the final corpus. Chatbot bias itself can manifest differently across cultural and linguistic contexts (Cabrero-Daniel & Sanagustín Cabrero 2023, 2; Narayan et al. 2024, 18 & 22), therefore the current literature may not fully reflect global diversity of user experience, but instead a homogenized view that primarily reflects Western contexts.

Therefore, findings that were related to user satisfaction as well as trust and engagement may have unintentionally prioritized culturally specific interpretations of what constitutes as “fairness” and “bias”. Studies conducted in other regions could have revealed different emergent forms of bias (such as bias tied to infrastructural exclusion), alternate user expectations or varying interaction patterns. Such nuances are likely underrepresented, thereby hindering this thesis’ ability to generalize conclusions across all user populations. Addressing this limitation in future research would likely promote digital inclusivity across chatbot systems while also building an equitable understanding of the effects bias has on different communities.

Then there is the data collection and analysis aspect which looks at another key segment being level of transparency of the methodology, particularly when it comes to conducting the systematic review and assessing the tools used for data visualization and theme representation. In this case, the search strategy could contain inherent limitations due to specific keyword dependencies and the scope of the used databases. Aside from this, the selection process itself is as time-consuming as it is error-prone (van Dinter et al. 2021, 1-2). Consequently, this could have introduced gaps in the coverage of keyword effectiveness

across the three databases that were used in this thesis as well as subjective judgements and/or interpretations when it came to finding the themes and eventual categorization.

Furthermore, the use of AI assistance, in this case for checking sentences for grammatical errors, simplifying complex explanations and also to generate an UpSet plot (although this idea was eventually scrapped in favour of a table), adds a dimension to the ethical discussion. The idea of using this form of visualization in the first instance stemmed from searching for an effective way to represent the thematic overlaps found during the data analysis. The Venn Diagram had proven ineffective because of the number of overlaps between the themes. In that wise, varying studies highlighted the effectiveness of UpSet Plots in displaying complex set interactions (Lex et al. 2014, 1984). Therefore, an alternative attempt to visualize the thematic overlaps was done using an UpSet plot created with the help of Anthropic AI (i.e, Claude). However, the static nature of the generated UpSet plot was insufficient in conveying the distribution of the 89 manually coded articles due to its lack of clarity.

As a result, a deliberate decision was made to present the overlaps in tabular form, as seen in Table 4, offering a clearer way to display the findings. This instance raises the question of how AI tools may or may not shape research interpretation even when they are not directly involved in data analysis. For example, while AI tools were not used to generate the themes, they could still skew explanations due to their influence on tone or emphasis of a particular explanation. Karjus (2025) argues that even neutral applications of AI, such as text correction and refinement, could reflect underlying biases embedded in training data. This highlights the importance of sustained human oversight, especially when integrating AI into traditionally objective tasks.

Following the analysis, there was also the results and interpretation aspect. To build a better understanding of how bias has evolved overtime, the evolution mapping strategy was used to categorize key periods in AI-chatbot history. The four generational phases allowed bias to be conceptualized in relation to different chatbot architectures and training styles, thus revealing different types of biases that have emerged and remained persistent across the different phases. Although, intersectional (compounding effects of race, gender, disability, etc) as well as community-specific biases is barely (if completely not) addressed in the eventual analysis and findings of this thesis. The ethical implications of using these chatbots are also discussed, as well as how it can be implied from Xie et al.'s (2024) study

that such implications are intertwined with user satisfaction. This therefore suggests that ethical principles related to the design and deployment of chatbots are a basis to how users interact and perceive AI systems. This could be something further explored in future research across varying domains if it has not already been fully explored.

The final aspect of this section is the relevance and contribution of this thesis. A major part of this section would be the importance of balancing stakeholders' perspectives within the review literature. It must be said that this thesis leaned more towards the user satisfaction, psychological response and trust, which is undoubtedly vital, but it could have been interesting to also pay attention to the developers, businesses and regulators perspectives as there is less recognition in the literature corpus. Although the developers are discussed, they are seen more as one of the root causes of bias in these systems, even though studies have revealed some users may deliberately exploit chatbots to produce odd or offensive content, as was the case with Tay (2016).

This imbalance could therefore inadvertently frame chatbot bias to only be a matter of user perception as opposed to the broader issue influenced by other dynamics, imperatives or even regulatory structures. Economic incentives, government models and developers' intent were underexplored in this thesis, which means that insights into institutional contributors to bias was mostly absent. Another part is with respect to the growing sophistication of bias mitigation strategies, which appear to mostly be evaluated in terms of computational performance. On top of this, there appeared to be insufficient ethical considerations on some of the strategies. Questions like who benefits from a strategy and who is left out or what constitutes a "successful" strategy is rarely fully addressed beyond the strategy's ability to reduce quantifiable bias. This implies that bias mitigation strategies need to be accompanied by some form of ethical framework in order to make a long-term transparency and inclusive impact, even if this may not necessarily be true.

With all this in mind, it is also vital to consider the researcher's position, and the tools employed all through the study. Admittedly the study attempted to achieve objectivity through transparent documentation and categorization, however, interpretations and decisions were influenced by the researcher. The use of AI tools like Claude and ChatGPT to assist with grammatical corrections and simplifying complex concepts and/or language did not directly influence data synthesis or theme development, but it could have still added a subtle interpretive influence (see appendix 3 for the Artificial Intelligence Assistance

Declaration). Although this will be explored in the subsequent sections, LLMs and other AI tools often reflect dominant linguistic norms and assumptions which could have unintentionally altered the tone or nuance of some ideas. The risk of this itself is minimal, but it reflects a certain importance attached to reflexivity when using said tools, no matter the context.

Having established the methodological approach and overview of the thematic analysis of this thesis, the following section presents the key findings that emerged during said analysis.

3 Findings

Following the research design and themes outlined in the previous chapter, this literature review synthesizes the notions and ideas presented in the selected articles on bias, user trust, and ongoing mitigation strategies within LLM-powered chatbots. Therefore, the general structure of this aspect of the thesis is based on the four themes highlighted in section 2.4.

3.1 Bias Identification and Evaluation in Chatbots

From the LLM perspective, the term bias is defined as the presence of systematic misrepresentations, attribution errors, or factual distortions that result in favouring certain groups or ideas, perpetuating stereotypes, or making incorrect assumptions based on learned patterns (Ferrara 2023, 2-3). That said, biases could manifest either subtly or in overt ways, thus likely affecting user satisfaction and perceptions of credibility. This section aims to explore the history of chatbot technology, then map out the historical trajectory of bias, before reviewing a comparative analysis of bias evaluation benchmarks.

3.1.1 History of Chatbots

Having been in development over the last fifty odd years, chatbots have reached a point where they are capable of demonstrating a greater capacity to understand and respond to questions effectively and quickly (Xue et al. 2023, 2). Although, in spite of these significant advances, concerns regarding algorithmic bias in LLM-powered chatbots have intensified, as recent studies have revealed how they impact varying aspects of chatbot design, functionality and overall user interactions (Xue et al. 2023, 6; Zhou 2024, 86).

In view of this, understanding the “why” and “how” behind something can provide valuable insights that inform how that thing is approached and managed. Consequently, to address these concerns regarding modern algorithmic bias, it is essential to first understand them within the broader context of the historical context of chatbot technologies. Tracing said history would enable a deeper appreciation of how biases have emerged and evolved over time as a potential result of their increased algorithmic complexity and training data (Beattie et al. 2022, 121–122; Xue et al. 2023, 6). Ultimately, looking at bias from this perspective could provide insights that guide ongoing and likely future efforts made towards bias mitigation in chatbot design and usage. This now leads to a more in-depth

analysis of the chatbot historical trajectory, which had previously been divided into four major phases. These phases were established based on existing literature, to simplify the key turning points in chatbot history.

The early 20th century (1900s-1950s) established the foundational concepts and groundwork for a wide array of technological advancements, including chatbot development. Most historical overviews on chatbots start with Alan Turing's 1950 test, otherwise referred to as "Turing Test", where a machine is assumed to be intelligent if it is able to indistinguishably mimic human communication (Xue et al. 2023, 3; Al-Amin et al. 2024, 6). However, Al-Amin et al. (2024, 5) explored an earlier foundational concept known as the "Markov Chain", developed by the mathematician Andrey Markov in 1906. Markov chains are simple statistical models based on stochastic processes which predict outcomes using previously observed data (see Almutiri & Nadeem 2022, 6). They are also still being used today in various NLP tasks such as text generation/prediction and part-of-speech tagging (see Almutiri & Nadeem 2022, 6). Although it is not explicitly mentioned, these two concepts placed together, based on my own interpretations, seem to have significantly shaped early chatbot technology.

That said, these foundational concepts were the base in which the early chatbots (1960s-1980s) emerged. More specifically, it was the Rule-Based chatbots era. They operated by pinpointing keywords in user queries and matching them against a set of their own predefined rules in order to generate responses (Xue et al. 2023, 6). Joseph Weizenbaum's ELIZA is considered to be the earliest form of a (rule-based) following its invention in 1966. It was programmed to imitate a therapist's open-ended questioning and mirror back perceived suitable responses to users, giving the illusion of understanding (Al-Amin et al. 2024, 6-7). People anthropomorphized it, even though it lacked real intelligence due to its limited knowledge base (Xue et al. 2023, 3). Nevertheless, ELIZA inspired subsequent research, thus leading to the eventual creation of PARRY by Kenneth Colby in 1972, Racter by Chamberlain and Etter in 1983, Jabberwacky by Rollo Carpenter in 1988 and TinyMUD by James Aspnes in 1989. Although these systems relied solely on rule systems and programming techniques, they each introduced unique innovations that established conceptual foundations for modern chatbots and language models.

Further down the timeline, the 1990s marked significant advancements in Natural Language Processing, which eventually evolved through the 2010s with the emergence of

pre-LLM Deep Learning chatbots. Still, it is important to note that the start of this period (1990s-2010s) was heavily influenced by the rapid expansion of the internet (Al-Amin et al. 2024, 9). Chatbots of this era began as retrieval-based ML-powered chatbots, that simulated human conversation by learning from user interactions, building a keyword-based repository, and using context matching to generate appropriate responses (Xue et al. 2023, 4 & 6).

Taking nearly three decades worth of chatbot innovation, Richard Wallace created the "Artificial Linguistic Internet Computer Entity", or A.L.I.C.E. for short in 1995 (Xue et al. 2023, 3; Al-Amin et al. 2024, 9). A.L.I.C.E. was a more advanced version of ELIZA, that used pattern matching techniques through Artificial Intelligence Markup Language (AIML) to imitate conversations (Al-Amin et al. 2024, 9). Shifting towards the 2010s, virtual assistants (such as Apple's Siri in 2011) and companions (such as Replika in 2017) became more popular. Systems developed at this point had IQ and EQ modules incorporated into their algorithms, making their responses to user queries more flexible and advanced (Xue et al. 2023, 4).

The culmination of these advancements then led to LLM-Powered Chatbots (2020 till present) which are now pre-trained on large datasets (Xue et al. 2023, 4). This current generation has drifted away from symbolic AI approaches and towards large scale natural language understanding and generation, permitting them to provide rich responses across a wide range of areas (Al-Amin et al. 2024, 15-16). Xue et al. (2023, 4) and Al-Amin et al. (2024, 16) as well as other scholars (Lippens 2024; Chan & Wong 2024; Babonnaud et al. 2024) further discussed how modern chatbots such as, ChatGPT, Gemini, Claude, Llama, amongst others, also possess large parameter scales, self-supervised learning capabilities, transformer architectures enhancing text processing and multimodal capabilities. These developments allow modern chatbots to reason, summarize, explain and independently generate content based on user input.

3.1.2 Bias Evolution Mapping: Persistent vs. Emergent Forms of Bias

The established historical context of chatbot development has shown the rising complexities of these systems and their pervasive integration into various aspects of everyday life (Xue et al. 2023, 4). Despite these advancements, there have been concerns regarding the potential negative impacts of bias on individuals and society (Ferrara 2024,

4). Various studies (Huang et al. 2019; Xue et al. 2023; Lippens 2024; Babonnaud et al. 2024; Zhou 2024; Urman & Makhortykh 2025) have explored how biases are able to either persist across chatbot generations or unexpectedly emerge as a result of a system's increased complexity.

Consequently, Xue et al. (2023, 6), and Ferrara (2024, 2) have recognized that chatbot biases (or generically in AI systems) can stem from three major sources; the data they are trained on, the algorithms that process the data, and the humans (designers or users) who interact with these systems. Differentiating which biases are persistent and which have emerged could play a role in overall effective management of these issues. Therefore, this section maps out the evolution of chatbot biases. This approach clearly indicates which biases have persisted across chatbot generations, identify new biases that have emerged as well as why and when they first became prominent.

That said, as the notion of AI was still new and mostly hypothetical between the 1900s and 1950s, this era was predominantly mathematical and relied mostly on stochastic processes as seen in the case of Markov Chains (Al-Amin et al. 2024, 5). Despite the fact that this is not explicitly mentioned in the literature used for this thesis, nor in Al-Amin et al.'s (2024) work, it can be assumed that biases may have arisen from early data sampling methods which would have directly affected probabilistic models and statistical inference methods (see Li & Zhang 2009). This is assuming the data sampling methods might not have represented varying scenarios or groups properly, thus leading to skewed outcomes. Additionally, Markov chain and other statistical models could contain inherent mathematical biases due to model assumptions, finite-sample effects, amongst other such factors (see Dette & Melas, 2010).

Moving towards the 1950s, the Turing Test is also subject to a few biases such as deeming human language as a definitive measure of intelligence (anthropocentric bias), reliance on subjective human judgement, and deception-based nature (a machine trying to deceive a human) amongst a few others (see Adams et al., 2016). In summary, although foundational-era biases do not appear to have been explicitly recognized as what we understand today to be "algorithmic biases", implicit mathematical and theoretical biases could have shaped the earliest stages of AI, which by extension laid out the groundwork for future biases to manifest more explicitly.

By the 1960s, rule-based chatbots relied on predetermined rules and/or content that had been manually created and embedded into its pattern-matching system (Xue et al. 2023, 6; Al-Amin et al. 2024, 6–8). There were no self-supervising learning capabilities at the time, meaning the content relied heavily on human input, and as a result, chatbot responses were limited to what their programmers deemed important or socially acceptable (Xue et al. 2023, 6). In simpler terms, chatbots depended entirely on the content provided by their programmers, which likely reflected their worldviews and own inherent biases.

Therefore, potential biases within these chatbots would have stemmed primarily from human interactions, i.e. designers or programmers (Xue et al. 2023, 6). Figure 4, (illustrating the evolution of chatbot bias), shows that the early stages of AI chatbots are marked by a predominance of *designer/programmer bias*, which spans from the Markov and ELIZA phases through to early learning systems. This indicates a time when and where bias was traceable to clearly defined sources (such as logic rules) which would have been easier to audit despite being subjective. Therefore, detection was relatively straightforward, but mitigation would have required changes to human-coded logic instead of just data-driven correction.

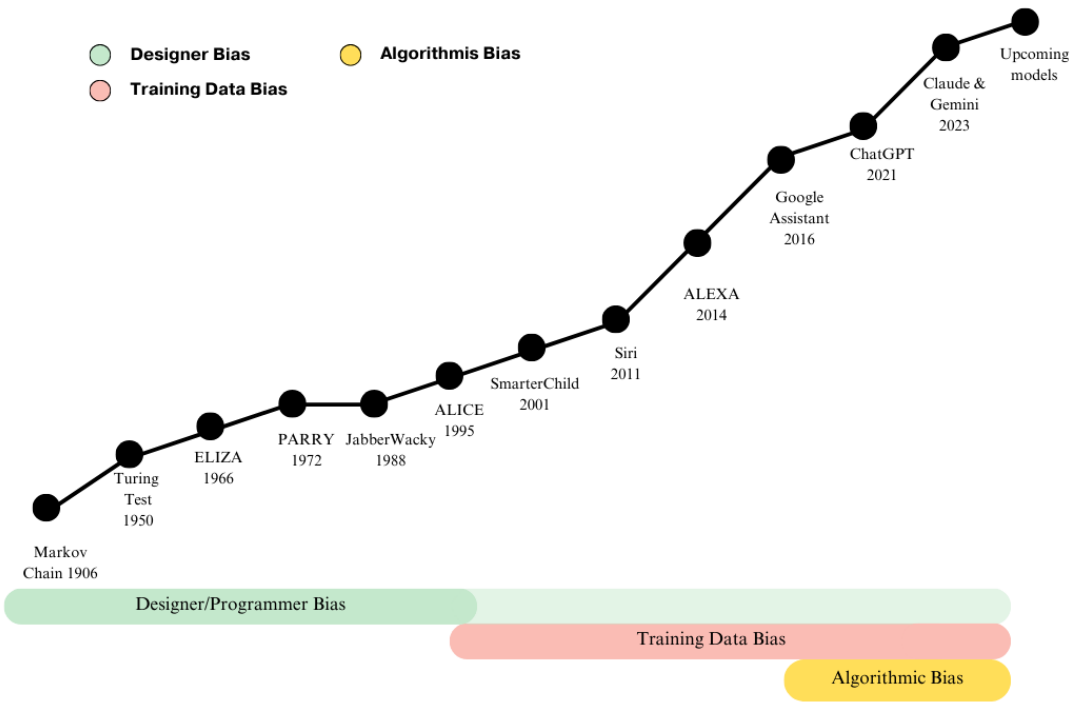

Figure 4. Evolution of Chatbot Bias Types

Retrieval-based ML-powered chatbots became more popular by the 1990s and were still similar to how earlier chatbots functioned, with the exception that they could learn from interactions and training data (Xue et al. 2023, 6). Although these chatbots were also susceptible to designer bias, figure 4 displays a turning point here with the surfacing of *training data bias*. This kind of bias primarily emerged from the data they were trained on. Data bias occurs when machine learning models are trained on incomplete or unrepresentative data (Ferrara 2024, 2). Additionally, with continued advancements in NLP techniques, new methods to reveal other forms of biases emerged. For instance, word embeddings such as Word2Vec showed how societal biases could become unintentionally embedded within language models (see Zhan, 2025). This marked a shift where bias detection now required more advanced diagnostic tools, (such as bias benchmarks or probing tasks) in case rule-based inspection became insufficient.

Finally, following the rise of LLM chatbots in 2020, there has been a significantly growing body of research surrounding the detection and mitigation of bias in LLM technologies (Chan & Wong 2024, 1). Compared to the previous years, LLM-chatbots are able to generate their own responses using ML and generative AI algorithms (Xue et al. 2023, 6). Despite this, biases in LLM-chatbots can majorly be traced from training data (i.e. data bias) and user interactions. Training data could potentially contain biases centred around outdated knowledge, misinformation, stereotypes and even hate speech (Babonnaud et al. 2024,1). LLMs and their related technologies are able to inherently reflect and amplify the biases present in their training data, (Chan & Wong 2024, 2; Lippens 2024, 2) thus causing inappropriate or harmful conversations (Babonnaud et al. 2024,1).

Figure 4 further illustrates that starting in the late 2010's, training data bias and algorithmic bias overlap, highlighting a compound bias environment. Algorithmic bias becomes increasingly dominant as the models' complexity grows, signifying a shift where the models' design becomes an active source of bias. This is further complicated by the black-box nature of these models, causing chatbot responses to become unpredictable, as the biases are more difficult to detect, interpret and thus control (Xue et al. 2023, 6; Ferrara 2024, 2). The evolution map in Fig. 4 thus outlines the historical progression of chatbots as well as the layered escalation of bias detection complexity. Over the years, bias in chatbots have shifted through distinct phases; initially being manually and explicitly or inadvertently encoded, then moving towards being embedded in training data and finally

manifesting as emergent algorithmically complex biases. This path also suggests that mitigation strategies would need to be multi-pronged and persistent in order to address the more complex sides of bias in chatbots.

Human Interaction bias, (biases that stem from how programmers or end-users interact with AI systems) specifically from the designer's perspective, appeared early on but is seen in Figure 4 to diminish overtime, although it does not totally disappear. Based on the articles that directly address sources of bias, designer bias (also known to be a part of the broader human interaction bias) has persisted through architectural choices (how the program is built), training objectives and datasets as they play a role in modern bias formation (Babonnaud et al. 2024, 3–4; Chan & Wong 2024, 1). Furthermore, the context in which an AI system such as the large language model, is deployed can alter how users engage with it, such as if a human intermediary is needed or involved in decision-making (Gallegos et al. 2024, 1107). The overall design of the user interface may also shift how the model's behaviour is perceived, either changing or reinforcing their assumptions, thus contributing to bias through interaction (Ferrara 2024, 3; Gallegos et al. 2024, 1107).

Training Data Bias on the other hand took root in the 1990s and persisted into the 2020s. Ferrera (2024, 2) highlighted how data is biased when it is unrepresentative of diverse cultural and social contexts, a challenge that continues to affect most chatbots. Additionally, modern LLM chatbots introduced a new form of training data bias where algorithmic selection from internet-scale datasets inherently hides certain information from users due to the volume of available information online (Xue et al. 2023, 11; Wang et al. 2024, 12). In other words, this new form of data bias does not allow users to access the remaining information it omits, thus introducing content invisibility and selection bias, which are more difficult to detect, audit or correct (Xue et al. 2023, 11; Wang et al. 2024, 12).

Then as seen in the Figure 4, *Algorithmic Bias* (which results from design and/or implementation choices that prioritize certain attributes over others (Ferrara 2024, 3)) became rampant in the LLM era due to increasingly complex model architectures. The complexity of these models in hand with real-time interactions and the different data sources would hinder the effectiveness of potential mitigation strategies (Ferrara 2024, 7).

Looking at chatbot bias from these perspective shows how it has shifted from being directly programmed in rule-based systems to implicit dataset embedding in LLMs. This has in turn transformed the way bias is analysed from assumably simple rule inspections to sophisticated interpretability research (Chan & Wong 2024). Additionally, chatbot bias has evolved from singular to compound forms in the sense that rule-based systems inherited bias from a single source (their designers/programmers), while modern LLM-chatbots are able to experience biases from different components such as the system's architecture and training data. Although, it does appear that as AI itself becomes more complex, so does the challenge of mitigating potential biases. Therefore, addressing bias, specifically chatbot bias as is the case of this paper, could require both persistent and emergent strategies.

3.1.3 Comparative Analysis of Bias Evaluation Benchmarks

Bias evaluation benchmarks serve a vital role in pinpointing the presence and nature of bias across different models and interactions. This section examines how bias evaluation benchmarks for AI chatbots, especially LLM chatbots, have developed over time and compare a few recent methods that can be used to assess bias. It provides insights into how benchmark design is able to shape bias detection and mitigation strategies.

3.1.3.1 Evolution of Bias Evaluation Benchmarks

Building on the discussion in section 3.1.2, assessing the extent to which a chatbot is biased is challenging and is further intensified by the way generative AI, such as chatbots in this case, are trained on vast and unverified data (Beattie et al. 2022, 120; Duncan & Mcculloh 2024, 687–688). Consequently, benchmarking frameworks have emerged as important tools to analyse and mitigate biases in these systems and to also inform the development and deployment of fair systems (Chan & Wong 2024, 2). Additionally, similar to the way bias in chatbots have evolved, there is a growing consensus that benchmarks must also adapt to the multidimensional nature of bias and fairness (Chan & Wong 2024, 1-2; Huang et al. 2024).

As with the previous sections, this section will attempt to briefly trace the progression of bias evaluation benchmarks in chatbots and examine how four more recent approaches (Babonnaud et al.'s (2024) qualitative auditing, Urman & Makhortykh's (2025) cross-lingual evaluation, Chan & Wong's (2024) BIG-Bench framework, Beattie et al.'s (2022)

subjective scoring and Duncan & McCulloh's (2024) machine learning classifier fit into and advance this trajectory.

The rule-based chatbots from the 60s till the 80s may have prioritized functionality over fairness since their decision-making processes were transparent (based on pattern-matching content) and their responses were controllable (Xue et al. 2023, 6). If bias was detected within these systems, it would have been assessed anecdotally through manual reviews since evaluations at the time were mostly subjective and primarily based on human judgment (Shieber 1994, 1-2). However, this began to change during the Pre-LLM Deep Learning Chatbots/NLP era, although this mostly occurs between 2010 and 2015, around the time data bias emerged.

Bias recognition and evaluation was mostly still informal like previous years, but there was a higher concentration on certain metrics (such as performance and accuracy) as opposed to systematically addressing bias related issues (Blakeney et al. 2021, 2-3). However, over the next five years, neural networks and transformers became a more dominant approach in NLPs, and this new scale and complexity exposed biases, prompting an increased emphasis on bias detection (Wolf et al. 2020, 1-2). This shift led to significant advancements in Natural Language Understanding (NLU) evaluations, with large-scale benchmarks like General Language Understanding Evaluation (GLUE) benchmark standardizing performance assessment (Wang et al. 2019).

In light of evaluations and benchmarks, AI ethics and bias awareness began to gain a significant amount of attention in the 2020s, especially following the launch of ChatGPT in 2022 and Gemini in 2023. Additionally, there were several methodological advancements such as real-world impact considerations (UNESCO 2022), contextual bias assessments (Sheng et al. 2021, 4275–4293) and the integration of qualitative and quantitative evaluation techniques like Fill-in-the-Blanks and Tree of Thoughts (Babonnaud et al. 2024, 195–203). On top of this, multidimensional approaches began to be explored as well as seen in the case with Google's BIG-Bench (Chan & Wong 2024) and Beattie et al.'s (2022, 117–123) Chatbot Bias Assessment Framework for instance, therefore advancing the studies on evaluation frameworks. These developments and advancements laid the groundwork for different kinds of evaluation strategies which will

be examined in detail in the next section, especially in relation to how benchmark frameworks capture and assess bias in LLM deployments.

3.1.3.2 Evaluation Methods and Benchmarks

Following the classification of bias in LLM chatbots based on their evolution patterns (i.e persistent versus emergent) and their sources, (i.e. algorithmic, training and human interaction), this section reviews how biases are evaluated using contemporary benchmarks. Guo et al. (2024, 9) discussed how evaluating bias, particularly in LLMs, requires a multidimensional approach that reflects the sociocultural as well as technical complexities inherent in these technologies. They further argued that evaluation process itself is as much an interpretive task as it is a technical one, thus needing continuous adaptation to various changing contexts.

Cravotta (2003, 57), defined a benchmark is a point of reference that can be used to measure and compare the value and/or quality of two or more similar alternatives, in the case of this thesis, it is in terms of bias and fairness in LLM chatbot performance. The papers for this analysis offered both quantitative and qualitative perspectives on bias evaluation methodologies as well as hybrid frameworks and machine learning classifiers. They reflect the multidimensional nature of the concept in addition to the evolving nature of bias in response to user engagement, model updates and sociocultural shifts.

First, Babonnaud et al.(2024) introduced a qualitative auditing approach that was designed to detect prejudice/biases by using simple prompts without directly soliciting harmful content (Babonnaud et al. 2024, 195). It was designed to be a form of ethical evaluation for LLMs. Their methodology had a two-pronged approach. The first, known as the self-assessment stage, involved exploring three distinct techniques used in various aspects of artificial intelligence to understand the rationale behind LLMs' outputs and to find potential prejudices/biases.

The first technique was the *Fill-in-the-Blanks (Sect. III-A)* where LLMs were tasked with completing sentences with missing words, after being trained on a list of predefined subjects (Babonnaud et al. 2024, 3). Next was the *Contextual Attribute Swap (Sect. III-B)* which analysed how sensitive and/or flexible the LLMs were by changing key attributes of a character and observing the changes in the LLMs' response (Babonnaud et al. 2024, 197). The third technique was the *Tree of Thoughts (Sect. III-C)* which was used to trace the

LLMs' reasoning and assess how biases propagate through multi-step responses (Babonnaud et al. 2024, 198). Once they concluded analysing the techniques they moved on to the second which was the *Human Auditing Stage*, which involved them (the authors) conducting a qualitative evaluation using standardized guidelines (Babonnaud et al. 2024, 197-198). Their findings revealed deeply ingrained stereotypes and prejudices in LLM outputs, especially in regard to gender, cultural and racial biases. However, the approach appears to be vulnerable to interpretive subjectivity as well as being labour-intensive since an individual would be required to manually assess the system's outputs. The scope is also narrow as it is focused on a particular subset of minority identities.

Next, Urman & Makhortykh (2025) on the other hand, adopted an open-ended questioning approach conducted in English, Ukrainian and Russian to see if LLM-chatbots were prone to political bias when responding to politically phrased prompts (Urman & Makhortykh 2025, 9). Compared to traditional benchmarks, their approach allowed greater flexibility in detecting potential biases, context-driven evaluation instead of the conventional multiple-choice assessments and cross-linguistic analysis which highlighted inconsistencies in the models' behaviours across the different languages and geopolitical narratives. Their approach overall revealed how rigid predefined benchmarks are becoming insufficient for bias detection, therefore emphasizing the need for more adaptable methodologies. Nevertheless, there are also drawbacks in the method's lack of standardization and difficulty in quantifying findings, which in turn limits reproducibility and scalability.

Chan & Wong (2024) on the other hand introduced a hybrid evaluation framework that combined both quantitative and qualitative measures in order to assess effective benchmarking practices as well as LLM model bias and fairness (Chan & Wong 2024, 1-2). They did this by using the Google BIG-Bench Benchmark, (a tool that assesses bias through real-world scenarios across various disciplines), as the primary evaluation tool. Then they paired this tool with the quantitative scoring metrics to systematically measure fairness and the qualitative assessments to complement the numerical bias detection and address potential gaps that had been overlooked by the quantitative metrics. While BIG-Bench is a valuable tool, the study highlighted a few limitations, specifically the tool's likelihood to overlook context-dependant but subtle biases and its reliance on predefined tasks that may not fully capture bias manifestations (Chan & Wong 2024). In light of this,

their study also emphasized the need for adaptable methodologies by advocating for the integration of underrepresented languages and cultural perspectives into future benchmark development.

Prior to this however, Beattie et al. (2022) focused more on exploring whether AI chatbots are capable of learning bias and if the bias they learn could be measured and thus mitigated (Beattie et al. 2022, 119). Their methodology involved developing the "Chatbot Bias Assessment Framework" that categorized chatbot responses into predefined bias classes and then AI response variability through non-deterministic evaluation (Beattie et al. 2022, 117). Yet, while it is helpful in addressing variability, the classification framework could risk oversimplifying complex and overlapping bias categories (Beattie et al. 2022, 122).

Finally, Duncan & McCulloh (2024) developed a machine learning classifier to detect political bias in LLM outputs. Their approach involved using Multinomial Naïve Bayes and SVM models to classify text data, splitting their text data (80% training / 20% test) to assess the degree of political leaning/bias, and then evaluating ChatGPT-4 responses, which revealed a higher proportion of liberal-leaning/biased responses (75%) (Duncan & Mcculloh 2024, 690). Their findings supported existing literature covered in this paper regarding the influence of training data on LLM outputs (i.e. data bias). They also highlighted the potential for algorithmic amplification of ideological biases, and in response to this, also emphasize the need for bias transparency as well as user awareness.

Table 5 displays a brief overview of the different bias evaluation approaches used in Large Language Models (LLMs) discussed above. Each study is categorized based on their approach, key strengths and limitations in order to directly compare each contribution to bias detection.

Table 5. Comparative Insights on Benchmark Evaluations

Study	Approach	Key Strengths	Limitations
Babonnaud et al. (2024)	Two-stage approach: LLM self-assessment. Human evaluation via fill-in-the-blanks, attribute swaps, and Tree of Thoughts.	Sheds light on implicit stereotypes & prejudices in LLMs (such as gender roles, cultural tropes). Avoids adversarial prompting.	Labor-intensive human analysis. Subjective interpretation risks. Present work limited to a specific list of minorities.
Urman & Makhortykh (2025)	Open-ended questioning	Captures nuanced ideological leanings. Reveals cross-lingual disparities/bias detection.	Limited to political contexts. Requires manual thematic coding. Difficult to standardize and quantify.
Chan & Wong (2024)	Hybrid (quantitative & qualitative). Google BIG-Bench benchmark with 360+ tasks	Standardized quantitative metrics. Broad coverage of bias types.	Misses context-specific biases
Beattie et al. (2022)	Conversational bias framework & numerical bias scoring across repeated queries	Accounts for real-world interactions. Mitigates LLM non-determinism via averaging. Quantifies subjective judgments.	Oversimplifies complex biases. Risk of category overlap. Scoring objectivity could be an issue.
Duncan & McCulloh (2024)	Machine Learning Classifier (SVM)	86 % accuracy with SVM on test data. Repeatable and Scalable.	Limited to political bias Potential for dataset bias

Together, the studies in Table 5 demonstrate that a single method to evaluate bias is not enough, as Guo et al. (2024) pointed out since each author employs some varied methodology or hybrid approach to conduct their studies. Also, while standardized benchmarks, are valuable, they should/need to be supplemented in some way by using real-world or context specific assessments, which is especially seen across all five studies. On top of this, there were a few shared limitations that arose across the five studies.

The first limitation is rooted in a noticeable trade-off between scalability and interpretation where qualitative audits such as Babonnaud et al.'s (2024) approach are too time-consuming and difficult to replicate even though they offer comprehensive understandings of latent bias. In contrast, Machine Learning Classifiers, such as the one referenced in Duncan and McCulloh's (2024) study, though efficient and repeatable may narrow their focus to easy classifiable biases over the more subtler ones.

The second limitation is an issue of contextual blind spots and rigid or pre-defined categorizations. In this case, Beattie et al. (2022), Chan and Wong (2024) and Urman and Makhortykh (2025) discussed in their own ways and based on their own methods that predefined benchmarking tasks may overlook or misrepresent varying regional, linguistic, or sociopolitical nuances. This would reduce sensitivity to compound or rising forms of bias that increasingly characterize LLM behaviour. The third limitation, minimal stakeholder inclusion, was also common gap amongst all five studies. In these studies, bias was framed from a data-centric or model perspective, with little to no feedback from diverse user groups. Such an omission could end up prioritizing technical performance over social impact and relevance.

Given these considerations, bias itself is perceived to be dynamic. Therefore, continuous benchmark refinement and interdisciplinary collaboration would be vital to address evolving biases.

3.2 Bias Mitigation Techniques

Although a vast majority of existing research has concentrated on identifying bias in AI systems, including LLM chatbots, their focus has also extended to varying techniques that can mitigate them, depending on their context (Xue et al. 2023, 14; Zhou 2024, 87; Guo et al. 2024,15). This section explores the evolution of these strategies, assesses their effectiveness across the three bias sources, and examines the inconsistencies and limitations that have persisted in spite of ongoing advancements.

3.2.1 Evolution of Mitigation Techniques

Like the biases themselves, bias mitigation techniques have also evolved overtime, reflecting shifting societal concerns on ethics and trust in AI technologies as well as technological advances. From the scripted, rule-based systems of the 1960s, till present day LLM chatbots, various studies have implied and revealed a progression where mitigation techniques have shifted from implicit containment to explicitly structured interventions (Blodgett et al. 2020; Beattie et al. 2022; Ernst et al. 2023; Shokrollahi 2023; Xue et al. 2023; Zhou 2024; Guo et al. 2024; Eisenmann et al. 2024).

In the initial stages of chatbot evolution, bias and potential mitigation techniques were not formalized concepts, which can be seen with the way the earlier foundational chatbots (e.g

ELIZA and PARRY) were designed. For instance, studies show that Weizenbaum's ELIZA, addressed the "problem of context", (the persistent issue of imbuing AI with contextual understanding and/or language) through its restrictive conversational design thereby avoiding deep semantic processing that could give room for biased or controversial outputs (Caldarini et al. 2022, 2; Xue et al. 2023, 3; Al-Amin et al. 2024, 6; Eisenmann et al. 2024, 2719).

Essentially, Weizenbaum's idea was rooted in a rule-based system where potential biases would likely have been contained through careful scripting and surface-level reflection strategies, thus controlling user interactions (Eisenmann et al. 2024, 2719). This method, although rudimentary and indirect, was also an intended public reflection of Weizenbaum's scepticism towards the superficiality of AI communication at the time (Eisenmann et al. 2024, 2720). Nevertheless, ELIZA still marked an early point where bias was more "managed" as opposed to being detected and corrected through the restriction of its conversational possibilities.

This inadvertent technique persisted into the first generation of commercial chatbots (such as basic FAQ or customer service chatbots) that began emerging in the 1990s. Bias management was still the trend during this period, enforced through restriction and predefined retrieval-based domains that would have been deemed to be non-controversial (Xue et al. 2023, 6-7). However, as suggested by Xue et al. (2023), Al-Amin et al. (2024), and Eisenmann et al. (2024), these enforced interventions likely addressed surface level harms and were not designed to tackle deeper concerns related to societal assumptions, expectations or biases. This made bias mitigation (or management) at this stage narrower and more focused on preventing brand risks instead of promoting fairness or trust.

However, this began to change with the emergence of large language models in the 2010s as open-domain dialogue generation made systematic biases more visible. This recognition has since sparked research into potential mitigation strategies over the last couple of years. Initial efforts of bias mitigation at this stage concentrated mostly on identifying overtly prejudiced outputs or discriminatory patterns in training sets although this could have been difficult since bias analysis itself is an inherently normative process as it involves subjective judgments (Blodgett et al. 2020, 5455 & 5460). As large language models, specifically from the chatbot perspective, have become increasingly integrated into real-world applications, the challenge of addressing bias has become complex, thus calling for

multifaceted and nuanced approaches (Ferrara 2024, 5; Zhou 2024, 86; Guo et al. 2024, 2-3). In response to this, bias mitigation techniques were categorized into three stages; the pre-processing, in-processing, and post-processing stages (Ferrara 2024, 5-6), which was first formalized in the IBM AI Fairness 360 paper (see Bellamy et al. 2018).

Since then, the categorization has been seen to also be applicable in addressing biases in chatbots, with the structure becoming the preparation (pre-model) stage, the development (intra-model) stage and the optimization (post-model) (Xue et al. 2023, 14; Guo et al. 2024, 15-17). The preparation stage focuses on preprocessing the data used in training these chatbot systems by using techniques like data augmentation, oversampling, under sampling and expert intervention, to ensure the training data is balanced and representative, thus reducing underlying biases but not totally removing them (Xue et al. 2023, 14; Guo et al. 2024, 15). The development stage looked more into intra-model techniques to mitigate biases during model design and training (Xue et al. 2023, 14).

Examples of such techniques include transfer learning (Guo et al. 2024, 16), adversarial training methods (Ernst et al. 2023), and Shokrollahi's (2023) quantum inspired intersectional bias mitigation. Then the final stage being the optimization stage occurred after the chatbot had been deployed, where post-model strategies were implemented to detect and mitigate biases that may have arisen from human interaction, i.e real-time mitigation through human feedback loops, reinforced calibration and projection-based methods (Xue et al. 2023, 14; Ferrara 2024, 7; Dai et al. 2024, 6440; Guo et al. 2024, 17). An example of such a technique is Narayan et al.'s (2024) Bias Intelligence Quotient, a framework that detects and neutralizes biases in model outputs.

Together, these stages attempt to ensure bias is addressed systematically across the entire AI life cycle. Be that as it may, other studies conducted by Blodgett et al. (2020), Talboy and Fuller (2023) and Ferrara (2023) amongst others have argued that there are still a set of fundamental challenges related to user perceptions, inadequate evaluation standards, unclear bias definitions depending on context and surface level mitigation strategies that do not necessarily address the root cause of the bias.

3.2.2 Effectiveness Analysis Across Different Bias Types

Considering the historical evolution of bias mitigation techniques, it becomes evident that the effectiveness of these interventions and techniques vary depending on the type of bias that is being addressed. Therefore, analysing mitigation effectiveness through the stages of bias mitigation reveals a certain pattern of success and limitation for each of the bias types. Table 6 displays a summary of the effectiveness of the bias mitigation stages across the bias types discussed in this thesis (data, algorithmic and human interaction).

Table 6. Summary of Effectiveness of Mitigation Stages across Bias Types

Bias Types	Preparation Stage	Development Stage	Optimization Stage
Data Bias	Highly effective but is limited by subjective bias judgement	Moderately effective but bias could remerge from fine-tuning loops	Limited effectiveness since it addresses mostly overt biases
Algorithmic Bias	Slightly effective but there are still risks of hidden generalization issues	Moderately effective as well but may reduce creativity	Deep structural biases may persist due to strategies being surface-level corrections
Human Interaction Bias	Limited effectiveness due to real world user unpredictability	Moderate to limited effectiveness depending on specific contexts	Fragile stage that risks reinforcing dominant norms even though it is essential

To begin, *Data Bias*, which occurs when training data reflects societal prejudices, is seemingly the most effectively addressed during the *preparation stage* (Xue et al. 2023, 14; Guo et al. 2024, 15-16). Based on studies conducted by Xue et al. (2023), Guo et al. (2024), Das and Sakib (2024) and Babonnaud et al. (2024) amongst others, there are different techniques to mitigate this specific type of bias, such as expert intervention, oversampling and dataset curation. Another example is Dai et al.'s (2024) suggestion of creating two versions of the same training sets where one possesses unprocessed and unfiltered data, while the other has been processed to remove surface-level biases.

These sets would then use ensemble learning to detect when a bias-sensitive correction is needed. Beattie et al. (2022) also highlighted the importance of implementing bias annotation protocols in the development of fair AI chatbot systems. These examples illustrate a few widely practised techniques that have been revealed to improve the fairness of training data. However, these techniques are not exhaustive, and ongoing research continues to propose additional strategies.

The *development stage* on the other hand exhibited mixed levels of effectiveness, as methods related to fine-tuning on augmented or curated datasets enhanced fairness, but did not fully eliminate residual biases (Talboy & Fuller 2023, 8-10; Xue et al. 2023, 14; Zhou 2024, 91). Guo et al. (2024) documented transfer learning techniques extended representation to low resource domains while the adversarial training methods explored by Ernest et al.'s (2023) displayed promise in finding and fixing hidden training biases.

However, these two techniques either preserve underlying biases from their source models or their effectiveness diminish in areas where data patterns are more complex. At the *optimization stage*, interventions for data bias are generally less effective at addressing intersectional or implicit biases as a result of their complexity and likely additional data requirement (Ferrara 2024, 6). Considering the above, data bias mitigation is shown to be effective and straightforward when it has been addressed in the early stages of the chatbot system's life cycle, however issues relating to subjectivity would likely remain.

Turning to *Algorithmic Bias*, intervention at the *preparation stage* would have limited success since the biases emerge from a model's architecture and its learning processes (Ferrara 2024, 2-3). Based on the ideas presented by Blodgett et al. (2020), Guo et al. (2024) and Ferrara (2024), data augmentation and curation could reduce the biases entering the model, but other forms of bias related to optimization and structure would still emerge during training. The interventions related to the *development stage* however showed stronger effectiveness. Techniques such as Shokrollahi's (2023) quantum-inspired intersectional bias mitigation approach demonstrated promising multi-dimensional control by changing how models process identity-related features. Similarly, Ernest et al.'s (2023) adversarial training methods and Guo et al.'s (2024) documented transfer learning techniques can also be used here since they both directly affect a model's behaviour and ability to suppress bias patterns while they learn.

Despite this, some studies such as (Talboy & Fuller 2023; Xue et al. 2023; Zhou 2024; Guo et al. 2024) imply that over-correction at this stage could negatively affect the model's creativity and fluency as well. Then the level of intervention or the strategies at the *optimization stage* possesses limited effectiveness due to the fact that post model strategies tend to mostly tackle surface-level issues alone (Ferrara 2024, 3; Dai et al. 2024). This can be seen in tools like Narayan et al.'s (2024) Bias Intelligence Quotient, which provides

sophisticated bias detection and correction capabilities, but still grapples with context dependent biases. In light of this, while algorithmic biases are known to be challenging to eradicate completely, mitigating them is more effective during the development stage.
As for *Human-Interaction Bias*, mitigation strategies and other forms of intervention at the *preparation stage* would be the lease effective since biases arises from user (designer or end-user) interaction in real-world contexts (Xue et al. 2023, 10-14). Furthermore, Eisenmann et al. (2024), Xue et al. (2023), and Guo et al. (2024), suggest that exposing models to diverse conversational styles does not necessarily mean they would be able to fully anticipate emergent biases despite the fact that the method can improve initial robustness within the model. *Development stage* interventions like adversial training, had moderate levels of effectiveness since it has been demonstrated that models trained on diverse interaction patterns are more capable of managing user-introduced biases (Xue et al. 2023, 11-12).

The *optimization stage* as seen in Table 6, is the most critical of all three stages for mitigating human-interaction biases. This is because post deployment systems that involve human feedback loops, projection-based recalibration methods and real-time moderation mostly try to dynamically adjust chatbot outputs (Xue et al. 2023, 11-13; Guo et al. 2024,17). These loops, however, indicate that user-driven feedback could inadvertently reinforce dominant prejudices and biases which would affect the overall fairness and neutrality of these models (Dingler et al. 2018, 1667; Xue et al. 2023, 12). Therefore, human-interaction bias would demand continuous real-time mitigation strategies, but said strategies are still prone to complexities, feedback bias and inherent (or not) user manipulation.

Further drawing on the information from table 6, it can be seen that there are also interdependencies between the stages that are capable of shaping a mitigation strategy's success. For data bias, table 6 illustrated a certain progression from the highly effective but limited preparation stage (which improves initial fairness) to limited effectiveness in the optimization stage (where fine-tuning could reintroduce bias). Then with algorithmic bias, the table demonstrated how mitigation is more focused on the way development stage decisions propagate. An inadequate preparation stage could lead to an overburdened development stage and inadequate optimization stage which could then obscure bias

patterns that emerge only during deployment. Finally, regarding human interaction bias, the table shows the progression from limited effectiveness in the preparation stage up to a delicate optimization stage that could end up reinforcing dominant norms. Based on the above analysis, all three bias types possess a specific strategy or set of strategies that vary in effectiveness based on the mitigation stage. Nevertheless, when reflecting on critiques by Blodgett et al. (2020), Talboy and Fuller (2023), Xue et al. (2023), and Guo et al. (2024), it is evident that there is no one size fits all in regards to mitigation strategies on top of the fact that no mitigation strategy guarantees the complete removal of a particular bias. Therefore, the effectiveness of any mitigation strategy would depend on its alignment and interaction with previous and future stages.

3.2.3 Inconsistencies and Limitations of Current Mitigation Strategies

Despite the advancements and effectiveness of the bias mitigation strategies in LLM chatbots, there are still notable inconsistencies and limitations. They span from the lack of universal definitions and context sensitivity, trade-offs between fairness, engagement and utility, reactive and surface-level mitigation techniques, feedback loops likely reinforcing prejudices and biases and scalability and operational challenges (Blodgett et al. 2020, 5460; Shokrollahi 2023, 5183–5184; Xue et al. 2023, 2, 6-7; Narayan et al. 2024, 2-3; Zhou 2024, 87; Dai et al. 2024, 6441–6442; Guo et al. 2024, 15; Das & Sakib 2024, 12). These challenges can influence the broader factors explored in this thesis, specifically user trust and satisfaction.

One of the more persistent limitations is related to the absence of a universally accepted definition of bias in AI systems. This is due to the fact that bias itself is inherently context-dependent and normative in nature, indicating that what constitutes as bias can vary across different situations (Blodgett et al. 2020, 5454; Ferrara 2024, 2). Consequently, this definitional vagueness results in inconsistent bias identification, measurement and mitigation across different chatbot systems, and other AI technologies, (Blodgett et al. 2020, 5460). Additionally, it is also important to note that a mitigation strategy that is effective in one context would likely not generalize to others, thus creating a gap between “debiased outputs” and real-world user perceptions.

Another limitation is centred around methodological inconsistencies, especially when it comes to trade-offs between bias mitigation and model performance. Xue et al. (2023), and

Zhou (2024), highlighted that models that are optimized aggressively for fairness could become overly evasive leading to noncommittal answers that reduce user experience. For instance, techniques like Chan and Wong's (2024) fairness-constrained fine-tuning and Ernest et al.'s adversial training could lead to outputs exhibiting reduced creativity and responsiveness even though they improve bias control during the development stage. These qualities (creativity and responsiveness) are central to user satisfaction in chatbot interactions (Guo et al. 2024, 44-45).

The third limitation explores current optimization stage strategies, especially those related to surface-level strategies like the Bias Intelligence Quotient (BiQ) (Narayan et al. 2024). Prompt-conditioning layers (Beattie et al. 2022) and bias-aligned persona conditioning (Dingler et al. 2018) also offer marginal interventions that are unable to fully prevent the emergence of bias due to complex or dynamic user interactions. These strategies mostly concentrate on detecting and correcting or suppressing potentially problematic outputs since they are limited in their ability to address the root cause of generative biases given it's normative nature (Blodgett et al. 2020; Ferrara 2024; Dai et al. 2024). Simply put, they are more likely to treat a bias symptom instead of the cause.

The fourth limitation is related to human-driven feedback approaches such as reinforcement learning with human feedback, which is a major aspect of most bias mitigation strategies (Xue et al. 2023, 14; Ferrara 2024, 7; Zhou 2024, 87; Guo et al. 2024, 17). Despite the effectiveness of this strategy, studies have revealed that feedback used to recalibrate chatbot behaviour has a tendency to reflect dominant societal prejudices and biases, thus marginalizing minority viewpoints (Dingler et al. 2018, 1665; Guo et al. 2024, 14 & 17). Additionally, models that are being adapted in accordance with user interactions also risk reinforcing the majority norms sited amongst the users, exacerbating biases instead of eradicating them especially in global contexts (Xue et al. 2023 11-12; Das & Sakib 2024, 11-12).

Therefore, mitigation strategies that depend on this feedback loop should be designed in such a way that the act of amplifying prejudices are avoided. This could be done through the implementation of diverse evaluation panels, employing continuous and rotational human oversight, or establishing clear bias detection metrics prior to deployment. Finally, the fifth limitation concerns scalability challenges due to difficulty of large-scale integration and high computational costs of seemingly more complex mitigation strategies

like the quantum intersectional bias correction (Shokrollahi 2023, 5184) and real-time conversational monitoring modules (Xue et al. 2023, 14). Organisations may also find such strategies to be operationally impractical especially when it comes to balancing performance, resource constraint and development speed.

Given these current limitations in chatbot bias mitigation, user trust, satisfaction and overall experience could be adversely affected. Studies conducted by Xue et al. (2023), Ferrara (2023), Zhou (2024), Guo et al. (2024), and Ferrara (2024), suggest that marginalized users have reported experiencing some degree of bias in "debiased" systems. These same studies also indicated how some users (especially those from marginalized communities) have reported feeling excluded or frustrated when a system appears to be inconsistent, or evasive or incapable of recognizing their needs. Such experiences are due to unpredictable interactions caused by inconsistent mitigation strategies which likely affects trust negatively while surface-level interference leads to unresponsive or evasive systems (Holliday et al. 2016, 167; Ferrara 2024, 5; Al-Shafei 2025, 423). These issues expose the gap between the technical fixes and real-world experience. Therefore, it can be implied that understanding user perceptions in chatbot interactions becomes critical for the long-term success of a mitigation strategy.

3.3 User Trust, Satisfaction, and Experience

As seen in previous sections, studies like Xue et al. (2023) and Guo et al. (2024) have shown that bias and its mitigation in chatbots (and other AI technologies) will likely continue to become more advanced and technical. Ultimately however, the effectiveness of any bias mitigation strategy is inadvertently judged by the user and the quality of their experience, especially in terms of their perceived trustworthiness and fairness (Borsci et al. 2021, 107; Gong et al. 2024, 146-147). If a user perceives a chatbot system as unfair or misaligned with their own expectation, it could lead to negative consequences that affect experience, trust, and even the overall adoption of the system (Duncan & Mcculloh 2024, 687 & 690; Kantharuban et al. 2024, 1-2). Additionally, from an ethnomethodological perspective (also known as EMCA perspective), users treat these interactions as socially situated practices by applying their own norms of accountability, conversational coherence, and varying levels of appropriateness (Eisenmann et al. 2024, 2728).

Therefore, the intersection of bias in LLM chatbots and user satisfaction presents an important study area with significant implications for both social interactions and technological development (Yao & Xi 2024, 1). This section explores how user expectations have evolved, the correlation between bias detection and user satisfaction metrics, the impact of biased interactions on user engagement and trust and how mitigation efforts can affect user engagement.

3.3.1 Evolving Expectations and Perceptions

User expectations regarding AI chatbots have undergone their own form of evolution as these technologies have become increasingly integrated into daily activities. Given the rigid nature of the earlier chatbot generations (i.e. rule-based and retrieval-based systems), user expectations were majorly task oriented. In other words, their expectations typically revolved around basic query handling and informational retrieval/generation, although when the need arose users would adjust said expectations accordingly (Caldarini et al. 2022, 8; Xue et al. 2023, 3 & 6; Al-Amin et al. 2024, 6-9; Yao & Xi 2024, 3). However, as chatbots and their applications continued to advance, especially in terms of their human likeness, user expectations began to shift towards the desire to interact with systems that are responsive, contextually and ethically knowledgeable, and emotionally intelligent (Nicolescu et al. 2022, 1-2; Wut et al. 2024, 9; Al-Shafei 2025, 412). A chatbot system or any other relevant user interface that displays these characteristics are seen to be more humanlike, which enhances overall user interaction experience (Al-Shafei 2025, 412).

This shift from basic functional interactions to modern day expectations of having a relatable digital assistant is further supported by Markovitch et al.'s (2024) research. They highlighted how previous surveys found that emotional intelligence and perceived empathy have significant influences on the loyalty and overall satisfaction of a user (Markovitch et al. 2024, 2). In that wise, obtaining accurate information is now just as important as the way it is delivered to the user in terms of the delivery method's alignment with the user's ethical and social expectations (Lee & Chan 2024, 90–93). This current trend can also be further explained from the Ethnomethodological Conversation Analysis (EMCA) perspective, which was initially explored in Harold Garfinkel's work on early chatbot models like ELIZA. The EMCA perspective offers detailed understandings of the various ways users interact with AI (Eisenmann et al. 2024, 2728). As such, Garfinkel observed

that users had the tendency to humanize these models because they employed their own understandings of generic social interactions to interpret the ones they had with a chatbot (Eisenmann et al. 2024, 2722–2723 & 2728).

Eisenmann et al. (2024, 2717) further discuss that the reasoning behind the users interpretative work, through an aspect of the EMCA, known as “trust conditions”. They are the inherent assumptions of mutual competence, engagement, and responsiveness between participants (human or machine) of a specific conversation or situation. Therefore, if a chatbot exhibits aspects related to the trust conditions such as transparency, competency in completing tasks, allowing user control and behaving predictably, the chatbot is more likely to foster coherent and smooth interactions (Holliday et al. 2016, 165; Eisenmann et al. 2024, 2717–2718). Although, if the chatbot is unable to meet these conditions as a result of biased responses for instance, users are likely to become distrustful or frustrated and may stop engaging with the chatbot (Holliday et al. 2016, 167; Al-Shafei 2025, 426). This sheds light on how users naturally project human interaction patterns onto AI systems by judging them based on responsiveness, engagement and mutual competence which would otherwise be reserved for human communication (Holliday et al. 2016, 166–167; Haugeland et al. 2022, 3-4; Li et al. 2023, 15).

Still, this paradox reflects an inherent risk in the impulse to over-humanize these chatbots (Virvou et al. 2024, 5; Al-Shafei 2025, 419). Although human-like traits can improve engagement as well as satisfaction, they could cloud the ethical and technical constraints in the systems as they lack true intentionality and understanding (Haugeland et al. 2022, 3; Ferrara 2023, 16; Xue et al. 2023, 9; Virvou et al. 2024, 5-6). A chatbot’s seeming empathy and/or ethical reasoning originates from pattern recognition rather than authentic comprehension (Xue et al. 2023, 9). Technical solutions to address this could lead to varying trade-offs such as less originality with alignment, more bias imbued in emotion mirroring and possibly perceived lack of empathy with neutrality (Nicolescu et al. 2022, 19-20; Yuan et al. 2023, 6–9; Xue et al. 2023, 15; Huang et al. 2024). From an ethical perspective, human-like AI may inherently encourage users to develop unhealthy attachments or place excessive trust its advice during critical situations, without fully comprehending the technology’s limitations (Virvou et al. 2024, 5-6).

These aspects reflect the earlier mentioned intersections between chatbots and user satisfaction, while also indicating the growing entanglement of experience, perceptions of

human-like AI systems and possibly ethical standards in human-machine interactions. On that note, considering that chatbots are increasingly integrated into everyday life, designers/programmers would need to navigate the line between using human-likeness and ensuring that expectations are realistically calibrated. Implementing contextual self-disclosure (chatbots communicating their shortcomings when needed) or employing progressive familiarity (chatbots gradually reveal their capabilities over-time) are ways to balance human-like AI and realistic user expectations.

3.3.2 Correlation between Bias Detection and User Satisfaction

As seen in the previous section, bias in chatbots can profoundly affect how users rate their satisfaction during their interactions, depending on the stakes and influences it may have in their lives (Yuan et al. 2023, 2). Recent studies have found that fairness has become a vital driver for user satisfaction in AI-powered chatbots, meaning that a user's response to perceived unfairness is likely to be negative (Xue et al. 2023, 15). It can therefore be assumed that even when chatbot outputs are factually correct, the presence or perception of bias could undermine overall trust which then leads to reduced satisfaction.

In light of this, Yuan et al. (2023) pointed out the importance of contextualizing user perceptions of bias by noting how the system's transparency (or explainability) and the user's background (how the user may perceive an output based on their past experiences) can influence how bias is understood and its effect on the user's satisfaction. That said, user interpretation can drive emotional and trust-related responses that shape satisfaction. For example, Wahbeh et al. (2023, 36 & 38) collected over 5000 public tweets about ChatGPT from "X" and found that most users reported neutral sentiments about bias in ChatGPT, followed by negative sentiments and positive sentiments. Ultimately these findings reflected negative user experiences or overall thoughts about ChatGPT at that point in time. This also aligns with Wuenderlich and Paluch (2017) study where they demonstrated that even subtle cues in language or tone, if perceived as biased or unfair, can erode a user's positive emotional engagement with a chatbot.

That said, being able to understand how users perceive and likely respond to bias in chatbot responses is important to evaluate user satisfaction. Biases in chatbot systems (data, algorithmic or human interaction bias) are capable of undermining trust and negatively affecting user experience. Wahbeh et al.'s (2023, 36) study also noted how the results of

their data analysis reported several types of bias, with data and algorithmic bias being the main concerns. The posts further expressed that algorithmic bias often becomes apparent in the model development phase, developers have the tendency to overlook this, causing models to inadvertently reinforce societal prejudices (Wahbeh et al. 2023, 37).

Still, bias manifests in different forms, and its detection could be shaped by multiple dimensions as seen in the previous sections. Data bias, usually stemming from unbalanced datasets, often leads to underrepresentation or entire exclusion of certain demographics, which is especially seen in sensitive sectors like healthcare (Xue et al. 2023, 9; Guo et al. 2024, 4). Algorithmic bias, on the other hand, could produce inconsistencies in chatbot responses, which users may interpret as unfair depending on the context of their initial query (Xue et al. 2023, 9; Guo et al. 2024, 4-5). Then Human-Interaction bias could significantly shape a user's perception of a chatbots output since the way responses are framed (conversational style), sequenced (phrasing), or emphasized (tone) during interactions contributes to a user's perceived fairness (Xue et al. 2023, 9; Guo et al. 2024, 5).

Additionally, the anthropomorphic design elements that are increasingly being incorporated into chatbot algorithms inadvertently amplify bias sensitivity seeing as these elements evoke higher expectations of accountability and fairness (Wuenderlich & Paluch 2017). Therefore, when an individual interacts with human-like chatbots, the phenomenon of anthropomorphism is able to change how bias is experienced. Ironically, if this same human-likeness was overextended, it could backfire if users perceive the emotional cues to be manipulative instead of empathetic (Al-Shafei 2025, 423), thus eroding trust and satisfaction.

Admittedly, enhancing a chatbot's emotional intelligence seems to be the most viable step towards deeper human trust and engagement, but it simultaneously makes perceived fairness more critical. This is the crux of anthropomorphic design, where the more a chatbot appears to be human, the higher the expectations for it to behave rational and fair, hence why it becomes so damaging when these systems fall short (Chan & Leung 2021, 9; Markovitch et al. 2024, 10; Al-Shafei 2025, 423). This sensitive point creates a paradox with anthropomorphic features heightening user satisfaction when bias is absent, but also amplifying disappointment when bias is felt. In addition to these heightened expectations, the way users react to perceived bias and unfairness is influenced by contextual factors

such as personal identity, technical literacy, and prior experiences with AI systems (Chan & Leung 2021, 1–10; Nicolescu et al. 2022, 1579; Al-Shafei 2025, 411–428).

However, measuring the above relationship between bias detection and user satisfaction presents methodological challenges. In some cases, studies have found that users only become aware of bias after the interaction has ended or when it is pointed out (Nicolescu et al. 2022, 17-18; Haugeland et al. 2022; Al-Shafei 2025, 427). Yet, despite this delayed recognition, user trust can still be diminished which in turn lowers satisfaction. Moreover, it is possible for dissatisfaction to also be accumulated gradually overtime through repeated biased interactions (Virvou et al. 2024, 3-4).

This displays a difference between bias that is perceived and that is objective as users may not consciously detect bias during these interactions but may still "feel" a certain discomfort that affects their satisfaction (Yuan et al. 2023, 7; Cabrero-Daniel & Sanagustín Cabrero 2023, 6–7). Yuan et al. (2023) and Wuenderlich and Paluch (2017) further discussed this by looking into a temporal dimension, where initial tolerance breaks down into eventual disengagement. As a result, effective assessment would require combining subjective feedback with behavioural indicators such as usage frequency and return rates. Tools like Borsci et al.'s (2021) Chatbot Usability Scale (a standardized tool to measure user satisfaction with AI chatbots) offers promising approaches for capturing this complexity.

With the above in mind, the correlation between bias detection and user satisfaction is shaped by a multifaceted interplay of user expectations, algorithmic features, and interaction contexts. These dynamics reflect a loop where the nature of chatbots become increasingly complex and users' expectations for fairness increases which then leads to a decreased tolerance for bias. The anthropomorphic elements here would boost engagement but could in return amplify the negative impact of even minor perceived bias on satisfaction. Consequently, minimal perceptions of bias could reduce user satisfaction, making fairness detection and user perception management a vital element in evaluating customer experience as well as the ethically effective deployment of a chatbot.

3.3.3 Impact Of Biased Interactions on User Engagement and Trust

Trust is known to be foundational in human–AI interactions, particularly in conversational agents that handle sensitive information delivery, mediate customer service, or aid in decision-making (Ng & Zhang 2025). A vast majority of users expect their chatbot interactions to be empathetic and reliable, but when these expectations are unmet users are likely to disengage or report lower satisfaction (Markovitch et al. 2024, 3-5; Al-Shafei 2025, 423). These adverse effects on trust, engagement and overall satisfaction have the tendency to persist on the long-term (Yuan et al. 2023, 1).

That said, an impact of biased interactions on user engagement stems from research by Chan and Leung (2021) and Wuenderlich and Paluch (2017). They highlighted an expectation gap where users anticipate human-like reasoning and understanding from chatbots. This expectation is fuelled by the increasingly anthropomorphic design of chatbots, which a typical user would view as the system being empathetic and socially intelligent (Markovitch et al. 2024, 4-5). Although when the interactions yield outputs that users perceive to be prejudiced, they experience cognitive dissonance that challenges the trust they had in the system (Chan & Leung 2021; Schmitt 2022; Markovitch et al. 2024, 10). The mismatch or dissonance here would lead users to reappraise their trust in the chatbot, whereby they question its alignment and reliability. This effect is amplified when chatbots give inconsistent responses or lack perceived empathy during interactions over time with users (Markovitch et al. 2024, 10; Al-Shafei 2025, 423). Therefore, trust erosion would not be immediate in most cases as users would only note the dissonance during repeated instances.

Building on this, interaction design also plays a role in influencing the bias and user engagement dynamic. Haugeland et al. (2022) and Al-Shafei (2025) demonstrated how alienating users by ignoring variability in communication styles, further reduce engagement and satisfaction through contextual (and cultural) misalignment in chatbot behaviour. This phenomenon is further exacerbated by underlying algorithmic biases stemming from the inattentive collection of data and overall processing practices (Wahbeh et al. 2023, 37).

With the increasing accumulation of these design and data oversights, it has been seen that they contribute to the predictable deterioration of trust. The deterioration follows a pattern

that include three determinants namely, predictability, fairness and competence. These are the foundations of trust in chatbots and are the qualities that are frequently compromised, thereby leading to bias (Schmitt 2022, 912). The VIRTSI model also portrayed how repeated biased interactions would gradually erode a user's desire or willingness to (re)engage with a chatbot, even if the more overt biases are absent (Virvou et al. 2024, 4-7).

Considering this, if the information being delivered by a chatbot is factually accurate, the way it is delivered could still leave users unsatisfied with their experience, assuming there is a perceived lack of empathy, or bias (El Gharbaoui et al. 2024, 1828–1829). These notions position trust as a critical moderating variable in determining user satisfaction and whether or not users continue engaging with a chatbot even after bias is detected (El Gharbaoui et al. 2024, 1829; Ng & Zhang 2025). However, when looking beyond individual interactions, user trust and satisfaction in chatbots can directly be impacted by the chatbot's parent company.

In other words, the cumulative effect of unchecked biased responses could potentially damage user-brand relationships. This is because users have the tendency to associate a chatbot's behaviour to the values of the organization that created it, while viewing personalization and communication competence as key indicators of its trustworthiness (Lee & Chan 2024, 93–94; Xie et al. 2024, 619–620). Taking the above findings into account, it is clear that biased interactions can diminish user trust and perceived fairness while also discouraging long-term engagement. As chatbots and their applications become more embedded into everyday life, the tolerance for bias would decline. This underscores the need for AI chatbot design approaches to be trust-sensitive and fairness-aware, to not only mitigate bias, but to preserve the relational integrity of human-AI interactions in the long run.

3.3.4 Impact of Mitigation Efforts on User Engagement and Trust

Following the exploration on how biased interactions could diminish user trust, satisfaction and engagement, it would be beneficial to investigate how mitigation strategies could rebuild or reinforce critical user outcomes. Since bias mitigation in AI chatbots has become more structured (following a three-stage process), its influence on user perception,

satisfaction, and interaction has deepened (Xue et al. 2023, 14-15; Ferrara 2024, 5-6; Guo et al. 2024, 15-18).

That said, a key impact area to begin with is the moderating influence of empathy on trust recovery. A study revealed how increasing a chatbot's perceived empathy through more empathetic communication (while avoiding excessive anthropomorphism) improved consumer evaluations of chatbot service (Markovitch et al. 2024, 10). This indicated that user trust depends on perceived empathy during an interaction and not just on the service experience. Such an impact would align with the optimization stage of the chatbot lifecycle, where strategies emphasizing empathy can compensate for prior negative encounters and strengthen or restore user trust. Another would be based on the expectation-perception gap and how it can affect user satisfaction according to Chan and Leung (2021). This would encompass strategies conducted during the development stage, such as Ernest et al.'s (2023) adversarial training, which could act as a proactive mitigation tool to manage user expectations and satisfaction.

A third area would be centred around interaction design and engagement in the chatbot's development phase. Haugeland et al. (2022), Wahbeh et al. (2023, 37) and Al-Shafei (2025, 423) discussed how design bias mitigation is able to affect user engagement. They found that engagement is likely to be higher when chatbots exhibit responsive, consistent and fair behaviour, aspects that usually require mitigation strategies. This supports the view that when a bias mitigation strategy is implemented early in the chatbot's lifecycle, said strategy can contribute to the perceived fairness of chatbot interactions. The fourth is related to trust, its dynamic nature and how potential biases could be mitigated during the optimization stage. Virvou et al.'s (2024) VIRTSI model exhibits how trust fluctuates and develops overtime in response to the chatbot's behaviour, user familiarity and potential feedback. Other optimization stage strategies such as the human-in-loop feedback help maintain trust and user confidence after the chatbot has been deployed, which is overall beneficial in maintaining user satisfaction and engagement.

Then the fifth aspect is related to the quality of the dialogue during the interaction as it serves as a key mediator of trust. In other words, user trust is improved and sustained when the chatbot is able to maintain transparent but coherent conversations (Ebubechukwu et al. 2024, 3 & 12). This matches other optimization strategies like Narayan et al.'s (2024) Bias Intelligence Quotient, which actively finds and neutralizes potentially biased outputs.

Having said that, mitigation strategies that span the preparation, development and optimization stages are capable of fostering user satisfaction and brand trust through emotional connection while also addressing technical concerns (Xue et al. 2023; Ebubechukwu et al. 2024; Guo et al. 2024). These impacts are more defined when the mitigation strategies prioritize the user and adapt to their needs while also being transparent, thus allowing users to trust and witness the fairness as it happens.

3.4 Evolving AI Ethics and Synthesizing Bias Mitigation with User Trust

This section integrates the findings from the previous discussions on bias evolution, mitigation strategies and user experience and satisfaction. It uses ethical frameworks to examine how the widespread adoption of LLM chatbots, into day-to-day activities has prompted the demand for responsible AI development (Alhajjar & Bradley 2022, 6; Babonnaud et al. 2024, 197; Guo et al. 2024, 18). Like bias and its mitigation strategies, ethical concerns are not static as they have evolved alongside shifting user perceptions, AI technology and the growing awareness of digital inequalities. Therefore, this section explores the historical development of ethics, and the relationship between ethical considerations in chatbot development and user trust/satisfaction.

3.4.1 Ethical Considerations in Chatbot Development

As seen in the previous 3.1 section, early chatbots like ELIZA were predominantly rule-based with limited capabilities and focused more on functionality with ethical concerns concentrated on simple issues like human transparency (Caldarini et al. 2022, 8; Xue et al. 2023, 3-4; Al-Amin et al. 2024, 7-8). These early systems worked within a specific scope where ethical discussions highlighted the necessity to inform users they were conversing with a machine (Luo et al. 2019).

However, overtime chatbots evolved into intricate LLM systems capable of mimicking human-like behaviour, thus introducing challenges (such as bias amplifications, erosion of trust, misinformation etc), which has also complicated the ethical landscape (Chan & Wong 2024, 2; Zhou 2024, 87; Feng et al. 2024, 544–545). This evolution suggests that *ethics* needs to be treated as an essential part of chatbot design, as briefly mentioned in the mitigation stages of section 3.2. As a result, concerns towards anthropomorphization, user

manipulation and psychological impact began to increase (Alhajjar & Bradley 2022, 72; Bryant 2023, 49; Guo et al. 2024, 18-20).

Consequently, the recent advancements urged the shift in ethical focus from just transparency to other considerations consisting of data privacy, emotional safety, potential harm and accountability (Nicolescu et al. 2022, 1579; Li et al. 2023, 1–24; Huang et al. 2024). This progression is further reflected by Feng et al.'s (2024) proposed CARE (Compassion, Accountability, Respect and Equity) Framework, which is a holistic approach to optimize chatbot impacts across their lifecycle. The model highlighted the need for developers to consider the broader roles of chatbots in addition to user-perceived impacts and situational context. Haugeland et al. (2022) and Al-Shafei (2025) further support this point as they showed that ignoring variability in communication styles and causing contextual (and cultural) misalignment in chatbot behaviour erodes user trust.

Moreover, research conducted by Müller et al. (2019) and El Gharbaoui et al. (2024) showed that ethical transparency is strongly correlated with user trust and satisfaction, especially when bias mitigation efforts can be seen and are understandable. The evolution of ethical considerations in chatbot development highlights a transition from simplistic rule-based ethics to multidimensional, user-sensitive ethical frameworks. As LLM-based systems become increasingly embedded in social and institutional contexts, the demand for continuous re-evaluation of ethical standards and best practices grows correspondingly.

3.4.2 Ethics Mediating Trust and Satisfaction in Bias Mitigation

The evolution of ethical frameworks in chatbot development has improved the understanding of their mediation of user trust and satisfaction. Given the growing significance of AI across varying domains, ethical considerations in its development are also becoming crucial. Therefore understanding the gaps in AI design and the consequence of biased data is important to address emerging challenges (Alhajjar & Bradley 2022, 73). This relationship is especially prominent from the bias detection and mitigation perspectives, where user perception of fairness, empathy, and system accountability has a direct impact on the quality of the user experience (Markovitch et al. 2024, 10).

As demonstrated in the earlier sections, technical interventions (such as algorithmic fine-tuning) could contribute towards bias mitigation but would not be enough if users are unable to recognize or understand it during their interactions with a chatbot. That said, a

system can be *fair*, but user trust could still erode if the system is unable to exhibit such fairness clearly, or in a way that aligns with a user's expectation. This leads to the notions of ethical explainability and transparency, which are two prominent themes of AI ethics and have been seen to directly influence user satisfaction (Sethy et al. 2023, 190). Research conducted by Müller et al. (2019) and El Gharbaoui et al. (2024) found that users report higher satisfaction and trust when chatbots acknowledge their own limitations. Users also appreciate these systems more when they are offered understandable and user-friendly explanations for how response outputs are generated (Cabrero-Daniel & Sanagustín Cabrero 2023, 2). These findings support the idea that while fairness should be embedded in a system's algorithm, it also needs to be perceivable by users during their interactions.

Feng et al.'s (2024, 537–548) CARE (Compassion, Accountability, Respect and Equity) Framework highlighted the importance of user-perceived ethics for the long-term success of chatbots. The framework argues that ethical AI should be noticeable in user interactions through emotionally intelligent designs, appropriate outputs and transparent accountability processes. For example, if a chatbot were to openly indicate that its output is biased, or that it rephrased said output, it is likely to strengthen user trust as well as mitigate potential harm.

This is closely linked to the points earlier raised on evolving user expectations and Heersmink et al.'s (2024) discussion on interactional flow/fluency and phenomenological transparency. An LLMs chatbot's apparent fluency fosters user trust based on the smoothness of the conversation and not on the actual understanding or fairness of the situation (Heersmink et al. 2024, 10). This presents an ethical challenge in the sense that even though interactional fluency improves usability, it could lead to over-trust (Virvou et al. 2024, 5-6). Ethical design would therefore be required to incorporate mechanisms that make fairness and limitations noticeable to users. In this context, ethical transparency becomes vital to avoid users from overestimating the system's epistemic capabilities or neutrality (Virvou et al. 2024, 5-6; Heersmink et al. 2024, 10).

This is compounded by user expectations as they increasingly expect chatbots to align with their own values and behave in what is deemed to be socially appropriate during a specific interaction (Wahbeh et al. 2023, 37; Al-Shafei 2025, 423). Studies like Fan et al. (2025) and Alhajjar & Bradley (2022) show that when users perceive a mismatch between their ethical expectations and the chatbot's behaviour (such as dismissing sensitive topics) the

user’s trust deteriorates, even if the system performs well on objective benchmarks. This goes in hand with research conducted by Sonboli et al. (2021), confirming that user trust and satisfaction are strongly influenced by perceived fairness and transparency, especially in AI-driven decision systems. Their findings highlighted that users react more favourably to systems that explain their recommendations and further justify them based on fairness principles. This insight supports the argument that bias mitigation is only effective when it is both ethically informed and user visible.

Collectively, this discussion emphasizes that ethics plays a mediating role in how bias mitigation strategies are accepted and/or adopted by users, and not just peripheral to chatbot performance. A technically fair model might not generate trust if its fairness isn't compatible with what users deem fair or important. This highlights the necessity of embedding ethics in communication and not just design. This would ensure that users understand, perceive, and experience ethical behaviour throughout their interaction. With chatbots becoming more prevalent in sensitive areas, prioritizing ethical considerations throughout its lifecycle will be crucial in upholding public trust, encouraging meaningful user engagement and preventing harm (Alhajjar & Bradley 2022, 73; Li et al. 2023, 15; Ferrara 2023, 13).

4 Conclusions

The aim of this thesis was to systematically examine how bias in LLM-powered chatbot affects user satisfaction, trust and engagement overtime. Through an extensive literature review and a multi-framework analysis the study identified persistent and emergent bias based on their source, evaluated some current mitigation strategies and analysed their implications for user perception. The findings revealed that chatbots have become more complex and human-like, which in turn has caused user expectations to shift from the initial task-based functionality to more empathetic, relatable and ethical alignment. In this context, user perception of detected bias does significantly affect overall satisfaction and trust depending on the context. On the other hand, when mitigation strategies are seen as user-centred and transparent, they indicate significant promise in rebuilding user engagement while fostering sustained trust.

Ultimately, the research affirms that bias is continuously evolving issue, shaped and affected by technological, cultural and social forces. Therefore, to ensure equitable and trustworthy chatbot interactions, developers, researchers and perhaps policy makers need to approach bias mitigation as a continuous effort. This thesis contributes to the growing body of literature at the intersection of algorithmic bias, AI ethics, user satisfaction, user experience and human-AI interaction. The contributions are organized into two categories, the first being theoretical and the second being practical.

4.1 Theoretical Contributions

First, this study advances the theoretical discourse on algorithmic bias in AI chatbots by adopting a historical and evolutionary lens to trace the emergence and persistence of bias across four chatbot generations. This is visually represented in the Bias Evolution Map (see Figure 4), which highlights and differentiates the persistent versus emergent bias types. It also maps out how these biases have been shaped by various changes in interaction patterns, training data and algorithmic design. Framing bias identification in this light supports adaptive and longitudinal approaches to taken to create bias mitigation strategies.

Then, this thesis also contributes to the growing body of human-computer interaction research by explicitly linking bias detection with changes in user satisfaction, trust and engagement over time. Synthesizing these insights across varying disciplines allowed for

the creation of a holistic framework for understanding how biased interactions are perceived and how mitigation strategies (when implemented correctly at the right time) could restore user confidence. This represents a shift from siloed approaches to aligning user trust, ethics and AI performance into a model that explains the link between fairness and user satisfaction in chatbot interactions. Additionally, the thematic overlaps revealed in this study (see Table 4) highlight the interconnected nature of the four themes discussed in this thesis: bias identification, bias mitigation, user perception and satisfaction, and ethical frameworks.

Finally, this thesis contributes to the ongoing refinement of bias evaluation benchmarks through critically assessing how applicable a few of them are to LLMs and LLM chatbots. The benchmark critique in Table 5 exhibited a few core limitations in existing tools; such as the subjectivity and limited scalability of Babonnaud et al.'s (2024) audits and the overreliance of predefined metrics in Chan and Wong's (2024) hybrid model. Consequently, analysing these benchmarks pointed out that some current evaluation methods require continuous refinement and interdisciplinary collaboration to address evolving biases. In other words, the critique highlighted the need for context aware and adaptive evaluation models.

Taken together, these contributions demonstrate that addressing bias in AI chatbots should consider pairing historical contexts with adaptable ethically grounded and user-sensitive frameworks.

4.2 Practical Contribution

On a practical level, this thesis offers valuable insights for developers, designers, organizations, and policymakers involved in chatbot deployment and governance.
By tracing the evolution of bias and evaluating potential mitigation strategies across different chatbot generations, this research highlights which interventions could be most effective for specific types of bias (e.g., data-driven, algorithmic, or human-interaction-induced bias). This can directly inform the development and tuning of future chatbot systems.
Then drawing on the effectiveness analysis of mitigation strategies from Section 3.2.2, the thesis emphasizes that there is no single mitigation approach that works on its own. However, a multi-layered and stage-specific intervention could be helpful in reducing

compounding bias effects in modern LLMs, depending on the context. This guidance proposes a bias-specific approach that customizes interventions to distinct causes of bias, thus allowing relevant stakeholders to focus on the most appropriate and efficient intervention for a particular case. Additionally, the findings show that transparent communication about bias risks, clear limitation disclosures, and continuous feedback integration are important to restore user trust after biased interactions.

These practices represent a shift from viewing transparency as a necessary checkbox requirement toward trust-building as an ongoing design and interaction strategy. This thesis highlights the importance of proactive disclosure, clear explanations and user feedback in (re)building user trust while also maintaining user engagement in sensitive services areas (e.g. healthcare).

In general, these insights could inform businesses and developers on the development of more ethical and user-sensitive chatbot systems in real-world applications, such as customer service, healthcare, education, and public administration.

4.3 Limitations

Although this study was comprehensive in nature, there are a few limitations that need to be acknowledged. To begin, the study makes predominant use of Western-sourced English-language research, which could have potentially introduced cultural and/or geographical bias into the findings. It could have also affected the way chatbot bias and user satisfaction was interpreted and thus framed, especially since underrepresented perspectives are excluded. Another limitation is rooted in the scope of the bias analysis. In other words, the analysis is focused on bias sources and high-level classifications (i.e, training data bias, algorithmic bias and human-interaction bias) without explicitly addressing community-specific or intersectional biases. Therefore, even though the review successfully traces the evolution of bias across the generations, not disaggregating the bias sources into specific categories (such as sentiment bias) could hide important insights while reducing the applicability of the findings in specific contexts.

A third limitation is the emphasis placed solely on user trust, satisfaction and engagement, with little to no actual attention on other stakeholders such as regulators and system programmers. This could reduce the finding's ability to capture broader bias drivers like institutional constraints, which could be vital in creating long-term mitigation strategies. Finally, this thesis relied solely on published literature and secondary data. This means that

the insights that are explored are subject to the methodologies and analytical scopes of the reviewed articles. Even though the thematic analysis gave rise to various patterns, the original limitations of the used studies could have influenced the synthesized conclusions.

4.4 Future Research Directions

While the thesis does advance understanding of certain notions in certain areas, there are other areas that could be explored:

Intersectionality and Compounding Bias: Although this thesis explores and maps out bias evolution across the chatbot generations, there is a lack of attention on intersectional biases. Therefore, future research could explore how these biases may affect diverse user groups in order to discover more nuanced vulnerabilities. Understanding the impact these biases have may be useful in developing benchmarks and/or mitigation strategies that can be tailored to complex-identity based interactions.

Cross-Cultural Bias and Multilingual Evaluations: The literature corpus used in this thesis was primarily in English and heavily Western centric. As seen in this thesis, chatbot responses and user perception differ from context to context, indicating that biases that are subtle or undetectable in one context may be noticeable in another. Cross-cultural studies could reveal culturally embedded or region-specific bias. Therefore, future studies could examine how bias manifests in LLM chatbots across different cultural/geographical contexts, in order to contribute to more globally inclusive chatbots

Conducting an Empirical Longitudinal User Studies: Although this thesis discussed the impact of bias and bias mitigation on the evolution of user satisfaction, empirical studies could also look into how satisfaction changes over extended interactions. In other words, how does repeated exposure to biased systems affect a user's overall satisfaction as well as perception, trust and engagement. This could further validate the claims presented in this thesis and provide richer contextual understanding of user satisfaction and trust trends. It could also contribute to trust modelling and assist developers in creating transparent and resilient models that can improve relationships with users over time.

Integrate Stakeholder Perspectives: The underrepresentation of other stakeholders aside from the user was a key limitation noted in this thesis. As this thesis focuses on user impact, future research could examine how other stakeholders, i.e. developers, platform holders

and policy makers perceive and manage bias. Addressing questions like “How do business incentives shape mitigation priorities” or “What compliance tools are effective to promote ethical chatbot usage” could result in more informed governance structures and even foster dialogue between relevant stakeholders.

To summarise, future research could build on the findings of this thesis by involving a diverse set of stakeholder groups while developing adaptable approaches to identify, evaluate and mitigate chatbot bias. This could deepen the ethical and practical relevance of AI fairness in conversational technologies.

5 Summary

This thesis examined the evolution of bias in LLM chatbots as well as the impact these biases have on user satisfaction through a systematic literature review of 89 peer-reviewed articles. Tracing chatbot development from foundational systems to modern LLM-powered conversational agents created an evolution map that highlighted both persistent and emergent biases. The research was organized around four core themes: bias identification and evaluation, bias mitigation techniques, user trust, satisfaction and experience, and ethical frameworks, which overlapped with each other at different points and in different ways.

The literature identified that chatbot bias, like most AI-related biases, originated from three main sources (training data, algorithmic design, and human interaction) and have become more nuanced and harder to detect and address overtime. A range of mitigation strategies were also reviewed, and they were found to have varying levels of success across the different bias types and depending on when they are deployed within the chatbot lifecycle. That said, the thematic overlaps found amongst the reviewed articles showed that bias is a deeply interdisciplinary issue as well as a technical challenge that intersects with user experience, system governance and ethical design.

The findings contribute to the understanding that chatbot bias not is not a fixed nor is it an isolated concept. Instead, it is a dynamic and historically rooted phenomenon that is shaped by various socio-technical and socio-cultural forces. This thesis offers researchers and developers a clearer lens through which bias can be evaluated by assessing a few benchmark tools while also empowering them to propose more transparent, user-centric, and ethically aware design practices for future chatbot development.

References

Adams, S. S. et al. (2016) – I-athlon: Toward a Multidimensional Turing Test – *AI Magazine*, Vol. 37, (1), 78–84

Akhtar, Z. B. (2024) – Unveiling the evolution of generative AI (GAI): a comprehensive and investigative analysis toward LLM models (2021–2024) and beyond – *Journal of Electrical Systems and Information Technology*, Vol. 11, (1), 22

Al-Amin, M. et al. (2024) – History of generative Artificial Intelligence (AI) chatbots: past, present, and future development – arXiv

Alhajjar, E. – and T. Bradley (2022) – AI Ethics: Assessing and Correcting Conversational Bias in Machine-Learning based Chatbots – *Workshop Proceedings of the 16th International AAAI Conference on Web and Social Media*, Vol. 2022, 67

Almutiri, T. – and F. Nadeem (2022) – Markov Models Applications in Natural Language Processing: A Survey

Al-Shafei, M. (2025) – Navigating Human-Chatbot Interactions: An Investigation into Factors Influencing User Satisfaction and Engagement – *International Journal of Human–Computer Interaction*, Vol. 41, (1), 411–428

Aninze, A. (2024) – Artificial Intelligence Life Cycle: The Detection and Mitigation of Bias – *International Conference on AI Research*, Vol. 4, (1), 40–49

Babonnaud, W. et al. (2024) – The Bias that Lies Beneath: Qualitative Uncovering of Stereotypes in Large Language Models – *Swedish Artificial Intelligence Society*, 195–203

Beattie, H. et al. (2022) – Measuring and Mitigating Bias in AI-Chatbots – *2022 IEEE International Conference on Assured Autonomy (ICAA)* – 117–123

Bellamy, R. K. E. et al. (2018) – AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias – arXiv

Blakeney, C. et al. (2021) – Measure Twice, Cut Once: Quantifying Bias and Fairness in Deep Neural Networks – arXiv

Blodgett, S. L. et al. (2020) – Language (Technology) is Power: A Critical Survey of "Bias" in NLP – In: Dan Jurafsky et al. (ed.) – *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* – 5454–5476

Borsci, S. et al. (2021) – The Chatbot Usability Scale: the Design and Pilot of a Usability Scale for Interaction with AI-Based Conversational Agents | Personal and Ubiquitous Computing

Brandtzaeg, P. B. – and A. Følstad (2018) – Chatbots: changing user needs and motivations – *Interactions*, Vol. 25, (5), 38–43

Braun, V. – and V. Clarke (2012) – Thematic analysis. – *APA Handbook of Research Methods in Psychology* – 57–71

Bryant, A. (2023) – AI Chatbots: Threat or Opportunity? – *Informatics*, Vol. 10, (2), 49

Cabrero-Daniel, B. – and A. Sanagustín Cabrero (2023) – Perceived Trustworthiness of Natural Language Generators – *Proceedings of the First International Symposium on Trustworthy Autonomous Systems* – 1–9

Caldarini, G. et al. (2022) – A Literature Survey of Recent Advances in Chatbots – *Information*, Vol. 13, (1), 41

Chan, M.-Y. – and S.-M. Wong (2024) – A Comparative Analysis to Evaluate Bias and Fairness Across Large Language Models with Benchmarks – Open Science Framework

Chan, W. T. Y. – and C. H. Leung (2021) – Mind the Gap: Discrepancy Between Customer Expectation and Perception on Commercial Chatbots Usage – *Asian Journal of Empirical Research*, Vol. 11, (1), 1–10

Cravotta, R. (2003) – Uncovering the truth in benchmarks – *EDN*, Vol. 48, (22), 57–64

Dai, S. et al. (2024) – Bias and Unfairness in Information Retrieval Systems: New Challenges in the LLM Era – *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* – 6437–6447

Dam, S. K. et al. (2024) – A Complete Survey on LLM-based AI Chatbots – arXiv

Das, A. B. – and S. K. Sakib (2024) – Unveiling and Mitigating Bias in Large Language Model Recommendations: A Path to Fairness – arXiv

Dette, H. – and V. B. Melas (2010) – A note on all-bias designs with applications in spline regression models – *Journal of Statistical Planning and Inference*, Vol. 140, (7), 2037–2045

Dingler, T. et al. (2018) – Biased Bots: Conversational Agents to Overcome Polarization – *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers* – 1664–1668

Dinter, R. van et al. (2021) – A decision support system for automating document retrieval and citation screening – *Expert Systems with Applications*, Vol. 182, 115261

Duncan, C. – and I. Mcculloh (2024) – Unmasking Bias in Chat GPT Responses – *Proceedings of the 2023 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* – 687–691

Ebubechukwu, I. et al. (2024) – Dialogue You Can Trust: Human and AI Perspectives on Generated Conversations – arXiv

Egger, M. et al. (2022) – *Systematic Reviews in Health Research: Meta-Analysis in Context* – John Wiley & Sons, Incorporated, Newark, UNITED KINGDOM

Eisenmann, C. et al. (2024) – "Machine Down": making sense of human–computer interaction—Garfinkel's research on ELIZA and LYRIC from 1967 to 1969 and its contemporary relevance – *AI & SOCIETY*, Vol. 39, (6), 2715–2733

Ekechi, C. C. et al. (2024) – AI-INFUSED CHATBOTS FOR CUSTOMER SUPPORT: A CROSS-COUNTRY EVALUATION OF USER SATISFACTION IN THE USA AND THE UK – *International Journal of Management & Entrepreneurship Research*, Vol. 6, (4), 1259–1272

El Gharbaoui, O. E. et al. (2024) – Chatbots and Citizen Satisfaction: Examining the Role of Trust in AI-Chatbots as a Moderating Variable – *TEM Journal*, 1825–1836

Ernst, J. S. et al. (2023) – Bias Mitigation for Large Language Models using Adversarial Learning

Fan, X. et al. (2025) – User-Driven Value Alignment: Understanding Users' Perceptions and Strategies for Addressing Biased and Discriminatory Statements in AI Companions – arXiv

Feng, C. (Mitsu) et al. (2024) – From HAL to GenAI: Optimizing chatbot impacts with CARE – *Business Horizons*, Vol. 67, (5), 537–548

Ferrara, E. (2023) – Should ChatGPT be Biased? Challenges and Risks of Bias in Large Language Models – *First Monday*

Ferrara, E. (2024) – Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies – *Sci*, Vol. 6, (1), 3

Følstad, A. – and P. B. Brandtzæg (2017) – Chatbots and the new world of HCI – *Interactions*, Vol. 24, (4), 38–42

Gallegos, I. O. et al. (2024) – Bias and Fairness in Large Language Models: A Survey – *Computational Linguistics*, Vol. 50, (3), 1097–1179

Gong, Z. et al. (2024) – Enhancing Trust in LLM Chatbots for Workplace Support Through User Experience Design and Prompt Engineering – *AHFE Open Access*, Vol. 143

Greyling, C. (2024) – A Short History of Chatbots. < https://cobusgreyling.medium.com/a-short-history-of-chatbots-42a92a4cf296 >, retrieved 2.4.2025

Guo, Y. et al. (2024) – Bias in Large Language Models: Origin, Evaluation, and Mitigation – arXiv

Haugeland, I. K. F. et al. (2022) – Understanding the user experience of customer service chatbots: An experimental study of chatbot interaction design – *International Journal of Human-Computer Studies*, Vol. 161, 102788

Heersmink, R. et al. (2024) – A phenomenology and epistemology of large language models: transparency, trust, and trustworthiness – *Ethics and Information Technology*, Vol. 26, (3), 1–15

Hiebl, M. R. W. (2023) – Sample Selection in Systematic Literature Reviews of Management Research – *Organizational Research Methods*, Vol. 26, (2), 229–261

Holliday, D. et al. (2016) – User Trust in Intelligent Systems: A Journey Over Time – 164–168

Holroyd, J. (2012) – Responsibility for Implicit Bias – *Journal of Social Philosophy*, Vol. 43, (3), 274–306

Howard, A. – and J. Borenstein (2018) – The Ugly Truth About Ourselves and Our Robot Creations: The Problem of Bias and Social Inequity – *Science and Engineering Ethics*, Vol. 24, (5), 1521–1536

Huang, D. et al. (2024) – Can chatbot customer service match human service agents on customer satisfaction? An investigation in the role of trust – *Journal of Retailing and Consumer Services*, Vol. 76, 103600

Huang, P.-S. et al. (2019) – Reducing Sentiment Bias in Language Models via Counterfactual Evaluation

Huang, Y. et al. (2024) – TrustLLM: Trustworthiness in Large Language Models – arXiv

Kantharuban, A. et al. (2024) – Stereotype or Personalization? User Identity Biases Chatbot Recommendations – arXiv

Karjus, A. (2025) - Machine-assisted quantitizing designs: augmenting humanities and social sciences with artificial intelligence - *Humanities and Social Sciences Communications*, Vol. 12, (1), 1-18

Kellaghan, T. (2010) – Evaluation Research – In: Penelope Peterson et al. (ed.) – *International Encyclopedia of Education (Third Edition)* – 150–155

Lame, G. (2019) – Systematic Literature Reviews: An Introduction – *Proceedings of the Design Society: International Conference on Engineering Design*, Vol. 1, (1), 1633–1642

Lee, F. Y. – and T. J. Chan (2024) – Establishing Credibility in AI Chatbots: The Importance of Customization, Communication Competency and User Satisfaction – 88–106

Lex, A. et al. (2014) – UpSet: Visualization of Intersecting Sets – *IEEE Transactions on Visualization and Computer Graphics*, Vol. 20, (12), 1983–1992

Li, J. et al. (2023) – Determinants Affecting Consumer Trust in Communication With AI Chatbots: The Moderating Effect of Privacy Concerns – *Journal of Organizational and End User Computing*, Vol. 35, 1–24

Li, W. – and C. Zhang (2009) – Markov Chain Analysis – In: Audrey Kobayashi (ed.) – *International Encyclopedia of Human Geography (Second Edition)* – 407–412

Lippens, L. (2024) – Computer says 'no': Exploring systemic bias in ChatGPT using an audit approach – *Computers in Human Behavior: Artificial Humans*, Vol. 2, (1), 100054

Long, H. A. et al. (2020) – Optimising the value of the critical appraisal skills programme (CASP) tool for quality appraisal in qualitative evidence synthesis

Lu, J. (2024) – Enhancing Chatbot User Satisfaction: A Machine Learning Approach Integrating Decision Tree, TF-IDF, and BERTopic – Preprints

Luo, X. et al. (2019) – Frontiers: Machines vs. Humans: The Impact of Artificial Intelligence Chatbot Disclosure on Customer Purchases – *Marketing Science*, mksc.2019.1192

Markovitch, D. G. et al. (2024) – Consumer reactions to chatbot versus human service: An investigation in the role of outcome valence and perceived empathy – *Journal of Retailing and Consumer Services*, Vol. 79, 103847

Meduri, S. (2024) – Revolutionizing Customer Service : The Impact of Large Language Models on Chatbot Performance – *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, Vol. 10, (5), 721–730

Müller, L. et al. (2019) – Chatbot Acceptance: A Latent Profile Analysis on Individuals' Trust in Conversational Agents – *Proceedings of the 2019 on Computers and People Research Conference* – 35–42

Naik, D. et al. (2024) – Large Data Begets Large Data: Studying Large Language Models (LLMs) and Its History, Types, Working, Benefits and Limitations – 293–314

Narayan, M. et al. (2024) – Bias Neutralization Framework: Measuring Fairness in Large Language Models with Bias Intelligence Quotient (BiQ) – arXiv

Navigli, R. et al. (2023) – Biases in Large Language Models: Origins, Inventory, and Discussion – *J. Data and Information Quality*, Vol. 15, (2), 10:1-10:21

Ng, S. W. T. – and R. Zhang (2025) – Trust in AI chatbots: A systematic review – *Telematics and Informatics*, Vol. 97, 102240

Nicolescu, L. 1 et al. (2022) – Human-Computer Interaction in Customer Service: The Experience with AI Chatbots—A Systematic Literature Review – 1579

Page, M. J. et al. (2021) – The PRISMA 2020 statement: an updated guideline for reporting systematic reviews – *Systematic Reviews*, Vol. 10, (1), 89

Petersen, K. et al. (2008) – Systematic Mapping Studies in Software Engineering

Ryan, M. J. et al. (2024) – Unintended Impacts of LLM Alignment on Global Representation – arXiv

Salama, M. et al. (2017) – Chapter 11 - Managing Trade-offs in Self-Adaptive Software Architectures: A Systematic Mapping Study – In: Ivan Mistrik et al. (ed.) – *Managing Trade-Offs in Adaptable Software Architectures* – 249–297

Schmitt, A. (2022) – Examining Trust in Conversational Systems: Conceptual and Empirical Findings on User Trust, Related Behavior, and System Trustworthiness – *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society* – 912

Sethy, A. et al. (2023) – AI: Issues, concerns, and ethical considerations – *Toward Artificial General Intelligence: Deep Learning, Neural Networks, Generative AI* – 189–211

Sheng, E. et al. (2021) – Societal Biases in Language Generation: Progress and Challenges – In: Chengqing Zong et al. (ed.) – *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* – 4275–4293

Shieber, S. M. (1994) – Lessons from a Restricted Turing Test – arXiv

Shokrollahi, O. (2023) – Intersectional Bias Mitigation in Pre-trained Language Models: A Quantum-Inspired Approach – *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management* – 5181–5184

Sonboli, N. et al. (2021) – Fairness and Transparency in Recommendation: The Users' Perspective – *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization* – 274–279

Sullivan, P. M. (2004) – Frege's Logic – In: Dov M. Gabbay and John Woods (ed.) – *Handbook of the History of Logic* – 659–750

Talboy, A. N. – and E. Fuller (2023) – Challenging the appearance of machine intelligence: Cognitive bias in LLMs and Best Practices for Adoption – arXiv

Tay: Microsoft issues apology over racist chatbot fiasco (2016) – *BBC News*. <https://www.bbc.com/news/technology-35902104 >, retrieved 6.3.2025.

UNESCO – Recommendation on the Ethics of Artificial Intelligence (2022) – UNESCO. <https://www.unesco.org/en/articles/recommendation-ethics-artificial-intelligence >, retrieved 17.3.2025.

Urman, A. – and M. Makhortykh (2025) – The silence of the LLMs: Cross-lingual analysis of guardrail-related political bias and false information prevalence in ChatGPT, Google Bard (Gemini), and Bing Chat – *Telematics and Informatics*, Vol. 96, 102211

Virvou, M. et al. (2024) – VIRTSI: A novel trust dynamics model enhancing Artificial Intelligence collaboration with human users – Insights from a ChatGPT evaluation study – *Information Sciences*, Vol. 675, 120759

Wahbeh, A. et al. (2023) – Perception of Bias in ChatGPT: Analysis of Social Media Data – *2023 IEEE Global Conference on Artificial Intelligence and Internet of Things (GCAIoT)* – 34–39

Wang, A. et al. (2019) – GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding – arXiv

Wang, Z. et al. (2024) – History, development, and principles of large language models: an introductory survey – *AI and Ethics*

Wolf, T. et al. (2020) – HuggingFace's Transformers: State-of-the-art Natural Language Processing – arXiv

Wuenderlich, N. – and S. Paluch (2017) – A Nice and Friendly Chat with a Bot: User Perceptions of AI-Based Service Agents – *ICIS 2017 Proceedings*

Xie, C. et al. (2024) – Does Artificial Intelligence Satisfy You? A Meta-Analysis of User Gratification and User Satisfaction with AI-Powered Chatbots. – *International Journal of Human-Computer Interaction*, Vol. 40, (3), 613–623

Xue, J. et al. (2023) – Bias and Fairness in Chatbots: An Overview

Yao, X. – and Y. Xi (2024) – Pathways linking expectations for AI chatbots to loyalty: A moderated mediation analysis – *Technology in Society*, Vol. 78, 102625

Yuan, C. W. (Tina) et al. (2023) – Contextualizing User Perceptions about Biases for Human-Centered Explainable Artificial Intelligence – *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* – 1–15

Zhan, Z. (2025) – Comparative Analysis of TF-IDF and Word2Vec in Sentiment Analysis: A Case of Food Reviews – *ITM Web of Conferences*, Vol. 70, 02013

Zhou, R. (2024) – Empirical Study and Mitigation Methods of Bias in LLM-Based Robots – *Academic Journal of Science and Technology*, Vol. 12, (1), 86–93

Appendices

PRISMA 2020 Abstract Checklist

Section & Topic	Item	Checklist Item	Tailored Response to this Thesis
TITLE	1	Identify the report as a systematic review.	The study should be titled to reflect it is a *Systematic Literature Review*.
BACKGROUND	2	Provide an explicit statement of the main objective(s) or question(s) the review addresses.	To review and map the historical trajectory of bias in LLM-powered chatbots, and the impacts it may have had on user satisfaction
METHODS	3	Specify the inclusion and exclusion criteria for the review.	Not explicitly stated in the abstract text provided. Should mention inclusion of studies on AI-powered chatbots, user satisfaction, and bias from past to present.
	4	Specify the information sources (e.g., databases, registers) used to identify studies and the date when each was last searched.	Sources not named; abstract should state which databases were searched and when.
	5	Specify the methods used to assess risk of bias in the included studies.	If not included in the abstract it can be overlooked and checked in the paper
	6	Specify the methods used to present and synthesise results.	If not included in the abstract it can be overlooked and checked in the paper
RESULTS	7	Give the total number of included studies and participants and summarise relevant characteristics of studies.	If not included in the abstract it can be overlooked and checked in the paper
	8	Present results for main outcomes, preferably indicating the number of studies and participants for each.	If not included in the abstract it can be overlooked and checked in the paper
DISCUSSION	9	Provide a summary of the limitations of the evidence included in the review (e.g.	If not included in the abstract it can be overlooked and checked in the paper

		risk of bias, inconsistency, and imprecision).	
	10	Provide a general interpretation of the results and important implications.	Findings in the study should inform the development/study of effective bias mitigation strategies and/or contribute to AI fairness and user experience design.
OTHER	11	Specify the primary source of funding for the review.	Was not applicable in this instance
	12	Provide the registration number and name of the registry.	Was not applicable in this instance

Artificial Intelligence Assistance Declaration

I, Olayinka Vicente, hereby declare that I have utilized the free version of the Artificial Intelligence (AI) tool, ChatGPT, in the preparation of my research thesis titled "Mapping the Impact of Bias in Large Language Model Chatbots on User Satisfaction", aiming to enhance the accuracy, efficiency, and depth of my work. This declaration aims to provide a clear and transparent illustration of the specific ways in which ChatGPT has contributed to my research, ensuring academic integrity and proper acknowledgment of technological assistance.

Scope of ChatGPT Utilization

Interpreting Complex Literature: Understanding complex concepts in academic literature can be difficult, however, ChatGPT can help researchers simplify these notions for better understanding. For instance, by using ChatGPT to provide a simple explanation to the "Bias Intelligence Quotient (BiQ)" as discussed by Narayan et al.'s (2024) would allow researchers to understand the concept and its relevance and/or relationship to bias mitigation strategies.

ChatGPT Input Command: Provide a simple explanation of the idea behind Narayan et al.'s (2024) "Bias Intelligence Quotient (BiQ)" Bias from their paper "Neutralization Framework: Measuring Fairness in Large Language Models with Bias Intelligence Quotient (BiQ)".

Clarifying Analysis Approaches suitable for Systematic Literature Reviews (SLRs): Researchers conducting SLRs usually face challenges in selecting and applying the appropriate analysis methods to synthesize diverse findings. However, with the aid of ChatGPT, researchers could receive clear explanations and illustrative examples of different analytical approaches used in SLRs. For example, thematic analysis identifies patterns, meta-analysis combines statistical results, or bibliometric analysis foe mapping research trends. The support allows researchers to easily align their strategy with their data types and review objectives.

ChatGPT Input Command: Explain the difference between thematic analysis, meta-analysis and narrative synthesis.

Paraphrasing Content Professionally/Academically: ChatGPT could help improve professional communication which is important for effectively conveying ideas in academic writings. For example, a paraphrased version of the definition of chatbots by Caldarini et al. (2022, 1) of chatbots as "intelligent conversational computer programs that mimic human conversation in its natural form" could be: "Caldarini et al. (2022, 1) defined chatbots as intelligent conversational computer programs that are designed to mimic natural human conversations to enable automated online guidance and support".

ChatGPT Input Command: Rephrase the below paragraphs following an academic tone for better clarity.

Clarity and Logical Progression: Maintaining clarity and a well-structured progression is important in scholarly writing to facilitate comprehension and engagement from readers. For example, In the context of this thesis, the order went from the introduction to the research gap and objectives, then the methodology, analysis of the findings and finally the synthesis, which enhanced the coherence and comprehensibility of the paper.

ChatGPT Input Command: Are the below paragraphs understandable and do they have a logical flow?

Ensuring Grammatical Accuracy: It is important to ensure grammatical precision and linguistic coherence in academic writing in order to enhance its readability and credibility. With the assistance of ChatGPT, researchers can refine sentence structures, correct grammatical errors and ensure consistency in the writing's style and tone; thus improving the overall quality of the writing.

ChatGPT Input Command: Review the below paragraph for sentence structure, grammatical accuracy and overall clarity. Also ensure the tone of the ideas remain formal and academic.

This Artificial Intelligence (AI) Assistance Declaration underscores a commitment to transparency, ethical integrity, and the responsible use of technology in advancing academic knowledge.

Olayinka Vicente
20.05.2025