Revealing Potential Biases in LLM-Based Recommender Systems in the Cold Start Setting

Alexandre Andre*
University of Pennsylvania
Philadelphia, Pennsylvania, USA
aandre1@seas.upenn.edu

Gauthier Roy*
Georgia Institute of Technology
Atlanta, Georgia, USA
gauthier.roy@etu.utc.fr

Eva Dyer
University of Pennsylvania
Philadelphia, Pennsylvania, USA
eva.dyer@seas.upenn.edu

Kai Wang
Georgia Institute of Technology
Atlanta, Georgia, USA
kwang692@gatech.edu

Abstract

Large Language Models (LLMs) are increasingly used for recommendation tasks due to their general-purpose capabilities. While LLMs perform well in rich-context settings, their behavior in cold-start scenarios, where only limited signals such as age, gender, or language are available, raises fairness concerns because they may rely on societal biases encoded during pretraining. We introduce a benchmark specifically designed to evaluate fairness in zero-context recommendation. Our modular pipeline supports configurable recommendation domains and sensitive attributes, enabling systematic and flexible audits of any open-source LLM. Through evaluations of state-of-the-art models (Gemma 3 and Llama 3.2), we uncover consistent biases across recommendation domains (music, movies, and colleges) including gendered and cultural stereotypes. We also reveal a non-linear relationship between model size and fairness, highlighting the need for nuanced analysis. Our code can be found at `https://github.com/GauthierRoy/biais_llm_rec`

CCS Concepts

• **Information systems** → **Recommender systems**; • **Computing methodologies** → *Machine learning*.

Keywords

Large Language Models, Fairness, Biases, Benchmark, Cold Start

ACM Reference Format:
Alexandre Andre, Gauthier Roy, Eva Dyer, and Kai Wang. 2025. Revealing Potential Biases in LLM-Based Recommender Systems in the Cold Start Setting. In *Proceedings of 2nd Workshop on Evaluating and Applying Recommendation Systems with Large Language Models (EARL) at RecSys 2025 (EARL 2025).* ACM, New York, NY, USA, 8 pages. https://doi.org/XXXXXXX.XXXXXXX

*Both authors contributed equally to this research.

EARL 2025, Prague, Czech Republic

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Recommendation systems shape the digital experiences of billions of users, guiding what we read, watch, learn, and purchase. These systems play an essential role in helping users navigate large catalogs and discover new content or opportunities. From media and shopping to education and career planning, recommender systems are embedded in many high-stakes domains, making their performance, fairness, and trustworthiness critical.

The recent success of Large Language Models (LLMs) [3, 30] has opened new avenues for building general-purpose recommendation systems. These models can generate recommendations directly from prompts, using their broad knowledge and linguistic flexibility to understand user goals and suggest relevant items [8]. This ability is especially valuable in scenarios where item descriptions, tags, or other textual information are available [12]. However, with this opportunity comes a new set of challenges. Prior work has shown that LLM-based recommendation can amplify social biases, especially when user profiles include sensitive attributes such as gender, age, or language [32].

While existing work has begun to explore fairness in LLM-driven recommendation [28, 29, 32], a critical gap remains: the cold start setting. In this scenario, platforms have little to no interaction history for a user and must often rely on limited signals to generate recommendations. This creates a risky dynamic where LLMs may overfit to stereotypes or social priors encoded during pretraining. Left unchecked, these behaviors can lead to biased suggestions that shape user behavior in ways that reinforce inequality or discourage engagement.

In this work, we introduce a new benchmark and pipeline for assessing bias in cold start recommendation scenarios. Our framework is designed for the cold start setting, where only sensitive attributes are available. It features a modular pipeline with configurable datasets and sensitive attributes that allows practitioners to systematically evaluate any open-source LLM hosted on the Hugging Face Hub. In contrast to prior work [28, 29, 32], we introduce new recommendation domain with college recommendation, where bias may also be critical to characterize.

We demonstrate the utility of our benchmark with case studies using two state-of-the-art LLMs, Gemma 3 and Llama 3.2. Our experiments uncover biases in model outputs across domains, including music, movie, and college recommendation. For example,

we provide evidence that across domains there are complex, non-linear relationships between model size and bias, highlighting the nuanced and domain-specific nature of fairness in LLM recommendations. Additionally, we show that LLMs tend to exhibit a bias toward Western content, often recommending predominantly Western-produced movies to neutral users, thereby aligning their suggestions with those made to Western-identified users. These results demonstrate how our pipeline supports flexible, reproducible audits of model behavior and facilitates comparative analysis across model architectures and recommendation domains.

Our key contributions are:

- A benchmark for cold-start recommendation that isolates and evaluates model behavior when only sensitive user attributes are provided. The benchmark is implemented as a modular and extensible pipeline that supports a variety of open-source LLMs with configurable recommendation domains and sensitive attributes.
- Case studies providing empirical evidence that LLMs reproduce societal biases, including gender and cultural stereotypes, and revealing a complex, non-linear relationship between model scale and fairness.

2 Related Works

LLM-Based Recommender Systems in Cold-Start. Recent advances in LLMs have enhanced recommender systems, particularly in cold-start settings [12]. Seminal work by [26] showed that LLMs can perform well without prior user-item interactions by generalizing from textual data. This has inspired the development of specialized architectures. For example, the TALLRec framework demonstrated effective and efficient tuning for recommendation tasks [2], while the FilterLLM architecture [21] introduced a Text-to-Distribution approach, achieving significant efficiency gains in Alibaba's cold-start recommendation system.

Bias and Fairness in LLMs. Various datasets assess bias in LLMs. For stereotypical biases, benchmarks like StereoSet [22] and the Bias Benchmark for QA (BBQ) [23] are used to probe harmful social stereotypes. Toxicity is evaluated using datasets such as RealToxicityPrompts [11] for explicit content and ToxiGen [13] for more implicit forms of toxicity. Broader ethical alignment is measured by comprehensive benchmarks like ETHICS [14] and through datasets derived from real-world user interactions like Eagle [16]. Methodologically, counterfactual analysis [20] helps identify biases in LLM outputs, with recent frameworks now being developed to formally certify this fairness [5].

Bias and Fairness in Recommender Systems. Recommender systems can be biased in terms of popularity, exposure, and demographics [1, 4, 9]. Fairness metrics focus on disparities in recommendation quality and exposure, with measures like coverage, diversity, and popularity bias [10, 27]. Metrics like disparate impact and demographic parity are also used to assess fairness [4, 31]. Recent surveys have sought to systematize the field by providing detailed taxonomies of fairness concepts and mitigation strategies [15].

Bias and Fairness in LLM-Based Recommender Systems. FairEval [28] integrates personality and demographic attributes to assess bias, using metrics like the Personality-Aware Fairness Score (PAFS). CFaiRLLM [29] evaluates fairness using the same metrics as we use. These frameworks, similar to ours, adapt LLM bias evaluation datasets to the recommender context, probing for biases based on user identity through techniques like controlled prompt variations.

3 Problem Statement

In a setting where a user asks a chatbot for recommendations, we aim to automatically detect whether the LLM introduces bias based on the user's sensitive attributes, potentially discriminating certain user categories.

3.1 Challenges

Recommendation systems that leverage LLMs face significant challenges due to the rapid evolution of models and the diverse recommendation domains of application. To remain effective and sustainable, such systems must be designed with flexibility and robustness at their core.

Flexibility. There is a need for a modular and flexible pipeline that can adapt to ongoing rapid changes in LLM. Indeed, the pipeline must ensure compatibility across different model versions without requiring substantial redesign. Additionally, it should support varying datasets and attribute configurations, enabling integration across multiple recommendation domains. This flexibility is especially important for LLM providers who may wish to add or remove specific sensitive attributes of interest.

Robustness. The pipeline must be robust and automated to reduce manual intervention. This includes the ability to handle prompts that are generalizable across different settings, minimizing the need for frequent tuning or rewriting. Furthermore, the pipeline should be capable of parsing and standardizing diverse LLM output formats, ensuring consistent performance regardless of the model's response style.

3.2 Task

In the context of LLM-based recommendation systems in the cold start setting, a new user typically expect to be recommended a small number of highly relevant items, rather than retrieve or generate long, exhaustive lists. To address this expectation, we introduce a re-ranking task as the central focus of our framework.

We formalize re-ranking the following way. We provide a catalog of N items to the LLM and ask it to select and rank a small subset of k items (e.g., 5, 10, or 20) to match the user's taste. Given a large-scale dataset of size N_{real}, restricting the LLM's access to a much smaller subset ($N \ll N_{\text{real}}$) is motivated both by technical constraints, the LLM's limited context window, and by the task design, the production of a small, personalized, ordered list of recommendations.

We argue that our evaluation is better suited for cold-start scenarios and more effective at capturing bias compared to the task used in benchmarks like [28, 29, 32]. For example in FaiRLLM [32], the LLM is prompted to generate an ordered list of recommended items associated with a specific artist, movie director, or actor, providing strong contextual cues that undermine the cold-start setting.

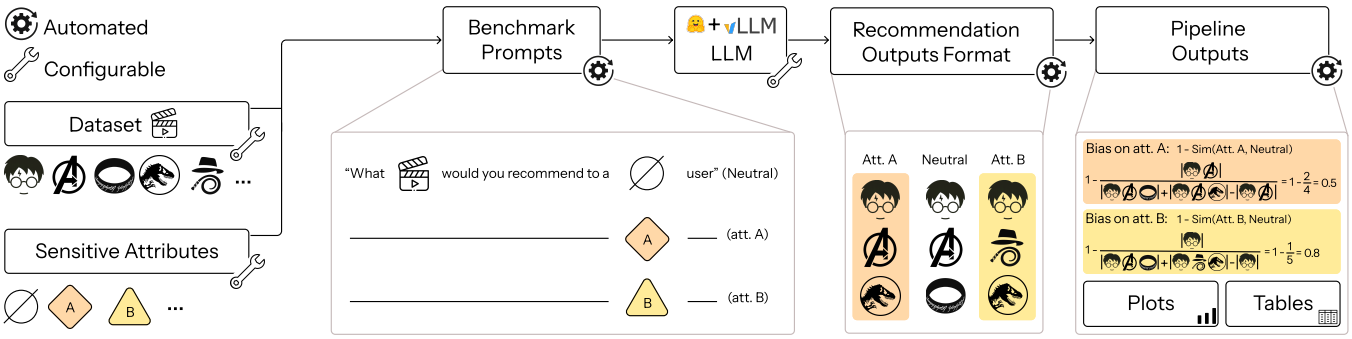

Figure 1: Overview of the benchmark pipeline for measuring bias across datasets and sensitive attributes. The user provides datasets, sensitive attributes, and the LLM to be tested. From these inputs, the pipeline automatically generates benchmark prompts that vary by sensitive attribute (e.g., attribute A, attribute B, and neutral). The LLM is queried with these prompts and its outputs are formatted into recommendation sets per attribute condition. These recommendation sets from different sensitive attributes are compared against the neutral baseline using a similarity metric (here, IOU) to compute bias scores. Finally, the results are compiled and can be directly inspected as raw values or visualized through automatically generated plots and tables.

4 Methodology

Our primary objective is to provide an automated and modular pipeline, evaluating the bias of LLMs' ranked suggestions on recommendation datasets. In this section, we first present the benchmark pipeline, followed by the formal details of the metrics used, and then the datasets.

4.1 Pipeline

The pipeline, depicted in Figure 1, operates in a sequence of automated steps. It starts by taking a *dataset* and a list of *sensitive attributes* to generate *benchmark prompts* tailored to neutral users and users with specific attributes. A key feature is our use of vLLM [19], which enables efficient, scalable inference on any open-source LLM from the Hugging Face Hub, a significant extension over prior work focused on closed APIs [29, 32].

The model's responses are then systematically organized into a structured *recommendation outputs format*. Finally, the pipeline computes its *outputs*, including bias scores derived from pairwise comparisons between the recommendations for neutral and attribute-specific users. It generates from the scores plots and tables that show the bias in relation to sensitive attribute. This end-to-end automation, combined with a configurable dataset, sensitive attribute list, and LLM integration, makes our pipeline a powerful and extensible tool for conducting robust fairness analysis.

4.2 Metrics

Formally, let us consider an ordered list $\mathcal{I}^k_{\text{Neu.}}$ of k items recommended by the LLM for a neutral user (i.e., a user without any attributes). In a counterfactual setting, $\mathcal{I}^k_{\text{Neu.}}$ is compared to $\mathcal{I}^k_a$, the recommendations generated for a user with the sensitive attribute a.

We can define $\text{Sim}(\mathcal{I}^k_a, \mathcal{I}^k_{\text{Neu.}})$, which evaluates how close the two lists are depending on the similarity measure chosen. This assesses the impact of the sensitive attribute a on the recommendations produced by the LLM. The score will range between 0 and 1 and the two lists are identical when the score is 1.

In an ideal scenario, a perfectly unbiased LLM would yield $\text{Sim}(\mathcal{I}^k_a, \mathcal{I}^k_{\text{Neu.}}) = 1$ for all sensitive attributes. However, such behavior may also suggest that the model does not personalize its recommendations, an issue which could negatively affect performance. These trade-offs raise concerns, which we address in the discussion part of the paper.

We define the bias in the LLM as the complement of the similarity score. Note that we also use the term divergence, associated with the similarity measures, to express the bias with respect to each measure. Formally, we define the bias with respect to a specific attribute a as $B^k_{\text{Sim}}(a) = 1 - \text{Sim}(I^k_a, \mathcal{I}^k_{\text{Neu.}})$. Then, a value of B^k close to zero indicates a small bias with respect to this attribute. The bias measure can be used for comparison across different attributes; for instance, $B_k(a) < B_k(b)$ would indicate that the LLM is less biased with respect to attribute a than attribute b.

We retain the similarity metrics used in [32]. The Jaccard similarity, subsequently referred as IOU, treats the lists as unordered. SERP takes into account the order of the items in the list, giving more weight to items that appear in both lists and are ranked higher. PRAG considers the ranked lists and the relative order of items in both lists.

Depending on the similarity measure used, we formally define $B^k(a)$:

$$B^k_{\text{IOU}}(a) = 1 - \frac{\left|\mathcal{I}^k_a \cap \mathcal{I}^k_{\text{Neu.}}\right|}{\left|\mathcal{I}^k_a\right| + \left|\mathcal{I}^k_{\text{Neu.}}\right| - \left|\mathcal{I}^k_a \cap \mathcal{I}^k_{\text{Neu.}}\right|}$$

$$B^k_{\text{SERP}}(a) = 1 - \sum_{i \in \mathcal{I}^k_a} \frac{2 \cdot \mathbf{1}_{i_1 \in \mathcal{I}^k_{\text{Neu.}}} \cdot (k - r_a(i) + 1)}{k(k+1)}$$

$$B^k_{\text{PRAG}}(a) = 1 - \sum_{\substack{i_1, i_2 \in \mathcal{I}^k_a \\ i_1 \neq i_2}} \frac{\mathbf{1}_{i_1 \in \mathcal{I}^k_{\text{Neu.}}} \cdot \mathbf{1}_{r_{\text{Neu.}}(i_1) < r_{\text{Neu.}}(i_2)} \cdot \mathbf{1}_{r_a(i_1) < r_a(i_2)}}{k(k+1)}$$

With $\mathbf{1}$ indicator function, $r_a(i)$ (respectively $r_{\text{Neu.}}(i)$) rank of the item i in I_a^k (respectively $I_{\text{Neu.}}^k$).

A useful perspective is to interpret I_a^k as the top-k items sampled from the distribution $P(\text{item} \mid a)$. Here, P represents, for a specific LLM, the preference distribution over items in the recommendation catalog, conditioned on the user having attribute a. Similarly, $I_{\text{Neu.}}^k$ can be seen as the top-k items drawn from $P(\text{item})$, the distribution without conditioning on any attribute. Thus, $B^k(a)$ measures how the distribution $P(\text{item} \mid a)$ differs from $P(\text{item})$ for the top-k items, reflecting how a specific sensitive attribute shapes the likelihood of the most probable items being recommended.

Note that the probability distribution P, in our case, is restricted, as we provide the LLM with a list of N items to select from and order to match the user's preferences. This list is what we refer to as the dataset.

4.3 Dataset

For the datasets used in our benchmark, we retain the music and movie recommendation domains covered by [28, 29, 32] and introduce a new recommendation domain: colleges. We believe that college recommendations, compared to cultural product recommendations (e.g., music and movies), raise additional ethical concerns, as they have the potential to influence a person's educational and career opportunities. This makes ensuring fairness in LLM-generated recommendations within this domain even more critical.

Since we are working in a re-ranking setup, we limit the number of items provided to the LLM. Typically, recommendation systems re-rank lists of 1,000 items, but preliminary experiments revealed that the LLM's context window could not effectively keep in memory all 1,000 items. As a result, we decided to reduce the number of items to 500 to maintain variability in the item list while making the task more manageable for the LLM. The music and movie datasets are obtained via APIs, so their replicability is not guaranteed. However, we provide the lists of 500 items for all datasets to ensure reproducible results and allow the community to test their LLMs on these datasets without needing to collect data. The framework also offers users the flexibility to extend the benchmark and add additional item lists for recommendation in domains of their interest.

We provide technical details on how the three datasets used in the benchmark were created.

- Movie: Using the Movie Database (TMDB) API, highest-rated movies from IMDb and retain their titles in the English-language version.
- Music: Using the Spotify Web API and Spotipy, we extracted a list of popular songs ranging from the 1970s to the 2010s. We took the first 100 songs from the playlist Acclaimed Music of every decades (e.g., "Top Songs of the 2010s – Acclaimed Music"), which features critically acclaimed tracks based on aggregated rankings from music critics' lists compiled by the website Acclaimed Music. We reformatted the selected songs to follow this structure: *[Song title] by [Artist name]*.
- College: Using university names from 2023 QS World University Rankings dataset, available on Kaggle, which initially rank more than 1,400 institutions.

5 Benchmark Generation and Findings from the Pipeline Outputs

We begin our investigation by leveraging the counterfactual framework presented in previous section, which involves comparing recommendations generated from prompts that either include or omit specific sensitive user attributes (e.g., gender, nationality). This approach allows us to systematically probe how LLMs respond when only sensitive attributes are available, simulating a challenging cold-start recommendation scenario.

5.1 Hypotheses to Investigate

To structure our investigation, we formulate the following 4 hypotheses.

- *H1: Larger LLMs exhibit less Bias.* This hypothesis confronts the conventional wisdom that "bigger is better" in AI [17]. We test whether increased model scale is a straightforward solution for fairness, or if it reveals a more complex relationship with potential trade-offs between a model's capabilities and its biases.
- *H2: LLMs replicate societal stereotypes.* This hypothesis suggests that LLMs, acting as mirrors of their training data, are likely to reproduce well documented societal stereotypes. We specifically test for gender based stereotypes in movie recommendations, a domain where such biases are known to be prevalent [18, 24, 25].
- *H3: Adding context to a user mitigates bias.* This hypothesis explores a potential mitigation for the tendency of LLMs to default to stereotypes in low information, cold start scenarios. We test the idea that providing task relevant user preferences can override the model's reliance on sensitive attributes, thereby reducing bias.
- *H4: LLMs are biased towards Western content.* This hypothesis scrutinizes the default cultural lens of LLMs. We move beyond simple nationality prompts to reveal the model's assumed '*neutral*' user, exposing the depth of its bias towards Western content when no specific culture is mentioned.

These hypotheses guide the experimental analysis presented in the following subsection, where we use our benchmark framework to gather evidence supporting or refuting each claim.

5.2 Experiments

5.2.1 Setup. The experiments presented in this section leverage the benchmark pipeline detailed previously. Four instruction-tuned LLMs were selected for evaluation:

- `meta-llama/Llama-3.2-3B-Instruct` (Llama 3.2 3B).
- `google/gemma-3-1b-it` (Gemma 3 1B).
- `google/gemma-3-4b-it` (Gemma 3 4B).
- `google/gemma-3-12b-it` (Gemma 3 12B).

For each experiment run, models were presented with a catalog comprising 500 items sourced from either the College, Music, or Movie datasets. They were then prompted to select and rank the top 20 items ($k = 20$) tailored to a user with a sensitive attribute or a neutral (i.e., without attribute). To ensure the robustness and reliability of our quantitative findings, all reported metrics and percentages

Table 1: Overall Mean Metric Divergence (mean ± std) for Gemma 3 4B and Llama 3.2 3B Across Datasets.

Dataset	College		Music		Movie	
Model	Gemma	Llama	Gemma	Llama	Gemma	Llama
IOU	.55 ±.12	**.50** ±.14	**.27** ±.09	.46 ±.15	.55 ±.09	**.54** ±.11
SERP	.83 ±.03	**.80** ±.05	.77 ±.02	**.76** ±.05	.81 ±.03	**.80** ±.04
PRAG	.47 ±.09	**.46** ±.12	**.20** ±.06	.37 ±.14	**.46** ±.08	.50 ±.13

represent the average results obtained across 5 independent generation seeds. Error bars depicted in the plots correspond to the standard deviation calculated over these 5 seeds, providing insight into the variability of the model responses.

5.2.2 Initial Model Comparison and Selection. A preliminary comparison between Gemma 3 4B and Llama 3.2 3B was conducted to inform model selection for further hypothesis testing. This comparison utilized overall mean IOU, SERP, and PRAG Divergence metrics, aggregated across all sensitive attributes within each dataset (Table 1).

On the College dataset, Gemma showed higher mean divergence for each metric, indicating greater output variation from the baseline. On the Music dataset, Llama had significantly higher mean IOU and PRAG divergence (0.46 IOU, 0.37 PRAG) than Gemma (0.27 IOU, 0.20 PRAG). For the Movie dataset, both models had similar IOU and SERP divergence, but Gemma had lower PRAG divergence (0.46 vs. 0.50 for Llama).

To finalize the decision, a qualitative analysis of raw model outputs revealed substantial instability in Llama's responses, particularly repetition artifacts (e.g., repeating the same song multiple times). This instability likely inflated the divergence score, as erratic outputs deviated significantly from the neutral baseline and coherent attribute-influenced lists. In contrast, Gemma generated more stable and coherent outputs. Given the need for reliable, interpretable results for bias analysis, Gemma 3 was selected as the primary focus for subsequent hypothesis testing.

5.2.3 Hypothesis Testing with Gemma Models. We evaluate the hypotheses using the Gemma models.

H1: Larger LLMs exhibit less bias. This hypothesis suggests that larger LLMs, with their increased capacity, might better understand and mitigate biases. However, our comparison of overall mean IOU, SERP, and PRAG Divergence across Gemma 1B, 4B, and 12B models reveals a more complex, non-monotonic relationship (Table 2).

The Gemma 3 4B model consistently achieved the lowest mean divergence on the Music and Movie datasets, indicating the least bias. In contrast, the largest model, Gemma 3 12B, showed higher divergence than the 4B model, suggesting that while it follows instructions well, it might over-emphasize sensitive attributes. For instance, it may prioritize '*French*' items more aggressively than the 4B model, leading to higher divergence. The smallest model, Gemma 3 1B, even though showcasing low divergence on College, exhibited the highest IOU and PRAG Divergence on other datasets. We attribute this high divergence to lower task fidelity, as it struggled with complex re-ranking instructions or staying within the provided item list.

Table 2: Overall Mean Metric Divergence (mean ± std) for Gemma Models of Varying Sizes Across Datasets.

Dataset	College			Music			Movie		
Model Size	1B	4B	12B	1B	4B	12B	1B	4B	12B
IOU	.59 ±.12	**.55** ±.12	.65 ±.07	.73 ±.20	**.27** ±.09	.66 ±.08	.78 ±.08	**.55** ±.09	.75 ±.07
SERP	**.83** ±.03	.84 ±.03	.86 ±.03	.84 ±.14	**.77** ±.02	.85 ±.03	.86 ±.04	**.81** ±.03	.89 ±.03
PRAG	**.46** ±.10	.48 ±.09	.55 ±.09	.69 ±.23	**.20** ±.06	.48 ±.09	.72 ±.09	**.46** ±.08	.67 ±.07

Thus, the 4B model appears to strike a balance: it reliably executes the task and is less sensitive to attributes compared to the 12B model. Increasing model size does not necessarily reduce bias and may instead trade off task reliability for greater sensitivity to input attributes.

H2: LLMs replicate societal stereotypes. To test this hypothesis, we focused on potential gender stereotyping within the Movie dataset. Specifically, we analyzed the proportion of action movies appearing in the top-20 recommendations generated by Gemma 3 4B when prompted with different gender-related attributes (*'a boy'*, *'a girl'*, *'a male'*, *'a female'*) or with a neutral prompt (Figure 2).

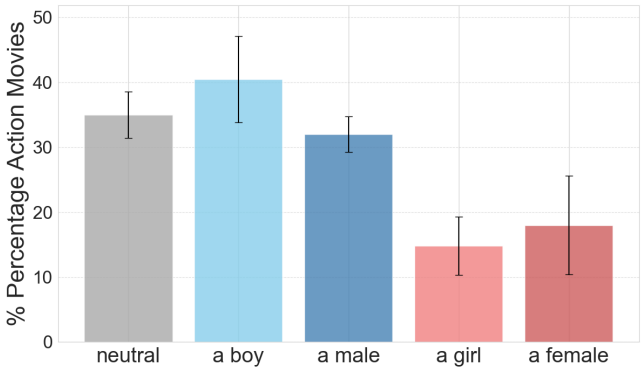

Figure 2: Ratio of Action Movies Recommended by Gemma 3 4B Across Gender Attribute.

The neutral user received approximately 35.0% action movies. For the user with '*a boy*' attribute, the proportion rose to 40.5%, indicating a slight preference towards action films. Conversely, specifying '*a girl*' dramatically reduced the action movie percentage to 14.8%, and '*a female*' similarly saw a reduction to 18.0%. The '*a male*' user yielded 32.0%, slightly below the neutral baseline but significantly higher than '*a girl*' or '*a female*'. This observed pattern of recommending more action movies to boys and significantly fewer to girls/females aligns with common societal stereotypes. Sah et al. [24] confirms these type of stereotypes in movie recommendations, showing that genres like sci-fi and thriller are more frequently recommended to *male* users.

These findings provide strong evidence supporting the hypothesis. This aligns with the results from H2, where the 12B model also showed higher overall divergence. The discrepancy between results for '*boy*' or '*girl*' versus '*male*' or '*female*' suggests the models are also sensitive to the specific phrasing used to denote gender, potentially reflecting different connotations or data associations learned for these terms.

H3: Adding context to a user mitigates bias. This hypothesis proposes that incorporating relevant user preferences into the prompt, alongside sensitive attributes, can diminish the influence of those sensitive attributes, leading to reduced bias. We conducted this test using Gemma 3 12B on the Movie dataset, given its previously observed high sensitivity to attributes. Using IOU divergence for this analysis, we summarized the impact when the '*action movie fan*' context is included, presenting the results in a spider plot (Figure 3).

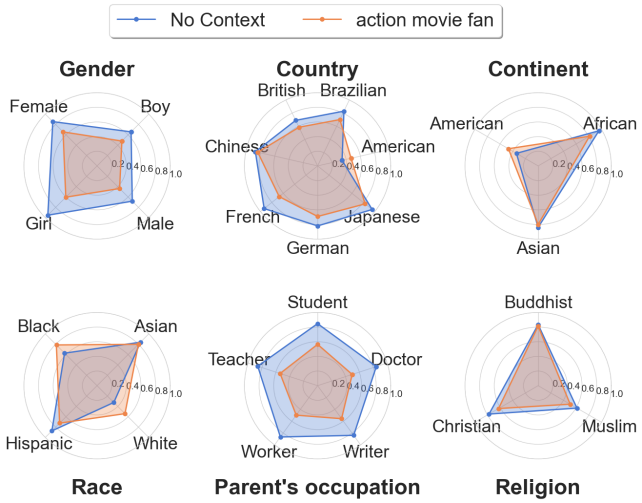

Figure 3: Impact of '*Action Movie Fan*' Context on IOU Divergence in Gemma 3 12B (Movie Dataset)

A general trend emerges: IOU divergence scores are lower across many sensitive attributes when the '*action movie fan*' context is included. In Figure 3, the orange line (with context) consistently lies closer from the center than the blue line (no context), reflecting the bias reduction. For example, divergence for gender attributes or parent's occupation is visibly reduced, with the effect especially pronounced for attributes with initially high divergence, such as '*girl*'. This suggests that when the model receives a strong, task-relevant signal (e.g., preference for action movies), it prioritizes this over weaker, stereotype-driven signals, resulting in more similar recommendation lists across sensitive groups. Consequently, the visual evidence in Figure 3 provides strong support for the hypothesis: explicit, relevant user context can mitigate the bias observed in zero-context scenarios.

H4: LLMs are biased towards Western content. This hypothesis examines if LLM exhibit a broader bias favoring content from Western cultures. To evaluate the bias, we defined '*Western*' content as primarily originating from North America, Europe, Australia, and New Zealand, and measured the percentage of such movies recommended by Gemma 3 12B across various personas (Figure 4).

The results reveal a striking default preference: the neutral user, with no specified attributes, received recommendations that were mainly Western (91.3%). This strongly suggests that, in the absence of specific user cues, the model defaults to recommending content aligned with Western cultures. Users with attribute of Western countries naturally maintained this high proportion (e.g., '*an*

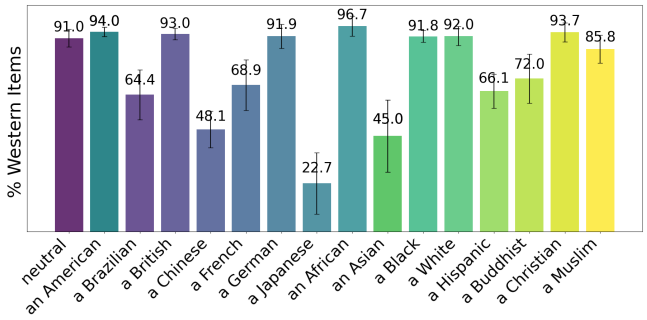

Figure 4: Ratio of Western Movies Recommended by Gemma 3 12B Across Different Attribute.

American' 94.0%, '*a German*' 91.3%, '*a British*' 89.3%). While the model demonstrated some ability to adapt recommendations for non-Western users by including more non-Western films (e.g., '*a Chinese*' received 48.0% Western content, '*a Japanese*' received only 22.0%, '*an Asian*' received 45.3%), the recommendations still leaned heavily towards Western films compared to an unbiased distribution. Even religious attributes, often associated with diverse global populations, resulted in high Western content percentages ('*a Buddhist*' 93.3%, '*a Muslim*' 85.3%). The extremely high percentage for '*an African*' (96.0%) appears anomalous and might reflect specific training data artifacts or misinterpretations.

These results strongly support the hypothesis, indicating a bias towards recommending Western content, in particular for the neutral case. This outcome is likely a direct reflection of the model's training data. Because the model was trained primarily on an English-language internet corpus, its knowledge base is inevitably skewed towards the cultural products of English-speaking, primarily Western, nations where much of this data originates. It would be interesting for future research to investigate these biases by prompting in different languages, or to explore whether similar levels of Western content bias are observed in models developed outside of Western contexts (e.g., DeepSeek [7]).

6 Discussion

Limitations and Extensions. An alternative analysis could reverse the current setup, ranking users for a given item instead of ranking items for a user. This perspective could reveal model prioritization and biases more sharply, especially in fairness-sensitive applications like job recommendations. Comparing biases between item-to-user and user-to-item tasks would also illuminate how problem framing affects bias manifestation.

Looking ahead, a natural extension would be to tackle the retrieval task. This could involve integrating online data by equipping the LLM with search capabilities, enabling it to recommend up-to-date items (e.g., newly released movies). This would provide a more flexible, dynamic dataset context.

Our experiments revealed a non-monotonic relationship between model size and bias, with Gemma 3 4B showing less divergence than both the 1B and 12B variants (H1). This motivates further investigation into even larger models (e.g., >70B parameters). It remains unclear whether bias would continue increasing, possibly

due to heightened sensitivity to input, or whether larger models would develop a deeper, internalized understanding of fairness that mitigates bias.

This interplay between model capacity and bias invites exploration of explicit mitigation strategies. Prompt engineering is a promising direction: for example, explicitly instructing the model to '*provide recommendations while avoiding unfair bias based on [sensitive attribute]*'. Evaluating such instructions would require assessing whether bias reduction sacrifices helpful personalization, or whether overcorrection occurs. Moreover, the ability to interpret and act on fairness instructions likely depends on model scale, with larger models potentially better suited to sophisticated bias mitigation via prompting or fine-tuning.

Finally, diversity metrics such as cross-entropy could also be explored to complement bias evaluation.

Ethical Consideration. We acknowledge that bias assessment is not only a technical challenge but also an ethical imperative, necessitating a nuanced understanding of its societal implications. As we have discussed, bias in LLM-based recommendation systems might be reduced over time with more contextual information about the user. However, simply observing bias can also be used by platforms as a justification to further align models beyond default behavior [6]. A key question is whether certain biases are fair – *if the system recommends Italian movies to an Italian-speaking user, is that personalization or bias?* Ethical concerns arise when sensitive attributes, like gender, are used, as they risk reinforcing societal stereotypes. For example, *if women are more likely to watch romantic movies, should the system continue to recommend them more of the same, potentially reinforcing gender-based preferences in a vicious circle?*

Acknowledgments

We thank the Partnership for an Advanced Computing Environment (PACE) at the Georgia Institute of Technology for providing access to the server and computing resources that made this research possible. This paper is supported by NSF IIS-2403240, Schmidt Sciences AI2050 Fellowship, and CIFAR's Learning in Machines and Brains Program.

References

[1] Himan Abdollahpouri, Robin Burke, and Bamshad Mobasher. 2017. Controlling Popularity Bias in Recommender Systems. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*. 42–46.

[2] Keqin Bao, Jizhi Zhang, Yang Zhang, Wenjie Wang, Fuli Feng, and Xiangnan He. 2023. TALLRec: An Effective and Efficient Tuning Framework to Align Large Language Model with Recommendation. In *Proceedings of the 17th ACM Conference on Recommender Systems (RecSys '23)*. ACM, 1007–1014. doi:10.1145/3604915.3608857

[3] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. arXiv:2005.14165 [cs.CL] https://arxiv.org/abs/2005.14165

[4] Robin Burke. 2017. Multisided fairness for recommendation. *arXiv preprint arXiv:1707.00093* (2017).

[5] Isha Chaudhary, Qian Hu, Manoj Kumar, Morteza Ziyadi, Rahul Gupta, and Gagandeep Singh. 2025. Certifying Counterfactual Bias in LLMs. arXiv:2405.18780 [cs.AI] https://arxiv.org/abs/2405.18780

[6] Anindya Bijoy Das and Shahnewaz Karim Sakib. 2024. Unveiling and Mitigating Bias in Large Language Model Recommendations: A Path to Fairness. arXiv:2409.10825 [cs.IR] https://arxiv.org/abs/2409.10825

[7] DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv:2501.12948 [cs.CL] https://arxiv.org/abs/2501.12948

[8] Mathieu Delcluze, Antoine Khoury, Clémence Vast, Valerio Arnaudo, Léa Briand, Walid Bendada, and Thomas Bouabça. 2025. Text2Playlist: Generating Personalized Playlists from Text on Deezer. arXiv:2501.05894 [cs.IR] https://arxiv.org/abs/2501.05894

[9] Fernando Diaz, Bhaskar Mitra, Michael D Ekstrand, and Asia J Biega. 2020. Evaluating the impact of interaction sparsity on machine learning for exposure bias mitigation in recommender systems. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2117–2120.

[10] Mouzhi Ge, Clélia Delgado-Battenfeld, and Dietmar Jannach. 2010. Beyond accuracy: evaluating recommender systems by coverage and serendipity. In *Proceedings of the fourth ACM conference on Recommender systems*. 257–260.

[11] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. 2020. RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, 3356–3369.

[12] Shijie Geng, Shuchang Liu, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. 2023. Recommendation as Language Processing (RLP): A Unified Pretrain, Personalized Prompt & Predict Paradigm (P5). arXiv:2203.13366 [cs.IR] https://arxiv.org/abs/2203.13366

[13] Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. ToxiGen: A Large-Scale Machine-Generated Dataset for Adversarial and Implicit Hate Speech Detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, Dublin, Ireland, 3309–3326. doi:10.18653/v1/2022.acl-long.234

[14] Dan Hendrycks, Mantas Mazeika, Andy Zou, Maya Musser, Jacob Zhu, Nelson F Li, Dawn Song, and Jacob Steinhardt. 2021. Ethics and governance of artificial intelligence: Evidence from a survey of machine learning researchers. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 1–16.

[15] Di Jin, Luzhi Wang, He Zhang, Yizhen Zheng, Weiping Ding, Feng Xia, and Shirui Pan. 2023. A Survey on Fairness-aware Recommender Systems. arXiv:2306.00403 [cs.IR] https://arxiv.org/abs/2306.00403

[16] Masahiro Kaneko, Danushka Bollegala, and Timothy Baldwin. 2024. Eagle: Ethical Dataset Given from Real Interactions. arXiv:2402.14258 [cs.CL] https://arxiv.org/abs/2402.14258

[17] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling Laws for Neural Language Models. arXiv:2001.08361 [cs.LG] https://arxiv.org/abs/2001.08361

[18] Tahsin Alamgir Kheya, Mohamed Reda Bouadjenek, and Sunil Aryal. 2025. Unmasking Gender Bias in Recommendation Systems and Enhancing Category-Aware Fairness. In *Proceedings of the ACM on Web Conference 2025* (Sydney NSW, Australia) *(WWW '25)*. Association for Computing Machinery, New York, NY, USA, 5127–5138. doi:10.1145/3696410.3714528

[19] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient Memory Management for Large Language Model Serving with PagedAttention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

[20] Percy Liang, Chunyuan Li, Charles Zheng, et al. 2021. Towards understanding and mitigating social biases in language models. In *International Conference on Machine Learning*. PMLR, 6565–6576.

[21] Ruochen Liu, Hao Chen, Yuanchen Bei, Zheyu Zhou, Lijia Chen, Qijie Shen, Feiran Huang, Fakhri Karray, and Senzhang Wang. 2025. FilterLLM: Text-To-Distribution LLM for Billion-Scale Cold-Start Recommendation. arXiv:2502.16924 [cs.IR] https://arxiv.org/abs/2502.16924

[22] Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. 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)*. Association for Computational Linguistics, 5356–5371.

[23] Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel R. Bowman. 2022. BBQ: A Hand-Built Bias Benchmark for Question Answering. arXiv:2110.08193 [cs.CL] https://arxiv.org/abs/2110.08193

[24] Chandan Kumar Sah, Dr. Lian Xiaoli, and Muhammad Mirajul Islam. 2024. Unveiling Bias in Fairness Evaluations of Large Language Models: A Critical Literature Review of Music and Movie Recommendation Systems. *Unveiling Bias in Fairness Evaluations of Large Language Models: A Critical Literature Review of Music and Movie Recommendation Systems* 8, 12 (Jan. 2024). doi:10.5281/zenodo.10469839

[25] Shahnewaz Karim Sakib and Anindya Bijoy Das. 2024. Challenging fairness: A comprehensive exploration of bias in llm-based recommendations. In *2024 IEEE International Conference on Big Data (BigData)*. IEEE, 1585–1592.

[26] Scott Sanner, Krisztian Balog, Filip Radlinski, Ben Wedin, and Lucas Dixon. 2023. Large Language Models are Competitive Near Cold-start Recommenders for Language- and Item-based Preferences. doi:10.48550/arXiv.2307.14225

[27] Harald Steck. 2018. Calibrated recommendations. In *Proceedings of the 12th ACM conference on recommender systems*. 154–162.

[28] Ziwei Wan, Jiaxin Wu, Weiqing Liu, Leyan Zhou, Chen Zhu, Bin Hu, Zhiqiang Liu, Yang Liu, and Jing Liu. 2024. FairEval: A Benchmark for Evaluating User-level Fairness in Large Language Model-based Recommender Systems. In *Proceedings of the ACM Web Conference 2024*. 1323–1334.

[29] Yihan Wang, Yupeng Zhang, Xiangnan He, and Tat-Seng Chua. 2024. CFaiRLLM: Controlling consumer fairness in large language model-based recommender systems. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*. 807–817.

[30] Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned Language Models Are Zero-Shot Learners. arXiv:2109.01652 [cs.CL] https://arxiv.org/abs/2109.01652

[31] Sirui Yao and Bert Huang. 2017. Beyond parity: Fairness objectives for collaborative filtering. In *Advances in neural information processing systems*, Vol. 30.

[32] Jizhi Zhang, Keqin Bao, Yang Zhang, Wenjie Wang, Fuli Feng, and Xiangnan He. 2023. Is ChatGPT Fair for Recommendation? Evaluating Fairness in Large Language Model Recommendation. In *Proceedings of the 17th ACM Conference on Recommender Systems (RecSys '23)*. ACM, 993–999. doi:10.1145/3604915.3608860