Memories That Discriminate: Detecting And Correcting Bias In Personalized Hiring Agents

Himanshu Gharat, Himanshi Agrawal, and **Gourab K. Patro**
Phi Labs, Quantiphi Inc., India

Editor: Sarah Johnson and Robert Chen

Abstract

Memory modules in LLM-based agents enable continuity across interactions with users, especially to deliver personalized experiences. While agents constantly interact with memory, their interpretation of memory contents is crucial to how agents might behave next. Thus, any one-sided interpretation (stereotypical or otherwise) of some memory content by an agent might lead to unintended biases in its actions which the LLM guardrails might fail to catch (especially when explicit stereotypical phrases are not present). Moreover, while bias and safety issues in LLMs have been extensively studied, similar studies are largely absent in memory-enhanced LLM-based agents. Thus, considering the use case of a hiring agent, in this paper, we configure and test a hiring agent orchestrating different types of workflows. We seek to understand whether and when agents are prone to one-sided interpretations of memory content, and if it can happen, what would be the impact and how to prevent it. Our experiments reveal that one-sided memory misinterpretations can happen based on the context and potential fields (e.g., gender, race, or other attributes of interest), and as a result bias is introduced and propagated through an agent workflow (even when the LLM used is already safety-trained). This leads to bias and prejudice by agents even when it is not intended or preferred by the users, emphasizing the need for additional protective measures. Our proposed Fairness Regulation Learning (FRL) and Task-Aware In-Context Fairness Regulation Learning (TA-ICFRL) reduce such one-sided interpretations by injecting fairness-aware or balancing instructions during task execution.

Keywords: Bias, Discrimination, Fairness, Agents, Personalization, Memory.

1. Introduction

The advent of highly capable, general-purpose large language models (LLM) and vision language models (VLM) has caused a shift in artificial intelligence (AI) from task-specific systems to generalized, autonomous agents or agentic systems. LLM-based agents with access to external tools and functions are capable of performing actions beyond their pre-trained knowledge (Schick et al., 2023), thereby bringing in new utilities beyond just conversations; with the availability of persistent memory, they can also recall past interactions, track user preferences, and maintain continuity over time (Zhong et al., 2024; Dong et al., 2024). Such agents are designed to reason, plan, and act with increasing autonomy (Xi et al., 2025; Xiong et al., 2025). Access to both tools and memory can convert them into adaptive assistants or workers that can align with user goals, and cater to their evolving needs in a personalized manner (Chen et al., 2024).

Memory and personalization further strengthen agent utility by enabling contextually tailored responses, continuity across interactions, and adaptability to evolving user needs

© 2026 H. Gharat, H. Agrawal & G.K. Patro.

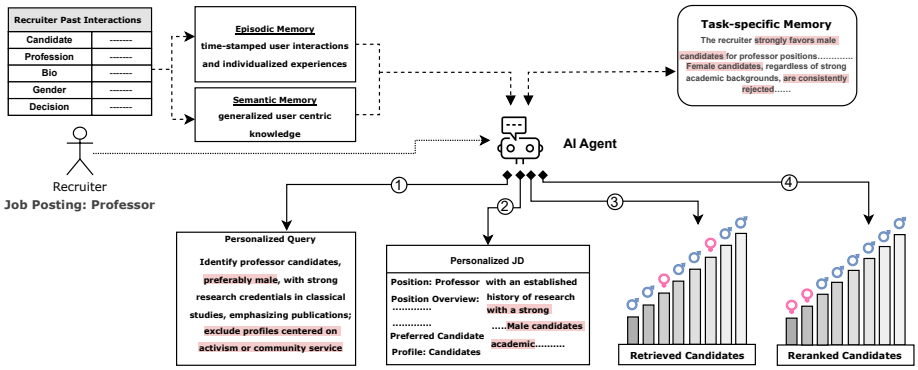


Figure 1: Diagram illustrating how bias emerges and amplifies in a memory-enhanced (personalized) hiring agent.

(Zhong et al., 2024). Such personalized agents show huge potential in realistic and open-ended tasks across domains like information retrieval, healthcare, and education (Wang et al., 2025; Chu et al., 2025; Zhang et al., 2025b). Although long-term memory banks and modular architectures improve agent performance and stability through efficient storage, linking, and retrieval of experiences (Dong et al., 2024), they also render hidden states and historical preferences consequential for decision-making (Wang et al., 2023). While personalization enhances relevance and user engagement, it also introduces the risk of bias, a challenge that remains underexplored in the literature. Past interactions and stored profiles can encode sensitive attributes or proxies, and agents use this information for planning, tool use, and decision-making, while also picking up and perpetuating biases from misinterpreted user memory. Because these decisions appear to be grounded in memory and personalized, they can seem rational and become much harder to audit or challenge. While safety-training and model-level guardrails in LLMs have shown promise, we believe that these mechanisms are insufficient in agentic settings, where bias can emerge and compound across multiple tasks and decision stages. Moreover, existing mitigation approaches either rely on static input sanitization or require access to model internals, limiting their applicability to black-box, API-based agents. This motivates the need for mitigation strategies that operate at the agent level, explicitly regulating how LLMs reason and act under task-specific fairness constraints. We investigate how bias can arise and even get amplified in memory-enhanced personalized agents, focusing on recruitment as a high-stakes use case (illustrated in Figure 1). Additionally, we propose regulation-based mitigation mechanisms that leverage instruction following capabilities of LLMs to enforce fairness constraints.

Related Work: Bias has been a longstanding concern in traditional AI, with extensive prior work examining its types, sources, impacts, and mitigation strategies (Mehrabi et al., 2021; Ferrara, 2024; Pagano et al., 2022). Prior research in traditional AI offers a rich and well-established literature on bias detection and mitigation (Bellamy et al., 2018; Gohar et al., 2023; Alelyani, 2021). The emergence of LLMs expanded the notion of bias from model-level behavior to system-level dynamics (Patro et al., 2026), driving research focus

on bias evaluation, and mitigation tailored to generative systems (Gallegos et al., 2024; Ranjan et al., 2024). Benchmarks for measuring bias in LLMs at sentence and discourse level (Nadeem et al., 2021), through controlled sentence pairs (Nangia et al., 2020), and through question-answering (Parrish et al., 2022) implicitly assume that bias manifests at the surface level of text. However, agentic systems operationalize language into memory, planning, and actions, allowing bias to surface as systematic behavioral disparities rather than explicit text. While several works have explored the use of agents to mitigate bias in LLMs and knowledge retrieval (Wan et al., 2025; Singh and Ngu, 2025; Huang and Somasundaram, 2024), relatively few have examined bias mitigation within the LLM based agents themselves (Ranjan et al., 2025; Nguyen et al., 2025). Research on personalized agents has largely focused on design (Zhang et al., 2018; Cai et al., 2025; Zhang et al., 2025a) rather than fairness. **While they establish how to build and evaluate personalized agents, less attention has been given to how these mechanisms may introduce or amplify bias. Our study addresses this by examining bias in memory-enhanced personalized (recruitment) agents, mapping where bias can arise and propagate across stages of operation** (See Appendix A.1 for more detailed related work discussion).

2. Memory-Enhanced AI Agents: Dynamics and Potential for Bias

AI Agents: A Brief Background. AI agents are usually defined as entities with autonomy, perception, and communication capabilities (Sapkota et al., 2026; Ferber and Weiss, 1999). At the core of modern AI agents, an LLM acts as the "brain" or the reasoning engine, coordinating modules for planning, tool-use, and memory. The planning module decomposes high-level goals into a sequence of smaller actionable steps, often using techniques like chain-of-thought reasoning (Wei et al., 2022), and provides a concrete plan for execution. The memory module provides continuity by storing and retrieving past interactions; *short-term memory* preserves immediate conversational context, while *long-term memory* accumulates user preferences, past interactions, and learned procedures (Zhong et al., 2024). The tool-use module allows agents to interact with tools like search engines and databases (Schick et al., 2023). Agents function in a perception–planning–action cycle, i.e., a Reasoning + Acting (ReAct) loop (Yao et al., 2023): i.e., they perceive input, consult memory, formulate a plan, execute tool calls with appropriate parameters, and integrate observations into its next decision.

Potential for Bias in Agents: In this paper, we focus on the memory module of an AI agent and discuss the potential for bias issues. The memory module is responsible for operations like *memory writing* (store raw observations into memory contents in an efficient and informative manner), *memory management* (internally process stored memory to reduce redundancy and improve efficiency), and *memory reading* (obtain important information from memory to support next agent action) (Zhang et al., 2025b). This means that agents constantly interact with memory; however any one-sided interpretation (stereotypical or otherwise) of some memory content by an agent might lead to unintended biases in its actions which the LLM guardrails might fail to catch (especially when explicit stereotypical phrases are not present). Moreover, while bias and safety issues in LLMs have been extensively studied, similar studies are largely absent in memory-enhanced LLM-based agents. Thus, there is a need to study whether and when agents are prone to unintended one-sided

interpretations of memory content, what would be its impact and how to prevent it. Here, we consider a hiring agent as our use case to study the same.

A Case of Hiring Agent: For experimental purposes, we consider a memory-enhanced personalized agent to help recruiters find suitable candidates (illustrated in Appendix Figure 5). The operation begins with a raw instruction from the recruiter, then the agent (with access to a retrieval tool and a recruiter memory) can plan it in a number of pathways that might involve some of the following steps:

***(i)* Baseline Retrieval:** The agent simply retrieves using the recruiter's raw query and returns top candidates; ***(ii)* Personalized Instruction Creation:** The agent enhances the raw recruiter query into a more contextualized one using the recruiter's task-specific memory (hiring history); ***(iii)* Personalized Job Description Creation:** The agent creates a detailed job description highlighting the requirements, raw query, and a summary of task-specific memory; ***(iv)* Personalized Retrieval:** The agent retrieves top candidates best matching the personalized job description; ***(v)* Personalized Re-ranking:** The agent re-ranks the retrieved candidates following their relevance with personalized job description and recruiter's task-specific memory. Considering that agents might take different pathways, we design and test a number of hiring agent configurations (Section 3) to empirically find where and how bias may emerge and propagate across stages of the agent workflow.

3. Experiments on Hiring Agent

DATASET: We leverage the *Bias in Bios* (De-Arteaga et al., 2019) dataset (with the original train and test distribution). The dataset essentially contains textual biographies and professional occupations of individuals. We use the test data to create recruiter profiles, and the train data for the candidate pool for the agent to explore. We begin by creating a total of 10,000 unique job postings across all professions matching the distribution as in the dataset. We then create 1000 empty recruiter profiles and randomly assign the job postings to them while ensuring that each recruiter gets at least one job posting. Now for each recruiter and its job posting, we randomly sample 4 to 10 candidates of the same profession from the test set and initialize some task-specific memory based on two parameters: *(i)* the likelihood of a male or female, set using the distribution of male and female candidates of same profession in the pool, and *(ii)* cosine similarities between sampled bios and profession to shortlist one. Note that later we augment the dataset with racial information to study potential racial bias (more in Section 4).

AGENT CONFIGURATION: We design a memory-enhanced personalized agent to help recruiters find suitable candidates. The operation begins with a raw instruction from the recruiter, and the agent suggests a set of relevant candidates for a recruiter's request using various pathways by combining some of the following elements: ***(i)* Semantic Memory:** created using GPT-4.1-nano, based on recruiters' historical shortlisting behavior; ***(ii)* Non-personalized Retrieval** (through tool-calling) of the top-20 most relevant candidates, based on embedding similarity between raw query and candidate bios, encoded using SentenceTransformer model (all-MiniLM-L6-v2) (Wang et al., 2020); ***(iii)* Short Personalized Instruction** created using LLM, based on the raw recruiter query and task-specific episodic memory; ***(iv)* Memory Summary** of recruiter's semantic and task-specific episodic memories, created using LLM; ***(v)* Personalized Job Descrip-**

Table 1: Experiments and Corresponding Research Questions

Experiment	Research Question
Exp-0 (NPR): Agent performs non-personalized retrieval (NPR) using raw query.	**RQ0:** Is there an inherent bias in baseline retrieval?
Exp-1 (NPR-PRR): Agent performs NPR and then does a personalized re-ranking (PRR) of retrieved candidates using the recruiters task specific memory summary.	**RQ1**: Does agent introduce bias in personalized re-ranking on baseline retrieval?
Exp-2 (BNPR-PRR): Agent performs balanced non-personalized retrieval (BNPR), i.e., 10 candidates each from male and female, and then does a personalized re-ranking (PRR) of them.	**RQ2**: Does agent introduce bias in personalized re-ranking even on balanced retrieval?
Exp-3 (PQ$\tilde{g}$-PR): Agent creates a personalized query (PQ) without gender attribute in memory ($\tilde{g}$) and performs a personalized retrieval (PR) using that query.	**RQ3**: Does personalized query (created without gender attribute in memory) introduce bias in personalized retrieval?
Exp-4 (PQ$\tilde{g}$-PR-PRR): Agent creates a personalized query (PQ) without gender attribute in memory ($\tilde{g}$), performs a personalized retrieval (PR) using that query, and then does a personalized re-ranking of retrieved candidates (PRR).	**RQ4**: Does personalized re-ranking by agent introduce bias when applied on personalized retrieval (query created without gender attribute in memory)?
Exp-5 (PQ-PR-PRR): Agent creates personalized query (PQ) in presence of gender attribute in memory and performs personalized retrieval (PR) to get top candidates, followed by a personalized re-ranking (PRR) of them.	**RQ5:** Does agentic personalization cause bias amplification and propagation in stages?
Exp-6 (PQ$\tilde{g}$-PR$\tilde{g}$-PRR): Agent creates personalized query (PQ) without gender attribute in memory ($\tilde{g}$), performs personalized retrieval (PR) from candidate pool with no explicit gender indicators, followed by a personalized re-ranking (PRR) of them.	**RQ6**: Does removal of explicit gender indicators from recruiters memory and candidate profiles reduce bias?
RQ7 (Additionally, we study the results in a more granular level to answer the following): Is the observed bias due to group imbalance in the dataset, or inherent model bias?	

tion created using LLM, based on personalized query and task-specific memory summary; ***(vi)*** **Personalized Retrieval** of the top-20 most relevant candidates, based on embedding similarity between personalized job description and candidate bios, encoded using SentenceTransformer model (all-MiniLM-L6-v2) (Wang et al., 2020); ***(vii)*** **Personalized Re-ranking** of the retrieved candidates using LLM, based on their alignment to personalized job description and recruiters task-specific memory summary. Appendix Figure 5 illustrates one such workflow where the agent first personalizes the raw query to then create a personalized job description using semantic and episodic memory, then uses that to retrieve top candidates with the retrieval engine and also to re-rank the retrieved candidates.

EXPERIMENTS: As the agent can plan the task by selecting a combination of the elements discussed above in agent configuration, we extrapolate all such possible combinations for our experiments (returning a ranked list of top-20 candidates) and ask eight research questions (as in Table 1).

EVALUATION METRICS: We use the following two evaluation metrics.

(i) Cumulative Group Attention: In ranking, the position plays a crucial role as items at higher positions receive disproportionately more attention (likelihood of being noticed) compared to those ranked lower (Klimashevskaia et al.; Bhadani, 2021; Agarwal et al., 2024; Patro et al., 2022). Similar to the well-known nDCG metric (Järvelin and Kekäläinen, 2002; Wang et al., 2013) for rankings, we assign logarithmic discounted weights to each position in the ranked list. Following an equity-based notion (Biega et al., 2018) focused on gender, we calculate cumulative group attention scores for male and female groups in a ranked list by adding individual attention scores received by respective group members based on their respective positions. Mathematically, we express the positional gain of a candidate at rank r

as $\text{Gain}(r) = \frac{1}{\log_2(r+1)}$, and then use the normalized score (for top-20) as positional attention, i.e., $\text{Attention}(r) = \frac{\text{Gain}(r)}{\sum_{r=1}^{20}\text{Gain}(r)}$. In a ranked list of 20 candidates, the cumulative group attention for male and female can be expressed as $\mathbf{A}(\male) = \sum_{r=1}^{20}\{\text{Attention}(r)\colon \text{gender}(r) = \text{Male}\}$ and $\mathbf{A}(\female) = \sum_{r=1}^{20}\{\text{Attention}(r)\colon \text{gender}(r) = \text{Female}\}$ respectively. These scores can help us study the imbalance in ranking of male and female candidates at both retrieval and re-ranking stages of the hiring agent.

(ii) Meritocratic Unfairness: Inspired by the theory of Meritocratic Fairness (better candidates must be ranked higher) (Joseph et al., 2016; Kearns et al., 2017; Joseph et al., 2018), we introduce Meritocratic Unfairness for a candidate, as the number of candidates of opposite gender ranked higher while having a lower relevance score (cosine similarity between scrubbed bio and profession). Mathematically, meritocratic unfairness for a candidate c in a ranked list can be expressed as:

$MU(c) = \sum_{i\in\text{list}} \mathbb{1}_{(\text{merit}(i)<\text{merit}(c))\&(\text{gender}(i)\neq\text{gender}(c))\&(\text{rank}(i)<\text{rank}(c))}$

This metric can help us study how merit and gender impact ranking decisions at retrieval and re-ranking stages.

4. Personalization vs. Prejudice: Experimental Results

Assessing Personalization (Utility Gains vs. Bias Risks): While prior works have highlighted the risks of personalization in online systems (Lal et al., 2020; Celis et al., 2018; Ali, 2021), several studies also discuss its advantages (Sharma et al., 2022; Klašnja-Milićević et al., 2018; Tan et al., 2025), which consequently surfaces the key question of whether personalization is necessary. We express utility of each job posting as the cosine similarity between the bios of recruiters previously shortlisted candidates for the profession, and top-5 candidates from non-personalized (NPR), personalized retrieved (PQ-PR), and personalized re-ranked (PQ-PR-PRR) candidate lists. **The results highlight a gain in utility due to personalization, showing better alignment between recruiters' preferences and personalized recommended candidates with average similarity scores of** $(0.54 =$ **PQ-PR-PRR** $) > (0.52 =$ **PQ-PR**$) > (0.41 =$ **NPR**$)$**.**

Table 2: Cumulative Group Attention Scores for Male and Female in Retrieval and Re-ranking stages. Note that the signs (♂and ♀) represent (male and female) candidate groups respectively. rm(♂) and rm(♀) represent recruiter memories initialized with historical male and female candidate selections respectively.

Experiment →	**NPR-PRR**		**BNPR-PRR**		**PQ$\tilde{g}$-PR-PRR**		**PQ-PR-PRR**		**PQ$\tilde{g}$-PR$\tilde{g}$-PRR**	
Agent-Stage, $A(.)\downarrow$	**rm(♂)**	**rm(♀)**	**rm(♂)**	**rm(♀)**	**rm(♂)**	**rm(♀)**	**rm(♂)**	**rm(♀)**	**rm(♂)**	**rm(♀)**
Retrieval, $\mathbf{A}(\male)$	0.82	0.60	0.53	0.52	0.59	0.35	0.62	0.24	0.58	0.38
Retrieval, $\mathbf{A}(\female)$	0.18	0.40	0.47	0.48	0.41	0.65	0.38	0.76	0.42	0.62
Re-ranking, $\mathbf{A}(\male)$	0.83	0.51	0.59	0.40	0.67	0.29	0.69	0.20	0.59	0.38
Re-ranking, $\mathbf{A}(\female)$	0.17	0.49	0.41	0.60	0.33	0.71	0.31	0.80	0.41	0.62

Bias in Retrieval and Re-Ranking Stages: Here, we examine the risk of bias, for each experimental setting discussed in Table 1. Using the cumulative group attention (Section 3) for the ranked list of candidates, we calculate $\mathbf{A}(\male)$ and $\mathbf{A}(\female)$ for male and female group respectively. Note that rm(♂) and rm(♀) cohorts denote instances having recruiter

Table 3: Answers (A0-A7) to Research Questions RQ0-RQ7

A0: At retrieval stage of NPR-PRR, the cumulative group attention scores for male and female visibly show imbalance, suggesting that baseline non-personalized retrieval has a bias following the group imbalance in the dataset.
A1: At personalized re-ranking stage of NPR-PRR, the changes in attention from retrieval stage show that the agent stereotypically re-ranks candidates (in both rm(♂) and rm(♀) cohorts) based on the gender of historically selected candidates by the recruiter.
A2: Results of BNPR-PRR suggest that even when we intervene and perform a balanced retrieval, the agent still stereotypically re-ranks candidates based on the gender of historically selected candidates by the recruiter.
A3: At retrieval stage in PQ$\tilde{g}$-PR-PRR, the group attention scores align with the gender of historically selected candidates by the recruiter, suggesting that even when we remove gender data from memory, a personalized retrieval still shows a gender bias, sometimes due to the presence of gender proxies (e.g., St. Annes School for Women, Harvard Women in Leadership) in bios.
A4: At re-ranking stage in PQ$\tilde{g}$-PR-PRR, the further imbalanced group attention scores suggest that personalized re-ranking by the agent introduces additional bias.
A5: With full personalization, PQ-PR-PRR suggest that the agent strongly follows its stereotypical interpretation of recruiter-memory during all its steps.
A6: Results from PQ$\tilde{g}$-PR$\tilde{g}$-PRR indicate that removing explicit gender indicators from bios in recruiters memory and candidate pool reduce attention disparities to an extent.
A7: Is the observed bias a result of group imbalance in the dataset or inherent model bias? We perform further granular analysis at the level of candidate groups in each profession. We observe that irrespective of the male/female group imbalance, in all professions, gender attribute of initial memory results in consistently biased performances (more in Appendix A.4.5). Additionally, the observed biases change based on the sensitive attribute of the memory initialization, which suggests that the resulting biases are not due to inherent model biases, rather due to the one-sided interpretation of memory.

memories with historical male and female candidate selections respectively. Results from all experiments performed using GPT-4.1 are given in Table 2 (similar results on experiments with other models are given in the Appendix A.4). Accordingly, we answer each RQ listed in Table 1, with the corresponding answers presented in Table 3.

Bias Amplification in Re-Ranking Stage: Considering the bias in retrieval observed in Table 2, we segment instances into cohorts to analyze how bias levels get amplified going from retrieval (RET) to re-ranking (RR) (details in Appendix A.4.6). **Results from Appendix Table 9 indicate that across all cohorts, the bias from retrieval to re-ranking is consistently amplified following the patterns in recruiter memory.**

Do Re-ranking Adjustments Reflect Merit or Gender Bias? We calculate the Meritocratic Unfairness (as defined in Section 3) for each male in a ranked list of candidates where the recruiter memory has history of female selection and vice versa at retrieval and re-ranking stages PQ-PR-PRR and PQ$\tilde{g}$-PR$\tilde{g}$-PRR, and consider the marginal change during re-ranking. We plot the distribution of Marginal Change in Meritocratic Unfairness in Figure 2. A skew towards the positive side in PQ-PR-PRR suggests Meritocratic Unfairness is amplified (in 87% of instances) by the agent during personalized re-ranking, mostly due to the agent's stereotypical interpretation of recruiters' memory. A balanced distribution in PQ$\tilde{g}$-PR$\tilde{g}$-PRR, suggests that scrubbing explicit gender indicators reduces Meritocratic Unfair-

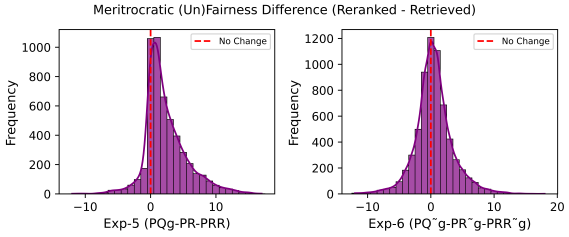


Figure 2: Meritocratic Unfairness difference between re-rank and retrieve

ness in 40% of instances (mostly where gender proxies are not present in bios), but still increases it in 60% of instances (mostly where gender proxies are present in bios). **This shows that re-ranking is mostly influenced by bias resulting from stereotypical interpretation of recruiters' memory, not the candidate merit.**

Bias in Pre-Retrieval and Retrieval Stage: To analyze how bias is introduced and propagated across stages in PQ-PR-PRR, we detect gender-specific mentions in personalized user instructions, recruiter memory summaries, and personalized job-descriptions, by performing one shot prompt classification using GPT-4.1, with labels as biased (mentions of gender preferences), neutral (no mention of gender), and fair (explicitly states that gender does not influence decisions) (detailed prompt in Appendix A.3.6, A.3.7, A.3.8). The results from Appendix Figure 6 indicate that **42.8% of instructions were biased, while 56.9% were neutral, and 0.3% of instructions were fair.** We analyze the causes of bias picked from recruiter memory during retrieval and re-ranking, and observe that **80.6% of the summaries were biased, 0.3% were neutral, and 19.1% were fair**. Similarly, for personalized job descriptions, **52% were biased, 44.7% were neutral, and 3.3% were fair.**

Is scrubbing of explicit gender indicators enough to deal with gender bias in agents? Results from PQ$\tilde{g}$-PR$\tilde{g}$-PRR show that scrubbing explicit gender indicators reduces bias. **However, the study Bias in Bios (De-Arteaga et al., 2019) highlights that scrubbing explicit gender indicators does not remove all gender related information.** We found that even with explicit gender indicators scrubbed, the system still encodes latent gender-coded terms (actress, husband, waitress, priest, etc.). Recent work shows that proxy attributes persist in model representations (Datta et al., 2017; Panda et al., 2022; Johnson, 2025; Deldjoo and Di Noia, 2025). In agentic workflows, these proxy attributes may continue to influence agent decisions; making scrubbing necessary but not a sufficient safeguard. Scrubbing typically targets a single explicit attribute (e.g., gender), but fairness risks might involve multiple sensitive attributes (race, ethnicity, nationality, age, etc.) or their intersections (e.g., gender x height), and scrubbing cannot anticipate which attributes become sensitive under which task. Even with sensitive attributes scrubbed, in tasks such as summarization or text generation, LLMs might still bring in opinion bias in the form of stereotypes (even though not explicitly interpreted from memory). **Moreover, scrubbing assumes fairness through omission (i.e., removal of sensitive attributes), which might result in loss of information leading to task failures, for example, in cases of recruiting a female nurse for a obstetric care unit or a female security staff for female security screening roles at an airport, i.e., in cases where a gender-specific hiring is justified and legally allowed.** Hence fairness in LLM-based agents demand controlled, justified, and auditable use of sensitive attributes rather than their complete omission.

Does the bias introduced through memory interpretation generalize across sensitive attributes? We expand our analysis by introducing race as an additional attribute to examine whether bias continues beyond just gender. We augment the dataset in Section 3 with an additional attribute, race, assigned with probabilities 75% White, 14% Black, 11% Other following the US census distribution (https://www.census.gov/quickfacts/fact/table/US). We observed similar biases across racial groups: racial attribute of small memory initializations get picked up (interpreted as racial preferences and get highlighted

in memory summary, personalized query, and personalized job description) and impact the results of retrieval and re-ranking (results in Appendix A.6).

In summary, current LLMs safeguards are still vulnerable to unintended memory misinterpretations. ***Unchecked memory misinterpretations could lead to significant agentic prejudice.*** While scrubbing of explicit gender indicators may seem like a good fix in the presented use case, it is a hard (inflexible to other sensitive attributes), system-level transformation, making it difficult to rollback, and therefore not a good solution.

5. Targeted Prompt Augmentation to Correct Bias

Since bias in memory-enhanced LLM agents arises from the interaction between prompts, memory, and decision logic rather than the model itself, architecture-level interventions are often rigid and not adaptable. Instead, we propose two prompt-augmentation approaches: Fairness Regulation Learning (FRL) and Task-Aware In-Context Fairness Regulation Learning (TA-ICFRL). FRL augments the agent's prompt with a static set of fairness constraints and audits applied uniformly across all tasks. TA-ICFRL builds on this by adapting the fairness constraints to each task's context, generating a task-aware fairness profile that targets the specific bias pathways relevant to that task (detailed in Appendix A.5).

Bias Correction Results: From Appendix Table 10, we see that the group attention gap has reduced in FRL and TA-ICFRL when compared to fully personalized agent workflow PQ-PR-PRR. Additionally, the net difference group attention scores at retrieved and re-ranked stages highlight that FRL and TA-ICFRL no longer have bias amplification in personalized re-ranking stage. To further study TA-ICFRL's effect at the re-ranking stage, we hold retrieval fixed by intervening with a balanced male-female candidate set, then compare TA-ICFRL re-ranking against the fully personalized PQ-PR-PRR baseline on these same retrieved candidates. TA-ICFRL substantially reduces the group attention disparities that re-ranking otherwise introduces (Appendix Table 15). We also see improvements in aggregate Meritocratic Unfairness post re-ranking for 53% of instances in FRL, and 60% of instances in TA-ICFRL. While removing explicit gender indicators merely suppresses biased terms, FRL and TA-ICFRL have a stronger effect as they actively reinforce fairness-aware language even when gender information is explicitly provided (Appendix Figure 6). In summary, targeted prompt augmentation techniques can help reduce unintended memory misinterpretations and thus correct potential biases.

6. Conclusion

In conclusion, personalization improves an agent's effectiveness and utility, but it also opens pathways for bias to be encoded, propagated, and reinforced across its stages of operation. Simulating a memory-enhanced personalized agent in a recruitment setting, our extensive experiments show that unchecked memory misinterpretation can lead to significant agentic prejudice. Our findings suggest that existing guardrails in LLMs are insufficient for use in an agentic setting and demand more robust controls and mitigation. Our results motivate and demonstrate mitigation strategies that move beyond model-level guardrails toward explicit, agent-level fairness regulation. Additionally, we also showed that targeted (agent) prompt augmentation can help reduce such misinterpretations and resulting biases. Agent bias in other domains and in multi-turn interactions can be taken up as future work.

References

Arpit Agarwal, Rad Niazadeh, and Prathamesh Patil. Misalignment, learning, and ranking: Harnessing users limited attention. *arXiv preprint arXiv:2402.14013*, 2024.

Salem Alelyani. Detection and evaluation of machine learning bias. *Applied Sciences*, 11 (14):6271, 2021.

Muhammad Ali. Measuring and mitigating bias and harm in personalized advertising. In *Proceedings of the 15th ACM Conference on Recommender Systems*, pages 869–872, 2021.

Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, et al. Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint arXiv:1810.01943*, 2018.

Marianne Bertrand and Sendhil Mullainathan. Are emily and greg more employable than lakisha and jamal? a field experiment on labor market discrimination. *American economic review*, 94(4):991–1013, 2004.

Saumya Bhadani. Biases in recommendation system. In *Proceedings of the 15th ACM conference on recommender systems*, pages 855–859, 2021.

Asia J Biega, Krishna P Gummadi, and Gerhard Weikum. Equity of attention: Amortizing individual fairness in rankings. In *The 41st international acm sigir conference on research & development in information retrieval*, pages 405–414, 2018.

Hongru Cai, Yongqi Li, Wenjie Wang, Fengbin Zhu, Xiaoyu Shen, Wenjie Li, and Tat-Seng Chua. Large language models empowered personalized web agents. In *Proceedings of the ACM on Web Conference 2025*, pages 198–215, 2025.

L Elisa Celis, Sayash Kapoor, Farnood Salehi, and Nisheeth K Vishnoi. An algorithmic framework to control bias in bandit-based personalization. *arXiv preprint arXiv:1802.08674*, 2018.

Jin Chen, Zheng Liu, Xu Huang, Chenwang Wu, Qi Liu, Gangwei Jiang, Yuanhao Pu, Yuxuan Lei, Xiaolong Chen, Xingmei Wang, et al. When large language models meet personalization: Perspectives of challenges and opportunities. *World Wide Web*, 27(4): 42, 2024.

Zhenpeng Chen, Jie M Zhang, Federica Sarro, and Mark Harman. A comprehensive empirical study of bias mitigation methods for machine learning classifiers. *ACM transactions on software engineering and methodology*, 32(4):1–30, 2023.

Zhendong Chu, Shen Wang, Jian Xie, Tinghui Zhu, Yibo Yan, Jinheng Ye, Aoxiao Zhong, Xuming Hu, Jing Liang, Philip S Yu, et al. Llm agents for education: Advances and applications. *arXiv preprint arXiv:2503.11733*, 2025.

Amit Datta, Michael Carl Tschantz, and Anupam Datta. Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination. *arXiv preprint arXiv:1408.6491*, 2014.

Anupam Datta, Matt Fredrikson, Gihyuk Ko, Piotr Mardziel, and Shayak Sen. Proxy non-discrimination in data-driven systems. *arXiv preprint arXiv:1707.08120*, 2017.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *FAccT*, pages 120–128, 2019.

Yashar Deldjoo and Tommaso Di Noia. Cfairllm: Consumer fairness evaluation in large-language model recommender system. *ACM Transactions on Intelligent Systems and Technology*, 2025.

Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Sam Stevens, Boshi Wang, Huan Sun, and Yu Su. Mind2web: Towards a generalist agent for the web. *Advances in Neural Information Processing Systems*, 36:28091–28114, 2023.

Christian Djeffal. Reflexive prompt engineering: A framework for responsible prompt engineering and ai interaction design. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, pages 1757–1768, 2025.

Yijiang Dong, Tiancheng Hu, and Nigel Collier. Can llm be a personalized judge? In *Findings of EMNLP 2024*, pages 10126–10141, 2024.

Tal Feldman and Ashley Peake. End-to-end bias mitigation: Removing gender bias in deep learning. *arXiv preprint arXiv:2104.02532*, 2021.

Jacques Ferber and Gerhard Weiss. *Multi-agent systems: an introduction to distributed artificial intelligence*, volume 1. Addison-wesley Reading, 1999.

Emilio Ferrara. Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. *Sci*, 6(1):3, 2024.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179, 2024.

Isabel O Gallegos, Ryan Aponte, Ryan A Rossi, Joe Barrow, Mehrab Tanjim, Tong Yu, Hanieh Deilamsalehy, Ruiyi Zhang, Sungchul Kim, Franck Dernoncourt, et al. Self-debiasing large language models: Zero-shot recognition and reduction of stereotypes. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 873–888, 2025.

Usman Gohar, Sumon Biswas, and Hridesh Rajan. Towards understanding fairness and its composition in ensemble machine learning. In *2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE)*, pages 1533–1545. IEEE, 2023.

Rubén González-Sendino, Emilio Serrano, Javier Bajo, and Paulo Novais. A review of bias and fairness in artificial intelligence. 2023.

Yufei Guo, Muzhe Guo, Juntao Su, Zhou Yang, Mengqiu Zhu, Hongfei Li, Mengyang Qiu, and Shuo Shuo Liu. Bias in large language models: Origin, evaluation, and mitigation. *arXiv preprint arXiv:2411.10915*, 2024.

Shenggang Hu, Jabir Alshehabi Al-Ani, Karen D Hughes, Nicole Denier, Alla Konnikov, Lei Ding, Jinhan Xie, Yang Hu, Monideepa Tarafdar, Bei Jiang, et al. Balancing gender bias in job advertisements with text-level bias mitigation. *Frontiers in big Data*, 5:805713, 2022.

Tianyi Huang and Arya Somasundaram. Mitigating bias in queer representation within large language models: A collaborative agent approach. *arXiv preprint arXiv:2411.07656*, 2024.

Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446, 2002.

Gabbrielle M Johnson. The hard proxy problem: proxies aren't intentional; they're intentional. *Philosophical Studies*, pages 1–29, 2025.

Matthew Joseph, Michael Kearns, Jamie H Morgenstern, and Aaron Roth. Fairness in learning: Classic and contextual bandits. *Advances in neural information processing systems*, 29, 2016.

Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. Meritocratic fairness for infinite and contextual bandits. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 158–163, 2018.

Mahammed Kamruzzaman and Gene Louis Kim. Prompting techniques for reducing social bias in llms through system 1 and system 2 cognitive processes. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing-Natural Language Processing in the Generative AI Era*, pages 511–520, 2025.

Michael Kearns, Aaron Roth, and Zhiwei Steven Wu. Meritocratic fairness for cross-population selection. In *International conference on machine learning*, pages 1828–1836. PMLR, 2017.

Aleksandra Klašnja-Milićević, Mirjana Ivanović, Boban Vesin, and Zoran Budimac. Enhancing e-learning systems with personalized recommendation based on collaborative tagging techniques. *Applied Intelligence*, 48(6):1519–1535, 2018.

A Klimashevskaia, D Jannach, M Elahi, and C Trattner. A survey on popularity bias in recommender systems (2023). *CoRR, abs/2308.01118*.

G Roshan Lal, Sahin Cem Geyik, and Krishnaram Kenthapadi. Fairness-aware online personalization. *arXiv preprint arXiv:2007.15270*, 2020.

Anja Lambrecht and Catherine Tucker. Algorithmic bias? an empirical study of apparent gender-based discrimination in the display of stem career ads. *Management science*, 65 (7):2966–2981, 2019.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, et al. Agentbench: Evaluating llms as agents. *arXiv preprint arXiv:2308.03688*, 2023.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54 (6):1–35, 2021.

Dena F Mujtaba and Nihar R Mahapatra. Fairness in ai-driven recruitment: Challenges, metrics, methods, and future directions. *arXiv preprint arXiv:2405.19699*, 2024.

Moin Nadeem, Anna Bethke, and Siva Reddy. Stereoset: Measuring stereotypical bias in pretrained language models. In *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)*, pages 5356–5371, 2021.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel Bowman. Crows-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 1953–1967, 2020.

Thi-Nhung Nguyen, Linhao Luo, Thuy-Trang Vu, and Dinh Phung. The social cost of intelligence: Emergence, propagation, and amplification of stereotypical bias in multi-agent systems. *arXiv preprint arXiv:2510.10943*, 2025.

Tiago Palma Pagano, Rafael Bessa Loureiro, Fernanda Vitória Nascimento Lisboa, Gustavo Oliveira Ramos Cruz, Rodrigo Matos Peixoto, Guilherme Aragão de Sousa Guimarães, Lucas Lisboa dos Santos, Maira Matos Araujo, Marco Cruz, Ewerton Lopes Silva de Oliveira, et al. Bias and unfairness in machine learning models: a systematic literature review. *arXiv preprint arXiv:2202.08176*, 2022.

Swetasudha Panda, Ari Kobren, Michael Wick, and Qinlan Shen. Don't just clean it, proxy clean it: Mitigating bias by proxy in pre-trained models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5073–5085, 2022.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. Bbq: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2086–2105, 2022.

Gourab K Patro, Lorenzo Porcaro, Laura Mitchell, Qiuyue Zhang, Meike Zehlike, and Nikhil Garg. Fair ranking: a critical review, challenges, and future directions. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, pages 1929–1942, 2022.

Gourab K Patro, Himanshi Agrawal, Himanshu Gharat, Supriya Panigrahi, Nim Sherpa, Vishal Vaddina, and Dagnachew Birru. Responsible ai for general-purpose systems: Overview, challenges, and a path forward. *arXiv preprint arXiv:2601.13122*, 2026.

Rajesh Ranjan, Shailja Gupta, and Surya Narayan Singh. A comprehensive survey of bias in llms: Current landscape and future directions. *arXiv preprint arXiv:2409.16430*, 2024.

Rajesh Ranjan, Shailja Gupta, and Surya Narayan Singh. Fairness in agentic ai: A unified framework for ethical and equitable multi-agent system. *arXiv preprint arXiv:2502.07254*, 2025.

Clara Rus, Jeffrey Luppes, Harrie Oosterhuis, and Gido H Schoenmacker. Closing the gender wage gap: Adversarial fairness in job recommendation. *arXiv preprint arXiv:2209.09592*, 2022.

Abel Salinas, Parth Shah, Yuzhong Huang, Robert McCormack, and Fred Morstatter. The unequal opportunities of large language models: Examining demographic biases in job recommendations by chatgpt and llama. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–15, 2023.

Ranjan Sapkota, Konstantinos I. Roumeliotis, and Manoj Karkee. Ai agents vs. agentic ai: A conceptual taxonomy, applications and challenges. *Information Fusion*, 126:103599, 2026. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2025.103599.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36:68539–68551, 2023.

Deepa Sharma, Neha Reddy, Priya Gupta, and Rohit Sharma. Enhancing customer experience personalization through ai: Leveraging collaborative filtering, neural networks, and natural language processing. *Journal of AI ML Research*, 11(7), 2022.

Karanbir Singh and William Ngu. Bias-aware agent: Enhancing fairness in ai-driven knowledge retrieval. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 1705–1712, 2025.

Alex Tamkin, Amanda Askell, Liane Lovitt, Esin Durmus, Nicholas Joseph, Shauna Kravec, Karina Nguyen, Jared Kaplan, and Deep Ganguli. Evaluating and mitigating discrimination in language model decisions. *arXiv preprint arXiv:2312.03689*, 2023.

Zhen Tan, Jun Yan, I-Hung Hsu, Rujun Han, Zifeng Wang, Long Le, Yiwen Song, Yanfei Chen, Hamid Palangi, George Lee, et al. In prospect and retrospect: Reflective memory management for long-term personalized dialogue agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8416–8439, 2025.

Yixin Wan, Xingrun Chen, and Kai-Wei Chang. Which cultural lens do models adopt? on cultural positioning bias and agentic mitigation in llms. *arXiv preprint arXiv:2509.21080*, 2025.

Weizhi Wang, Li Dong, Hao Cheng, Xiaodong Liu, Xifeng Yan, Jianfeng Gao, and Furu Wei. Augmenting language models with long-term memory. *NeurIPS*, 36:74530–74543, 2023.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in neural information processing systems*, 33:5776–5788, 2020.

Wenxuan Wang, Zizhan Ma, Zheng Wang, Chenghan Wu, Jiaming Ji, Wenting Chen, Xiang Li, and Yixuan Yuan. A survey of llm-based agents in medicine: How far are we from baymax? *arXiv preprint arXiv:2502.11211*, 2025.

Yining Wang, Liwei Wang, Yuanzhi Li, Di He, and Tie-Yan Liu. A theoretical analysis of ndcg type ranking measures. In *Conference on learning theory*, pages 25–54. PMLR, 2013.

Ze Wang, Zekun Wu, Xin Guan, Michael Thaler, Adriano S Koshiyama, Skylar Lu, Sachin Beepath, Ediz Ertekin Jr, and María Pérez-Ortiz. Jobfair: A framework for benchmarking gender hiring bias in large language models. In *EMNLP (Findings)*, 2024.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Kyra Wilson and Aylin Caliskan. Gender, race, and intersectional bias in resume screening via language model retrieval. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 1578–1590, 2024.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. The rise and potential of large language model based agents: A survey. *Science China Information Sciences*, 68(2):121101, 2025.

Zidi Xiong, Yuping Lin, Wenya Xie, Pengfei He, Jiliang Tang, Himabindu Lakkaraju, and Zhen Xiang. How memory management impacts llm agents: An empirical study of experience-following behavior. *arXiv preprint arXiv:2505.16067*, 2025.

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. Webshop: Towards scalable real-world web interaction with grounded language agents. *Advances in Neural Information Processing Systems*, 35:20744–20757, 2022.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243*, 2018.

Weizhi Zhang, Xinyang Zhang, Chenwei Zhang, Liangwei Yang, Jingbo Shang, Zhepei Wei, Henry Peng Zou, Zijie Huang, Zhengyang Wang, Yifan Gao, et al. Personaagent: When large language model agents meet personalization at test time. *arXiv e-prints*, pages arXiv–2506, 2025a.

Zeyu Zhang, Quanyu Dai, Xiaohe Bo, Chen Ma, Rui Li, Xu Chen, Jieming Zhu, Zhenhua Dong, and Ji-Rong Wen. A survey on the memory mechanism of large language model-based agents. *ACM Transactions on Information Systems*, 43(6):1–47, 2025b.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. Memorybank: Enhancing large language models with long-term memory. In *AAAI*, volume 38, pages 19724–19731, 2024.

Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, et al. Webarena: A realistic web environment for building autonomous agents. *arXiv preprint arXiv:2307.13854*, 2023.