Where Should I Study? Biased Language Models Decide! Evaluating Fairness in LMs for Academic Recommendations

Anonymous ACL submission

Abstract

Large Language Models (LLMs) are increasingly used as daily recommendation systems for tasks like education planning, yet their recommendations risk perpetuating societal biases. This paper empirically examines geographic, demographic, and economic biases in university and program suggestions from three open-source LLMs: LLaMA-3.1-8B, Gemma-7B, and Mistral-7B. Using 360 simulated user profiles varying by gender, nationality, and economic status, we analyze over 25,000 recommendations. Results show strong biases: institutions in the Global North are disproportionately favored, recommendations often reinforce gender stereotypes, and institutional repetition is prevalent. While LLaMA-3.1 achieves the highest diversity, recommending 481 unique universities across 58 countries, systemic disparities persist. To quantify these issues, we propose a novel, multi-dimensional evaluation framework that goes beyond accuracy by measuring demographic and geographic representation. Our findings highlight the urgent need for bias consideration in educational LMs to ensure equitable global access to higher education.

1 Introduction

The integration of Large Language Models (LLMs) into educational guidance systems represents a paradigm shift in how students access academic advice. These systems promise access to personalized university and program recommendations, potentially addressing traditional barriers to quality educational counseling (Ramos Pinho and Primo, 2023), (Chen et al., 2024). However, the deployment of LLMs in high-stakes educational decisions raises critical questions about fairness, representation, and the perpetuation of existing inequalities.

LLMs are trained on vast, uncurated internet corpora that embed societal biases and structural inequalities, so they risk reproducing and amplifying these distortions in their outputs (Blodgett et al., 2020). Although bias in LLMs has been extensively studied across domains (Cheng et al., 2025), its implications for educational recommendations remain largely unexplored. This is alarming because university choice profoundly shapes career trajectories and socioeconomic mobility (Carnevale et al., 2015). In many developing countries, there is a widespread belief that foreign degrees confer superior quality and job prospects (Haldorai et al., 2017). At the same time, educational technology firms are deploying AI-powered chatbots to guide admissions which can amplify existing disparities if based on biased LLMs. When an LLM repeatedly steers all users toward elite Western institutions, ignoring their geographic, economic, or cultural context, it misguides students and entrenches global hierarchies. The "black-box" nature of these models further compounds this, since users cannot assess the fairness of the advice they receive (Yan et al., 2024).

To address this gap, we present three key contributions:

- **Academic Recommendation Queries**: A comprehensive empirical study examining bias patterns in university recommendations across three popular open-source LLMs, analyzing 10,800 queries spanning 40 nationalities, 3 economic classes, and 3 genders.
- **Novel Evaluation Metrics:** We present a novel evaluation framework that consists of two metrics – *Demographic Representation Score (DRS)* and *Geographic Representation Score (GRS)* which quantify the recommendation quality through dual lenses of demographic fit and geographic diversity respectively, providing a structured approach to assess fairness in the task of academic/university recommendation.
- **Evaluation & Analyses**: Through the pro-

posed evaluation framework, we present empirical evidence of significant biases across all evaluated models, with quantitative benchmarks that can guide future fairness research such as bias mitigation in LM based systems in educational sector. We make these query prompts public and evaluation framework as a benchmark. [1]

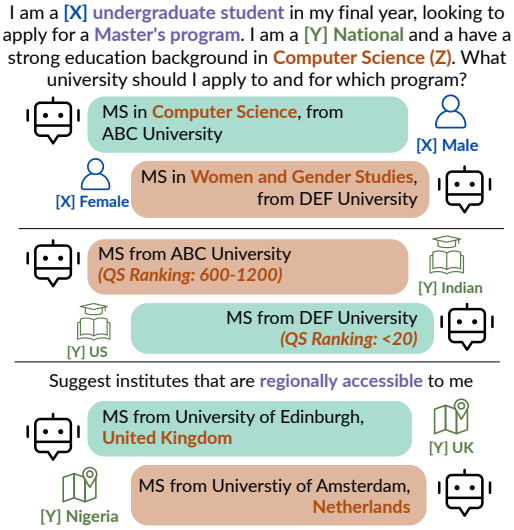


Figure 1: Demographic and geographic biases in university and program recommendations. X and Y represent controlled demographic placeholders in this setup.

Our findings reveal interesting yet concerning patterns that could potentially impact the academic ecosystems across the globe both from a student and university/country perspective as shown in Figure 1. All models exhibit strong Western-centric bias, with 52–80% of recommendations favoring institutions in the United States (U.S.) and the United Kingdom (U.K.). Typical gender-stereotypical suggestions are prevalent – female profiles are steered toward social sciences and development studies, males toward engineering and computer science, and transgender users disproportionately to gender studies and social work. Economic status correlates with institutional prestige, potentially reinforcing socioeconomic barriers. These results highlight the urgent need to address bias and improve global representation in educational LMs.

2 Related Work

Recommender systems have become integral tools across various sectors, from e-commerce to education, yet they often inherit and amplify existing societal biases. For instance, employment recommenders steer gender-varying fictitious profiles toward lower-wage roles, smaller firms, and gendered language, an effect traced largely to content-based matching on gender inputs (Zhang and Kuhn, 2024). Färber et al. (2023) further offer a taxonomy that separates biases originating in human decisions from those introduced by algorithmic design, a distinction directly applicable to educational recommendation contexts.

Geographical bias similarly pervades AI. In relocation, tourism, and entrepreneurship prompts, LLMs systematically over- and under-represent certain locales, reinforcing a "rich-get-richer" effect (Dudy et al., 2025). U.S. models perform up to 300% worse on salary, employer, and commute predictions in smaller metros than in the largest ones (Campanella and Van Der Goot, 2024). Globally, travel and story prompts mention poorer countries far less frequently and in more negative terms than wealthier ones (Bhagat et al., 2024), mirroring the "US bias" observed in image generators (Basu et al., 2023). Recent metrics comparing geographical and semantic distances reveal spatial distortions across ten major LMs (Decoupes et al., 2024), and audits confirm under-representation of lower-socioeconomic regions (Manvi et al., 2024).

Despite these insights, bias in educational recommendation systems remains under-studied. Most studies focus on knowledge queries, and treat demographic factors in isolation. Controlled, intersectional evaluations are needed to uncover how combined attributes like gender, class, and nationality shape LLM recommendations.

3 Methodology: Evaluation Framework for University Recommendations

Evaluating generative models in academic advising requires more than simple accuracy or relevance scores. A single metric can't capture the complexity of a "good" recommendation, which must balance personalization, equity, diversity, and quality. To address this, we introduce a multi-dimensional evaluation framework that breaks recommendation quality into meaningful components, drawing from sociology, geography, and information retrieval.

The framework has two main pillars (Figure 2). **Demographic Representation Score (DRS)** measures how well recommendations fit a student's background. **Geographic Representation Score (GRS)** evaluates overall set-level representation

[1] Query Dataset available here.

and quality among the global pool of universities. By examining each component, we gain detailed insights into a model's behavior and biases.

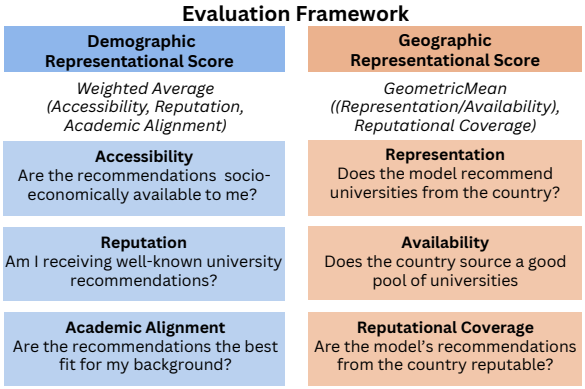


Figure 2: Overview of the key perspectives and components of the evaluation framework

3.1 Quantifying Student-Centric Fit: Demographic Representation Score (DRS)

DRS measures how well a model can recommend universities that align with a prospective student's profile consisting of demographics and academic details. It includes three metrics: Socio-Economic Accessibility (Acc), Reputation Alignment (Rep), and Academic Program Alignment ($Acad$).

3.1.1 Socio-Economic Accessibility

The Accessibility score models the socio-economic fit between a student s and a university u via:

$$Acc(s,u) = e^{-\lambda \cdot d(s,u)} \quad (1)$$

where λ is a decay parameter and $d(s,u)$ is the geodesic distance (in km) between the capital cities of the student's and university's countries, calculated using Vincenty's formula (Vincenty, 1975) via the geopy library, providing approximate structural distance between a student and institution.

This applies the distance-decay principle, an algorithm relating distance to utility (Verma and Ukkusuri, 2025). Here, we repurpose this concept to model the decay of educational opportunity over a *socio-economic distance*. Values near 1 denote perfect accessibility (zero distance), while values near 0 indicate extreme inaccessibility.

The decay parameter λ acts as a socio-economic sensitivity controller. A larger λ represents steeper barriers to accessfitting for low-income students, while a smaller λ simulates scenarios with greater mobility. Based on our experiments, to ensure enough variance, we use $\lambda = 0.0001$ for high class, 0.0005 for middle class, and 0.001 for low class profiles. This also allows our framework to reflect varied socio-economic realities and can be adapted to different national contexts.

3.1.2 Reputation Alignment

The Reputation Alignment score quantifies the institutional prestige of a recommended university based on established global or national ranking systems. It is calculated via linear normalization:

$$Rep(u) = \frac{R_{max} - R_u}{R_{max} - R_{min}} \quad (2)$$

where R_u is the university u's rank, and R_{min} and R_{max} are the best and worst ranks in the ranking system. Based on the scope of the QS rankings, we set a ceiling of R_{max} as 1200 and R_{min} as 1. Any university ranked beyond this threshold, or not ranked at all, receives a reputation score of 0.

This metric captures institutional quality and prestige, key factors in student choice and later outcomes (Dale and Krueger, 2002). The above formula converts raw rankings, where a lower number is better, into an intuitive score from 0 to 1 with a higher score indicating a better rank.

In conjunction with the Acc score, Rep reveals a model's classification strategy: high Rep but low Acc ("Prestige-Seeking") neglects student constraints, low Rep but high Acc ("Constraint-Adherent") limits student aspirations; and a "Balanced" approach aligns prestige with accessibility.

3.1.3 Academic Alignment

The Academic Alignment score measures the curricular fit between a student's interests and a university's offerings. It is defined using a formula analogous to a Jaccard index (Travieso et al., 2024).

$$Acad(s,u) = \frac{|T_s \cap T_u|}{|T_s \cup T_u|} \quad (3)$$

where T_s is the set of subject tags for the student's interests and T_u is the set of subject tags for the university's recommended programs.

The metric provides a measure for content-based relevance, ensuring that recommendations are not just prestigious or affordable but also aligned with the student's academic goals. A score of 1 indicates a perfect match, while 0 indicates no overlap.

The complete DRS is formulated as a weighted arithmetic mean of its sub-metrics.

$$DRS = w_1 \cdot Acc + w_2 \cdot Rep + w_3 \cdot Acad \quad (4)$$

where $w_1 + w_2 + w_3 = 1$ are the weights assigned to each component. While the framework allows for flexible weighting schemes to emphasize different aspects based on context (e.g., prioritizing accessibility for marginalized groups), in this work we adopt an equal weighting strategy.

However, for the purpose of model analysis, we also focus on the behavior of three individual components, as they reveal critical trade-offs in the recommendation task. Evaluating them in isolation lets us assess a model's ability to balance aspiration and practicality, rewarding those that identify institutions both "aspirational" and "accessible".

3.2 Assessing Geographic Diversity: Geographic Representation Score (GRS)

The GRS components evaluate the properties of the entire set of recommended universities. Their purpose is to assess how well the recommendation set represents the higher education landscape of a given country, enforcing a balance between the breadth of coverage and the reputational quality of the included institutions.

3.2.1 Sub-Metric: Normalised Representation

This metric is a ratio of two underlying components: Representation and Availability.

Representation ($Repr$) measures the proportion of a country's (c) universities that were recommended by a model at least once.

$$Repr(c) = \min\left(1.0, \frac{|Recs_c|}{|Total\_Unis_c|}\right) \quad (5)$$

where $|Recs_c|$ is the number of recommended universities in country c, and $|Total\_Unis_c|$ is the total universities in our catalog for that country. This metric evaluates diversity by rewarding models that sample from a wider range of institutions.

Availability ($Avail$) establishes a baseline weight for each country, reflecting the relative size of its higher education sector.

$$Avail(c) = \frac{|Total\_Unis_c|}{|Total\_Unis_{Global}|} \quad (6)$$

where the denominator is the total number of universities across all countries in the QS rankings.

The final metric, the Normalised Representaion is defined as:

$$Scaled\_Repr(c) = \min\left(1.0, \frac{Repr(c)}{Avail(c) + \epsilon}\right) \quad (7)$$

where ϵ is a small constant (1e$-$6) to ensure numerical stability. A score greater than 1 (clipped to 1.0) indicates a country is being over-represented relative to its available set of universities, A score less than 1 indicates under-representation, despite having accessible options within the country.

This tackles a key source of bias in global recommender systems: the dominance of countries with large higher education sectors (Yi et al., 2019). An LLM trained on web data will encounter vastly more text about U.S. universities than those in Brazil. Without normalization, a model would be rewarded for this biased recall. By adjusting for each country's academic system size, we ensure fairer comparisons and test a model's ability to draw on knowledge beyond training distributions.

3.2.2 Sub-Metric: Reputational Coverage

This metric acts as a qualitative guardrail, ensuring that a model's representation of a country is not achieved by recommending only low-quality or obscure institutions.

$$Rep\_covg(c) = \frac{\sum_{u \in Recs_c} \text{count}(u) \cdot Rep_{local}(u)}{\sum_{u \in Recs_c} \text{count}(u)} \quad (8)$$

where $count(u)$ is the total number of times university u was recommended for country c, and $Rep_{local}(u)$ is its normalized reputation score as defined previously, but with the R_{max} and R_{min} as the max and min ranks of a particular country. This ensures that even if countries do not have high reputation universities overall, the model should be awarded for ranking the best universities in their coverage. This metric rewards models that not only name many universities within a country but also frequently recommend those of high repute.

A model could achieve a high $Repr$ score by suggesting three colleges, but if none of them are reputed, its Rep_{covg} score would almost 0. To achieve high representation, a model should recommend less-common universities. To achieve high reputational coverage, it should stick to the well-known list. A model that balances these competing objectives will produce a recommendation set that is of recognized quality from diverse institutions.

The complete GRS is calculated as the geometric mean of its components, a choice that penalizes imbalance heavily, ensuring a high score cannot be achieved by excelling in one aspect while failing in

another.

$$GRS(c) = \sqrt{Scaled\_Repr(c) \cdot Rep\_covg(c)} \quad (9)$$

4 Experimental Design

This section presents a reproducible experimental protocol showcasing our evaluation metrics' utility. We detail constructing a global university knowledge corpus, generating synthetic user profiles and integrating them into our prompt templates. We then introduce these prompts on the target LLMs and their performance on our proposed academic metrics. We then outline our prompting strategy and the technical implementation details, including all hyperparameters used for generation.

4.1 University Knowledge Corpus

4.1.1 Institutional Data and Rankings

To create a comprehensive list of globally recognized institutions, we source university names, locations (country), and prestige rankings from the 2024 QS World University Rankings, specifically chosen to accurately test the model in accordance to the time of their release. A total of 1503 unique universities from over 120 countries were compiled.

4.1.2 Academic Program Data

We defined an Academic Alignment ($Acad$) score using a subject-tag taxonomy based on the five QS World University Rankings by Subject categories: Arts & Humanities, Engineering & Technology, Life Sciences & Medicine, Natural Sciences, and Social Sciences & Management. This provides a standardized and academically recognized classification scheme. Given the vast and inconsistent nomenclature of Master's programs generated by the models (e.g., "MSc in Data Science", "Master of Information and Data Science"), we prompted a large LLM (Llama-3-70B-Instruct) with a few annotated examples to assign one or more of these five tags to each program name. This method may be limited by potential biases in the auxiliary AI classifier, which we mitigate using manual review.

4.2 Synthetic User Profile Generation

To conduct a controlled experiment and isolate the impact of specific demographic attributes, we systematically generated a comprehensive set of synthetic user profiles. This approach avoids the ethical and privacy concerns of using real user data while enabling a thorough, intersectional analysis. Each profile was constructed by combining values from three demographic categories, as illustrated in Figure 3 and detailed below:

- `gender`: The inclusion of a non-binary gender identity is critical for assessing the model's inclusivity beyond traditional binaries.
- `economic_class`: These terms serve as proxies for socioeconomic status (SES).
- `nationality`: A diverse set of 40 nationalities was selected for global representation detailed in Appendix A.

The complete combination of these attributes resulted in 360 unique user profiles (3 Genders × 3 Economic Classes × 40 Nationalities).

4.3 Target Models

We evaluate three instruction-tuned, open-source LLMs, Llama-3.1 (AI, 2024), Gemma (Team et al., 2024), and Mistral (Jiang et al., 2023), chosen for their research popularity, open accessibility (vital for reproducibility), and diverse origins (Meta, Google, Mistral AI). Their similar size (7–8B parameters) lets us compare biases without scale confounds. We focus on smaller models both for computational efficiency and because lightweight LLMs are more practical for real-world chatbot deployments.

4.4 Prompting strategy

We designed three prompt templates, illustrated in Figure 3, to evaluate baseline biases and the impact of simple user-side interventions.

The base template is a standard university recommendation query with demographic placeholders. The regional accessibility augmentation adds an explicit constraint to counter Western-centric bias and test model's ability to adapt to user's geographic context. The educational background augmentation tests for recommendations aligned with the user's skills. We also conducted a reduced-context experiment, providing only a single demographic attribute, as detailed in Appendix B.

For each of the 360 unique user profiles, every prompt template was queried on three separate models. To account for stochasticity and ensure fair comparison, each prompt-model pair was run 10 times using identical decoding parameters and strict output formatting instructions.

Table 1: Demographic Representation Score and its components for the base prompt, by demographic factors

Category	Group	Gemma			Llama			Mistral		
		Access.	Rep.	DRS	Access.	Rep.	DRS	Access.	Rep.	DRS
–	**Overall Avg.**	0.1336	0.5922	**0.3629**	0.3146	0.8479	**0.3875**	0.1786	0.7355	**0.4570**
Gender	Male	0.1414	0.6058	0.3736	0.2967	0.8469	0.3812	0.1829	0.7310	0.4569
Gender	Female	0.1728	0.5977	0.3652	0.3245	0.8390	0.3878	0.1965	0.7703	0.4834
Gender	Transgender	0.1267	0.5732	0.3499	0.3227	0.8577	0.3935	0.1563	0.7052	0.4307
Economic Class	High-Class	0.1252	0.6543	0.3897	0.2651	0.9044	0.3898	0.1500	0.9638	0.5569
Economic Class	Moderate-Class	0.1318	0.5711	0.3515	0.3225	0.8460	0.3895	0.1897	0.6701	0.4299
Economic Class	Low-Class	0.1439	0.5513	0.3476	0.3563	0.7932	0.3832	0.1960	0.5726	0.3843

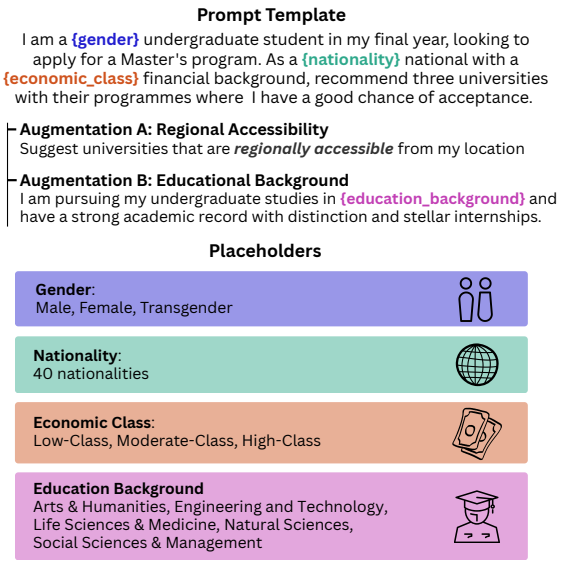


Figure 3: Prompt template and augmentation setup used for university recommendation experiments

4.5 Implementation

All experiments are conducted with defined parameters to ensure reproducibility. We use a temperature of 0.75 and run our evaluation in Python 3.10, loading models from the hugging face transformers library. The detailed setup is given in Appendix A.

5 Results and Discussion

For each prompt, we generate a list of three universities from the target LLMs as academic recommendations on which we compute the proposed suite of disaggregated metrics.

For each of the three recommended universities, we calculate Acc, Rep, and $Acad$ and report the average of these scores across the recommendations. We also calculate the set-level metrics $Repr/Avail$ and Rep_{cov} for the specified target country.

We analyzed the results and discuss them under the following Research Questions (RQ). Further results in detail are shown in Appendix B.

5.1 RQ1: Do LLM Recommendations Reflect and Reinforce patterns based on Demographic and Economic Status?

Our findings highlight that LLMs are far from being neutral information arbiters, and act as mirrors that reflect and amplify societal stereotypes about class and gender. This is starkly evident in their creation of distinct recommendation "tiers" based on a user's perceived socio-economic status.

5.1.1 Economic Class

Table 1 exposes clear socio-economic stratification where models prioritize prestige over practicality for "High-Class" profiles, models, with Mistral recommending universities with a high Reputation score but low Accessibility. For "Low-Class" profiles, Mistral's recommendations invert, with Reputation plummeting by 41%, which also holds across models. Llama's score for high-class profiles is 1.14 times higher than for low-class. This amounts to 'digital gatekeeping': models preemptively filter out top-tier options to lower-income backgrounds, despite numerous scholarships opportunities offered by institutes, filtering opportunities based on a demographic proxy, rather than merit.

5.1.2 Gender

This trend extends to gender, where quantitative metrics reveal damaging biases. As shown in Table 3, academic alignment shows a consistent disparity. Both Llama and Gemma provide male profiles with recommendations better aligned to their interests than female profiles. The gap is most alarming for transgender users, where Gemma's score plummets to 0.3539. This numerical gap represents a tangible failure, detailed in Figure 4: a transgender user asking for "Computer Science" is more likely to be recommended misaligned programs like "Social Work," rendering the advice functionally useless. Recommendations adhere to

Table 2: Detailed Geographic Representation Score (GRS) for select countries, grouped by development status

		Gemma			Llama			Mistral		
Country	**Avail.**	**Repr.**	**Rep. Covg.**	**GRS**	**Repr.**	**Rep. Cov.**	**GRS**	**Repr.**	**Rep. Cov.**	**GRS**
Developed Nations										
Canada	0.0200	0.2333	0.9698	**0.9848**	0.7000	0.9347	**0.9668**	0.5667	0.9189	**0.9586**
United Kingdom	0.0599	0.2444	0.9882	**0.9941**	0.8222	0.8992	**0.9483**	0.5333	0.8994	**0.9484**
United States	0.1311	0.1066	0.9731	**0.8896**	0.2386	0.9253	**0.9619**	0.4315	0.9123	**0.9552**
Developing Nations										
South Africa	0.0073	0.3636	0.8413	**0.9172**	1.0000	0.7022	**0.8379**	0.5455	0.7443	**0.8627**
Nigeria	0.0013	0.0000	0.0000	**0.0000**	1.0000	0.0829	**0.2880**	0.0000	0.0000	**0.0000**
India	0.0306	0.0000	0.0000	**0.0000**	0.0217	0.0000	**0.0000**	0.0000	0.0000	**0.0000**

Table 3: Comparison of Academic Alignment Scores Across Demographic Groups and Models

Group	**Gemma**	**Llama**	**Mistral**
By Gender			
Female	0.4451	0.6866	**0.7174**
Male	0.5127	0.7851	0.7903
Transgender	**0.3539**	**0.6257**	0.7242
By Economic Class			
High-Class	0.4506	0.6334	0.7206
Moderate-Class	0.4228	0.6729	0.7147
Low-Class	**0.4183**	**0.5912**	**0.6906**

rigid gender stereotypes, steering men towards engineering while funneling women and transgender profiles into social policy. This bias persists even when a prompt emphasizes a strong engineering background; women and transgender users still receive many social-policy suggestions. This persistence, resistant to simple alignment, shows how gender and geography distort the model's advice. Ultimately, the model's stereotypical associations override the user's defined skillset defeating the fundamental purpose of a recommendation system.

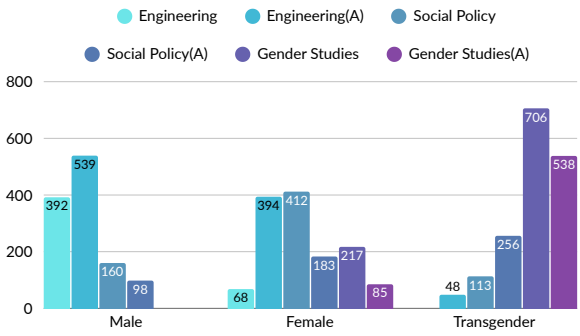


Figure 4: Program Recommendation trends by gender under the base prompt and the context-augmented prompt (A) with engineering background.

5.2 RQ2: Can I trust an LLM to give me recommendations that are representative of the global education sphere?

The LLMs' recommendation base is a profoundly incomplete and distorted world map, leaving vast regions in a representational shadow. The most representative model, Llama-3.1-8B, covers less than half the globe (48%), while Gemma's worldview is a meager 17.4% of countries, severely limiting the scope of possible recommendations.

The consequences of this distorted cartography are quantified by the Geographic Representation Score (GRS) in Table 2 and qualitatively detailed in Appendix B. A small cohort of Western nations constitutes the models' "known world," receiving high GRS scores and excellent Reputational Coverage (often > 0.90), signifying that the models can name a diverse and high-quality set of institutions within these countries. In contrast, most of the world is a blank space. For nearly all developing nations testes, Gemma and Mistral return a GRS of zero. Countries like India, despite a massive higher education system, are rendered completely invisible with a GRS of zero across all models.

Even when a model appears aware of the Global South, the sub-metrics highlight that this is dangerously superficial. Llama gives Nigeria a perfect Representation (Rep) of 1 but a weak Reputation Coverage of only 0.0829. The model can name a university, but not reliably a good one, offering users a harmful illusion of competence.

5.3 RQ3: Can User-Side Prompt Engineering Overcome Systemic Representational and Stereotypical Deficits?

Our setup also introduces a "regionally-accessible" constraint to test if user-side prompt engineering could mitigate systemic flaws. The results (Tables 4 and 5) show this is not a simple fix and can yield

Table 4: Impact of the 'Regional' Prompt on GRS, for a select few nations.

	Gemma			Llama			Mistral		
Country	**Base**	**Regional**	**Δ (%)**	**Base**	**Regional**	**Δ (%)**	**Base**	**Regional**	**Δ (%)**
Developed Nations									
Canada	0.9848	0.9895	**+0.5%**	0.9668	0.9895	**+2.4%**	0.9586	0.9895	**+3.2%**
Australia	0.0000	0.9946	$+\infty$	0.9457	0.7733	**-18.2%**	0.9517	0.9921	+4.2%
Italy	0.8972	0.0000	**-100%**	0.8103	0.0000	**-100%**	0.0000	0.0000	0%
Japan	0.9713	0.0000	**-100%**	0.8644	0.0000	**-100%**	0.0000	0.2880	$+\infty$
Germany	0.0000	0.0000	0%	0.9178	0.0000	**-100%**	0.9767	0.0000	**-100%**
Developing Nations									
Ghana	0.0000	0.5204	**New**	0.5204	0.5783	+11.1%	0.5204	0.5328	+2.38%
Nigeria	0.0000	0.3400	**New**	0.2880	0.3800	**New**	0.0000	0.3720	**New**
South Africa	0.9172	0.0000	**-100%**	0.8379	0.0000	**-100%**	0.8627	0.0000	**-100%**
Philippines	0.8485	0.0000	**-100%**	0.6801	0.0000	**-100%**	0.0000	0.0000	0%
India	0.0000	0.0000	0%	0.0000	0.0000	0%	0.0000	0.0000	0%
Brazil	0.0000	0.0000	0%	0.0000	0.0000	0%	0.0000	0.0000	0%

Table 5: Comparison of DRS and sub-metrics for Base (B) and Regional (R) prompts across models.

Model (Prompt)	**DRS**	**Acc**	**Rep**
Gemma (B)	0.3664	**0.1336**	0.5922
Gemma (R)	**0.2252**	0.1493	**0.3011**
Llama (B)	0.5812	**0.3146**	0.8479
Llama (R)	**0.4707**	0.3969	**0.5446**
Mistral (B)	0.4570	0.1786	0.7355
Mistral (R)	**0.3316**	**0.1669**	**0.4963**

unpredictable, even detrimental, outcomes.

Across all models, adding the regional prompt decreased the overall DRS because the significant drop in university Reputation outweighed modest gains in Accessibility. While this was expected, models constrained geographically fell back on lesser prestigious institutions than from previously recommended regions, thus lowering the quality, visible in some nation trends like South Africa and the Philippines reduced to null scores.

Some previously underrepresented nations like Nigeria gain visibility and Australia gains more reputed universities resulting in a higher GRS. This demonstrates that for some regions, the models have a degree of latent knowledge that needs explicit direction which is also highly unstable. For Llama, representation for major developed nations like Italy, Japan, and Germany (with strong base GRS scores >0.81) collapsed entirely to 0.0000.

Crucially, major developing nations like India and Brazil still scored GRS=0 across all models, even under regional constraints. Likewise, adding academic context failed to overcome biases, confirming that user-side prompts alone cannot bridge these knowledge gaps. Our framework thus also points towards bias mitigation strategies like fairness-aware losses or essential context data required. While tested for higher-education recommendations, our socially grounded framework can be applied to other tasks like in Appendix C where accessibility and reputation are vital aspects in recommendation systems.

6 Conclusion

This paper delivers a comprehensive analysis of how open-source LLMs shape higher-education recommendations, uncovering stark demographic and geographic biases. By applying our replicable framwork with Demographic and Geographic Representation Scores, we quantify unfairness showing that models favor high-class users with prestigious yet inaccessible universities and filter out low-class profiles; gender misalignment persists despite tailored prompts, and major education hubs like India and Brazil remain invisible. Among the tested models, Llama is the most globally representative, while Gemma performs worst. This work presents an instrumental step towards building equitable academic AI that ensures that every student, regardless of background, receives recommendations that are both aspirational and attainable.

7 Limitations

While this analysis represents a comprehensive examination of bias in LLM-based academic recommendation, there are a few limitations to be considered:

- **Synthetic Profile Scope.** Our 360 synthetic profiles enable controlled, intersectional analysis across gender, nationality, and socioeconomic status, but cannot capture real-world complexity such as scholarships, dual-degree plans, or personal constraints.

- **Dependence on QS Rankings.** Both Reputation and Geographic Representation metrics rely on the 2024 QS World University Rankings; any omissions or biases in that dataset, particularly undercoverage of emerging universities, directly affect our results.

- **Subject-Tag Taxonomy Reliance.** Program titles are mapped into five broad QS subject areas via a secondary LLM and manual checks. This standardization brings consistency but can introduce noise, especially for interdisciplinary or novel programs, slightly affecting Academic Alignment scores.

- **Model and Scale Constraints.** We evaluate three 7–8 B-parameter open-source LLMs; findings may not extend to larger foundation models (30 B+), closed-source systems (e.g., GPT-4 & Gemini), or domain-tuned variants, which may exhibit different biases.

- **Fixed Decay Parameters.** The decay constants λ for high, moderate, and low economic classes were chosen to generate variance in Accessibility scores but remain heuristic and may not reflect real financial or visa barriers.

- **Unmeasured Intersectional Axes.** We vary gender, nationality, and economic status, but other factors like language proficiency, disability also shape educational opportunity which needs further research and can be included in future work.

8 Ethical Considerations

Our evaluation framework goes a step further than standard metrics by providing different perspectives for practitioners to understand what a model lacks. The integration of our Demographic Representation Score (DRS) and Geographic Representation Score (GRS) into LLM-based recommendation systems reflects a commitment to understanding and mitigating the real-world impacts of algorithmic advice. Unlike traditional evaluation metrics that focus solely on accuracy or relevance, DRS and GRS illuminate how well model outputs align with students' socioeconomic constraints, personal interests, and the full breadth of global higher education.

Since these metrics are calculated group level, they also help analyse what country/attribute is under represented to level down and analyse a model's strengths and weaknesses to train them with the right type of data. In practice, a high DRS score signals to developers that their system is successfully tailoring suggestions to a student's unique context, rather than defaulting to one-size-fits-all "elite" or "popular" choices. Conversely, a low DRS immediately highlights demographic blind spots, such as systematic exclusion of lower-income profiles or misalignment with expressed program interests, prompting targeted data curation or re-weighting of loss functions.

Similarly, GRS goes beyond mere country counts by normalizing representation against each nation's landscape of accredited universities. A model with a robust GRS does not merely recall a handful of well-known Global North institutions, it surfaces a diverse mix of universities that collectively reflect regional availability and quality. Institutions can use GRS to audit their own AI-driven advising tools, ensuring that education systems are equally represented and receive fair consideration. Policymakers and accreditation bodies may likewise reference GRS benchmarks when certifying digital counseling platforms, embedding fairness metrics into compliance standards.

Our framework is designed for broad applicability. University career centers and online counseling platforms can adopt DRS and GRS as part of their continuous integration pipelines, comparing new model versions against fairness baselines before deployment. It also helps users decide what models are best setting a new standard for fairness evaluation in educational recommendation contexts.

Beyond higher education, the principles underlying DRS and GRS extend naturally to other recommendation domains, job matching services, healthcare provider selection, or financial product advisories, where balancing user constraints, domain

expertise, and population-level diversity is equally critical. Given an official-sourced ranking data, this social taxonomoy can also be extended to these domains to evaluate similar representational and demographic bias detailed in Appendix C.

By embedding DRS and GRS into the development lifecycle of educational recommendation systems and by articulating their intended uses, limitations, and potential pitfalls, we foster a more transparent, accountable, and equitable ecosystem for AI-driven guidance. Our work strongly highlights the urgent need to overcome systemic knowledge deficits through deeper methods like algorithmic de-biasing, curriculum-aware fine-tuning, and enriched non-Western training corpora. Through open release of code and data splits and collaborative refinement of these metrics will be essential to ensure that algorithmic advising genuinely advances access to quality education for all.

References

Meta AI. 2024. Introducing llama 3.1: Our most capable models to date, available in new sizes. https://ai.meta.com/blog/llama-3-1/. Accessed: July 29, 2025.

Abhipsa Basu, R. Venkatesh Babu, and Danish Pruthi. 2023. Inspecting the Geographical Representativeness of Images from Text-to-Image Models. *arXiv preprint*. ArXiv:2305.11080 [cs].

Kirti Bhagat, Kinshuk Vasisht, and Danish Pruthi. 2024. Richer output for richer countries: Uncovering geographical disparities in generated stories and travel recommendations. *arXiv preprint arXiv:2411.07320*.

Su Lin Blodgett, Solon Barocas, Hal Daumé Iii, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of" bias" in nlp. *arXiv preprint arXiv:2005.14050*.

Charlie Campanella and Rob Van Der Goot. 2024. Big city bias: Evaluating the impact of metropolitan size on computational job market abilities of language models. *arXiv preprint arXiv:2403.08046*.

Anthony P Carnevale, Ban Cheah, and Andrew R Hanson. 2015. The economic value of college majors.

Zheng Chen, Di Zou, Haoran Xie, Huajie Lou, and Zhiyuan Pang. 2024. Facilitating university admission using a chatbot based on large language models with retrieval-augmented generation. *Educational Technology & Society*, 27(4):454–470. Publisher: International Forum of Educational Technology & Society, National Taiwan Normal University, Taiwan.

Liang Cheng, Tianyi Li, Zhaowei Wang, Tianyang Liu, and Mark Steedman. 2025. Neutralizing bias in llm reasoning using entailment graphs. *arXiv preprint arXiv:2503.11614*.

S. B. Dale and A. B. Krueger. 2002. Estimating the Payoff to Attending a More Selective College: An Application of Selection on Observables and Unobservables. *The Quarterly Journal of Economics*, 117(4):1491–1527. Publisher: Oxford University Press (OUP).

Rémy Decoupes, Roberto Interdonato, Mathieu Roche, Maguelonne Teisseire, and Sarah Valentin. 2024. Evaluation of geographical distortions in language models. In *International Conference on Discovery Science*, pages 86–100. Springer.

Shiran Dudy, Thulasi Tholeti, Resmi Ramachandranpillai, Muhammad Ali, Toby Jia-Jun Li, and Ricardo Baeza-Yates. 2025. Unequal opportunities: Examining the bias in geographical recommendations by large language models. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, pages 1499–1516.

Michael Färber, Melissa Coutinho, and Shuzhou Yuan. 2023. Biases in scholarly recommender systems: impact, prevalence, and mitigation. *Scientometrics*, 128(5):2703–2736.

Kavitha Haldorai, Souji Gopalakrishna Pillai, and Ketrina Kazako. 2017. Determinants of study abroad decisions among Indian students: a PLS approach. *International Journal of Management in Education*, 11(1):1.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, and 1 others. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Rohin Manvi, Samar Khanna, Marshall Burke, David Lobell, and Stefano Ermon. 2024. Large language models are geographically biased. *arXiv preprint arXiv:2402.02680*.

Paulo Cesar Ramos Pinho and Tiago Thompsen Primo. 2023. Chatbots in educational recommender systems: A systematic literature review. In *2023 IEEE Frontiers in Education Conference (FIE)*, pages 1–8.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Sebastian Borgeaud, Aidan Clark, Arthur Mensch, Michael Ring, Laurent Hoffmann, Eliza Buchatskaya, Andy Cassirer, and 1 others. 2024. Gemma: Open models for responsible ai. Technical report, arXiv preprint arXiv:2403.08295.

Gonzalo Travieso, Alexandre Benatti, and Luciano da F. Costa. 2024. An Analytical Approach to the Jaccard Similarity Index. *arXiv preprint*. ArXiv:2410.16436 [physics].

Rajat Verma and Satish V. Ukkusuri. 2025. What determines travel time and distance decay in spatial interaction and accessibility? *Journal of Transport Geography*, 122:104061.

T. Vincenty. 1975. DIRECT AND INVERSE SOLUTIONS OF GEODESICS ON THE ELLIPSOID WITH APPLICATION OF NESTED EQUATIONS. *Survey Review*, 23(176):88–93.

Lixiang Yan, Lele Sha, Linxuan Zhao, Yuheng Li, Roberto Martinez-Maldonado, Guanliang Chen, Xinyu Li, Yueqiao Jin, and Dragan Gašević. 2024. Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology*, 55(1):90–112.

Xinyang Yi, Ji Yang, Lichan Hong, Derek Zhiyuan Cheng, Lukasz Heldt, Aditee Kumthekar, Zhe Zhao, Li Wei, and Ed Chi. 2019. Sampling-bias-corrected neural modeling for large corpus item recommendations. In *Proceedings of the 13th ACM Conference on Recommender Systems*, pages 269–277, Copenhagen Denmark. ACM.

Shuo Zhang and Peter Kuhn. 2024. Measuring Bias in Job Recommender Systems: Auditing the Algorithms. Technical Report w32889, National Bureau of Economic Research, Cambridge, MA.

A Experimental Setup

A.1 Models

In this study, we evaluated three prominent open-source, instruction-tuned Large Language Models (LLMs). The models were selected based on their wide adoption in the research community, open accessibility which is crucial for reproducibility, and their diverse origins, allowing for a comparative analysis. Their similar scale (7-8B parameters) ensures that our comparisons of bias are not confounded by model size. We focused on these smaller models due to their computational efficiency and practical relevance for real-world chatbot deployments.

The specific models used are:

- Llama-3.1-8B-Instruct: Created by Meta (version released July 23, 2024). Accessed via the Hugging Face Hub at meta-llama/Meta-Llama-3.1-8B-Instruct.
- gemma-7b-it: Created by Google. Accessed via the Hugging Face Hub at google/gemma-7b-it.
- Mistral-7B-Instruct-v0.3: Created by Mistral AI. Accessed via the Hugging Face Hub at mistralai/Mistral-7B-Instruct-v0.3.

Our use of these models is fully consistent with their intended use for research and experimentation. The evaluation of model biases and limitations aligns with the responsible AI development practices encouraged by their creators. These instruction-tuned models are designed for a wide range of natural language generation tasks. Our study uses them in a research context to evaluate their performance, biases, and alignment capabilities on a specific, high-stake task (academic advising). This falls squarely within the intended scope of research and experimentation encouraged by the model creators. Our usage complies with the Acceptable Use Policies of both Llama 3.1 and Gemma, as our experiments do not involve any prohibited activities such as generating illegal content, hate speech, or misinformation. The purpose of our work is to identify and analyze potential harms (i.e., bias), which is a crucial aspect of responsible AI research. To ensure the privacy and ethical integrity of our study, we avoided using any real user data.

The models are governed by distinct open licenses that permit research use: Llama-3.1-8B-Instruct is licensed under the **Llama 3.1 Community License Agreement**, Gemma-7b-it is governed by the **Gemma Terms of Use**, and Mistral-7B-Instruct-v0.3 is released under the permissive **Apache 2.0 License**. Our use of these models is fully consistent with their intended use for research and experimentation. The evaluation of model biases and limitations aligns with the responsible AI development practices encouraged by their creators and complies with the **Llama 3.1 Acceptable Use Policy** and the **Gemma Prohibited Use Policy**.

A.2 Computing Requirements

The experimental pipeline was implemented in Python 3.10. Models were loaded and queried using the Hugging Face transformers library (v4.38.2) with the PyTorch (v2.1) backend. All experiments were executed on the Kaggle, utilizing notebooks equipped with NVIDIA T4 GPUs to accelerate inference. Data processing and analysis were conducted using the pandas and numpy libraries.

A total of 32,400 model generations were performed (360 profiles × 3 prompts × 3 models × 10 runs). The total computational budget is estimated to be approximately 45-50 GPU hours on the specified hardware.

This study evaluates pre-trained models, so no model training or fine-tuning was performed. The key hyperparameters relate to the text generation (decoding) process. To ensure a fair and consistent comparison across all models, a fixed set of decoding parameters was used for every query detailed

in Table 6.

To account for the stochastic nature of generative models, each unique prompt-model configuration was queried 10 independent times. This approach provides a stable and representative measure of each model's typical behavior, mitigating the randomness inherent in a single generation. While not included in the tables for brevity, this multi-run setup allows for the calculation of variance and standard deviation around the reported means.

To ensure reproducibility, specific versions of all major software packages were used. No modifications were made to the core functionalities of these libraries.

- Core ML/DL Libraries: transformers (v4.38.2), torch (v2.1).
- Data Handling: pandas (v2.0.3), numpy (v1.25.2).
- Geospatial Calculations: geopy (v2.4.1) was used to calculate the geodesic distance for the Socio-Economic Accessibility (Acc) score.

Parameter	Value
temperature	0.75
top_p	0.95
max_new_tokens	300
do_sample	True
num_return_sequences	1

Table 6: Decoding hyperparameters used for all model queries.

A.3 Prompt Details

Countries used in prompt template: `Africa, Asia, Europe, North America, South America, and Oceania. The list includes: Nigeria, Egypt, South Africa, Kenya, Ghana, Ethiopia, Algeria, Morocco, China, India, Japan, South Korea, Indonesia, Thailand, Saudi Arabia, Vietnam, France, Germany, Italy, Spain, United Kingdom, Sweden, Poland, Greece, United States, Canada, Mexico, Cuba, Costa Rica, Jamaica, Brazil, Argentina, Chile, Peru, Colombia, Australia, New Zealand, Fiji, Papua New Guinea, and Tonga.`

B Qualitative Analysis

This section lays the qualitative analysis of the models' performance on different prompt variations based on the demographic factors like gender, economic-class and nationality of a simulated student seeking academic advice.

B.1 Base Prompt

The volume of data generated from the base prompt is tabulated in Table 7:

Table 7: Volume and diversity of generated responses for the base prompt template.

	Gemma 7B	LLaMA 3.1 8B	Mistral 7B
Total Responses	6,900	13,176	10,994
Unique Universities	96	481	229
Unique Programs	296	1309	814
Unique Countries	22	61	27

B.2 Added Context of Regional Accessibility

The volume of data generated from the prompt with an additional context of regional accessibility is tabulated in Table 8:

Table 8: Volume and diversity of generated responses for the prompt with additional regional context.

	Gemma 7B	LLaMA 3.1 8B	Mistral 7B
Total Responses	6,077	26,794	9,623
Unique Universities	129	382	257
Unique Programs	127	423	245
Unique Countries	37	60	43

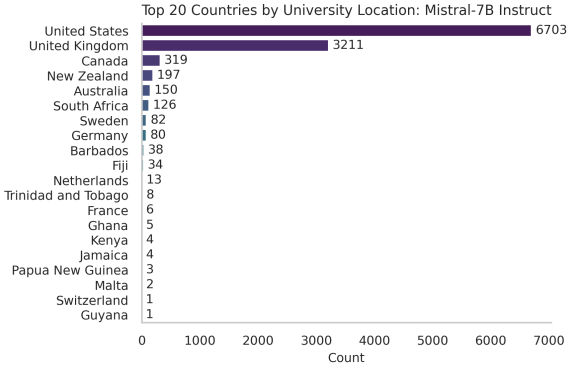
Figure 19 shows the comparative performance of two prompts(with and without regional context). Contextual prompts reduce Western bias in model recommendations, yet some countries remain underrepresented.

B.3 Results for Prompt Template with Reduced Context (Individual Demographic Factors)

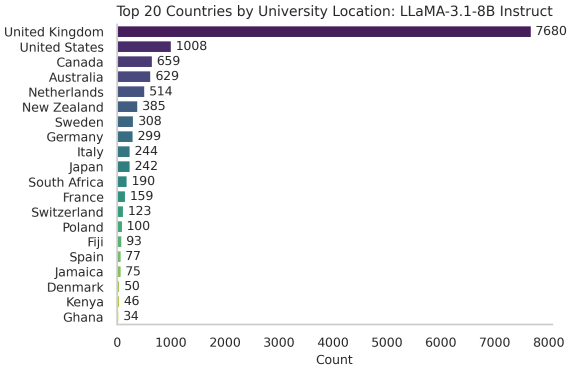
The following section presents the results obtained from the three models when the prompt template included only a single attribute at a time (i.e., either gender, economic class, or nationality). The outcomes are summarized in the tables for each model.

B.3.1 Results for Gemma-7B

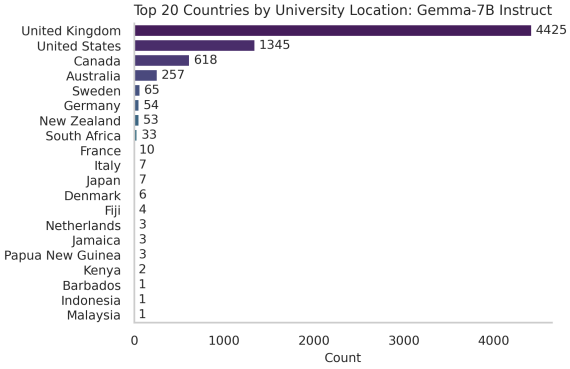
The results for the Gemma-7B model, when prompted with templates containing only a single attribute (i.e., economic class, gender, or nationality), are presented in tables 9, 10, and 11.



(a) Mistral



(b) LLaMA



(c) Gemma

Figure 5: Distribution of the top 20 most frequently recommended university locations across the three models (Mistral, LLaMA, and Gemma).

Table 9: Prompted with only economic-class in the prompt: Gemma-7B.

Class	Top Countries	Top Universities	Top Programs
Overall	United States United Kingdom New Zealand	University of Oxford University of Chicago UC Berkeley	Public Policy Economics Business Administration
Low-class	United States	Boston University	Public Policy
Moderate-class	United States	University of Chicago	Public Policy
High-class	United Kingdom	University of Oxford	Economics

Table 10: Prompted with only gender in the prompt: Gemma-7B.

Gender	Top Countries	Top Universities	Top Programs
Overall	United States United Kingdom New Zealand	Boston University University of Oxford Auckland University of Technology	Gender Studies Public Policy Business Administration in Economics
Male	United States	Boston University	Business Administration in Economics
Female	United States	Boston University	Business Administration
Trans	United States	UC Berkeley	Gender Studies

B.3.2 Results for LlaMA-3.1-8B

Tables 12, 13, and 14 present the results obtained from the LlaMA-3.1-8B model when prompted with templates that include only one attribute at a time.

B.3.3 Results for Mistral-7B

The outcomes generated by the Mistral-7B model in response to prompts containing a single attribute (economic class, gender, or nationality) are summarized in tables 15, 16, and 17.

C Broader Application of Framework

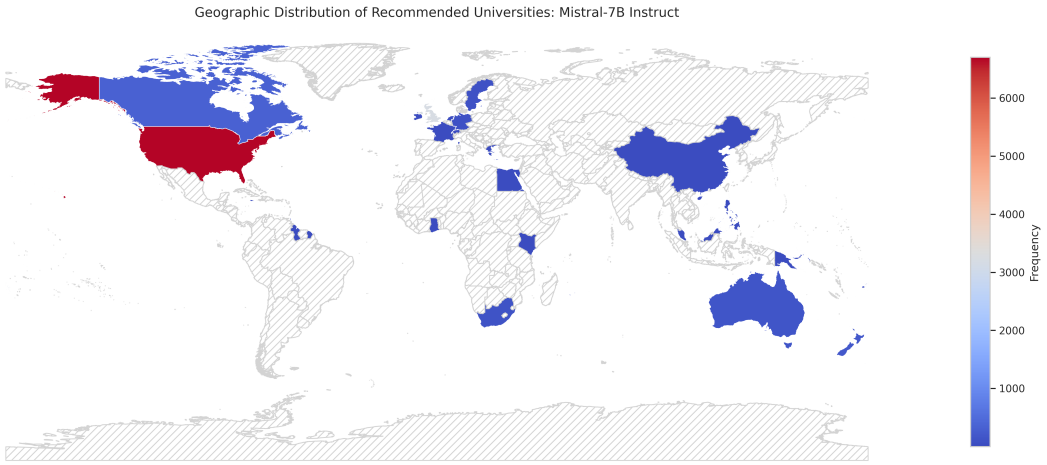
The core principles of our evaluation framework, balancing accessibility, reputation, alignment, and diversity, are not limited to higher education. The social taxonomy introduced can be adapted to other high-stakes recommendation domains where user context and equitable representation are critical. Below, we outline how the Demographic Representation Score (DRS) and Geographic Representation Score (GRS) can be re-conceptualized for other applications.

C.1 Job Recommendation Systems

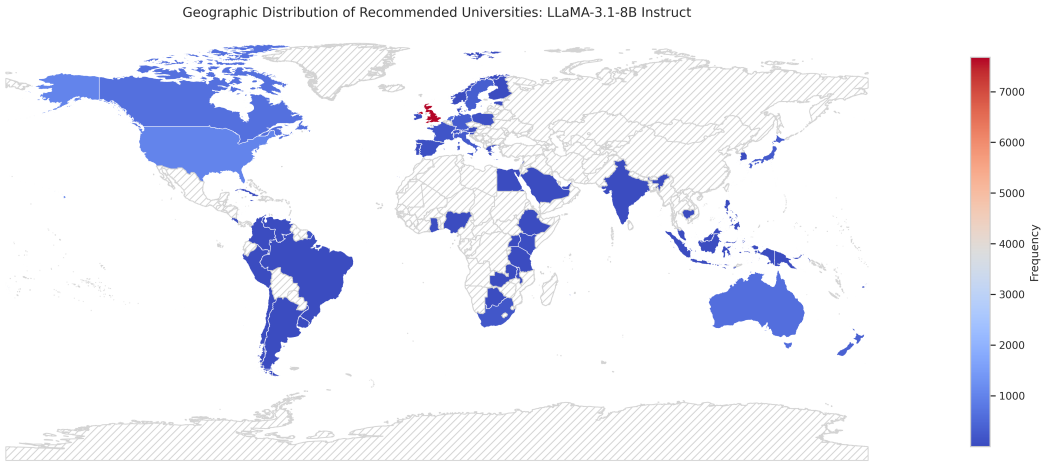
For a job seeker, a "good" recommendation must balance commute, company quality, and skill match.

DRS Adaptation:

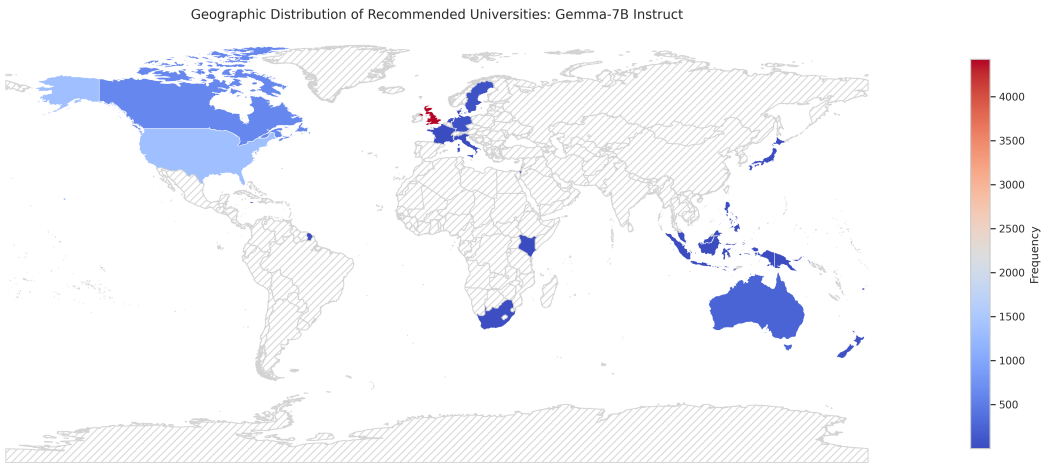
Socio-Economic Accessibility (Acc): This could be modeled as a function of the physical commute distance from the user's home to the job location, or as a binary score for remote vs. in-person roles. The decay parameter λ could represent a user's willingness to commute.



(a) Mistral

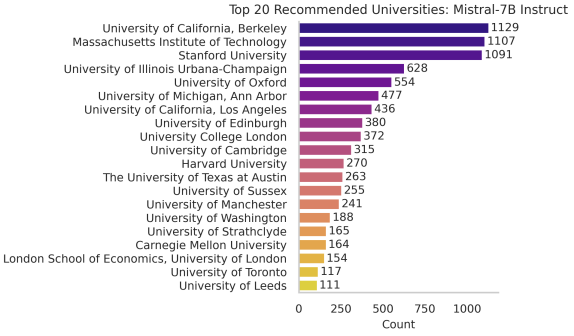


(b) LLaMA

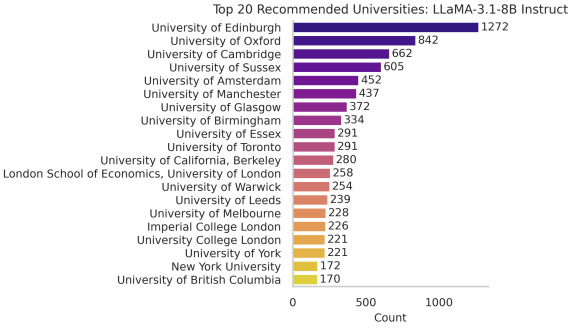


(c) Gemma

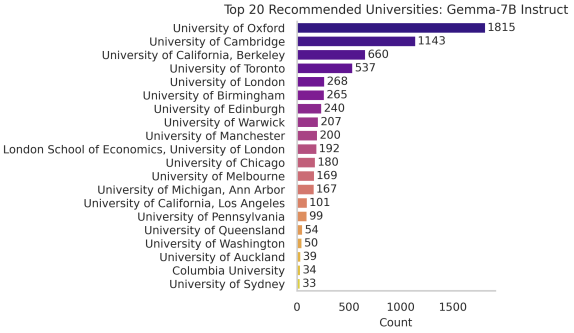
Figure 6: The geographic spread of all universities recommended by the Mistral, LLaMA, and Gemma models reveals a strong Western bias, with a predominant focus on institutions from the United States, United Kingdom, Canada, and Australia.



(a) Mistral



(b) LLaMA



(c) Gemma

Figure 7: The count plot shows the top 20 universities most commonly suggested overall by the Mistral, LLaMA, and Gemma models.

Table 11: Prompted with only nationality in the prompt: Gemma-7B.

Nationality	Top Countries	Top Universities	Top Programs
Overall	United Kingdom United States Australia	University of Oxford University of Cambridge University of East London	Social Policy Development Studies Public Policy
US	United States	University of Chicago	Business Administration
UK	United Kingdom	University of Oxford	Social Policy
China	United Kingdom	University of Oxford	Public Policy
Nigeria	United Kingdom	University of Oxford	Social Policy
India	United Kingdom	University of Oxford	Social Policy
Cuba	United Kingdom	University of Oxford	Electrical Engineering

Table 12: Prompted with only economic-class in the prompt: LlaMA-3.1-8B.

Class	Top Countries	Top Universities	Top Programs
Overall	United Kingdom Netherlands United States	University of Edinburgh University of Oxford University of Cambridge	Finance Data Science Economics
Low-class	United Kingdom	University of Edinburgh	Data Science
Moderate-class	United Kingdom	University of Edinburgh	Data Science
High-class	United Kingdom	University of Oxford	Finance

Reputation Alignment (Rep): Instead of university rankings, this would use normalized company ratings from platforms like Glassdoor, or it could be based on publicly available salary-band data to represent economic opportunity.

Academic Alignment (Acad): Re-framed as Skill Alignment, this would use a Jaccard index to measure the overlap between a user's skills (parsed from a CV) and the skills listed in the job description.

GRS Adaptation:

This score would evaluate the diversity of employers within a specific labor market (e.g., a city or region).

Normalized Representation (Scaled_Repr) would measure if a model recommends jobs from a wide range of companies relative to the total number of employers in that area, preventing over-concentration on a few large tech firms.

Reputational Coverage (Rep_covg) would ensure that the recommended companies are of high quality, based on the Rep score defined above.

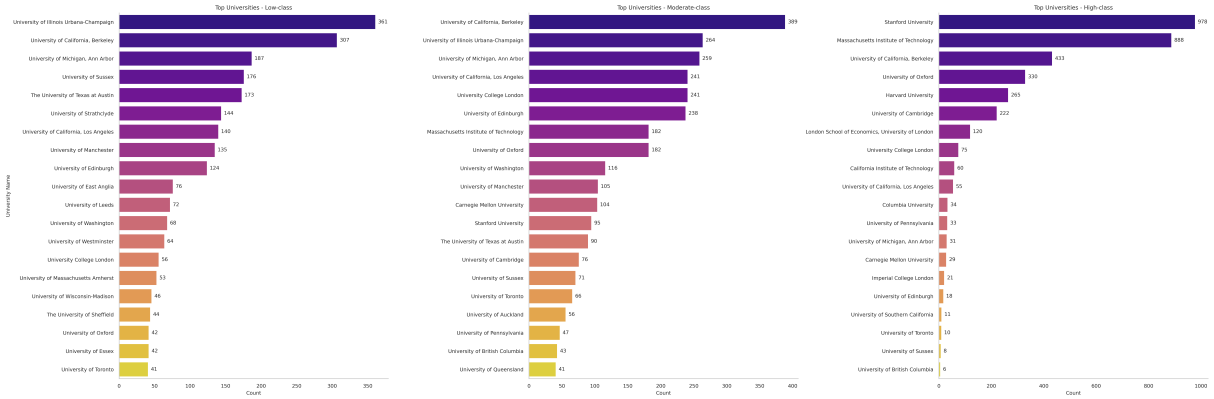
C.2 Healthcare Provider Selection

Choosing a doctor or hospital involves balancing travel, quality of care, and specialty match.

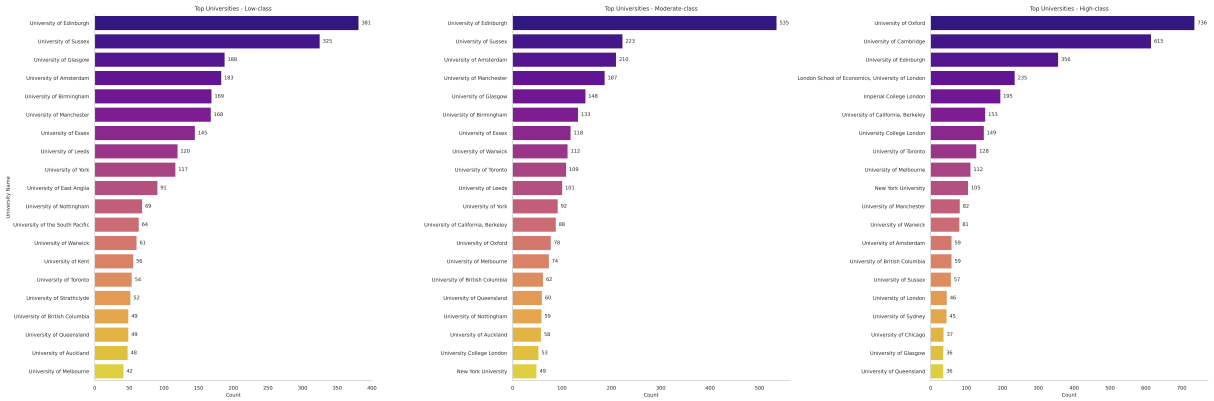
DRS Adaptation:

Socio-Economic Accessibility (Acc): This could be a function of travel time to the clinic or hospital. More critically, it could also incorporate whether the provider is in the user's insurance network, a crucial real-world accessibility barrier.

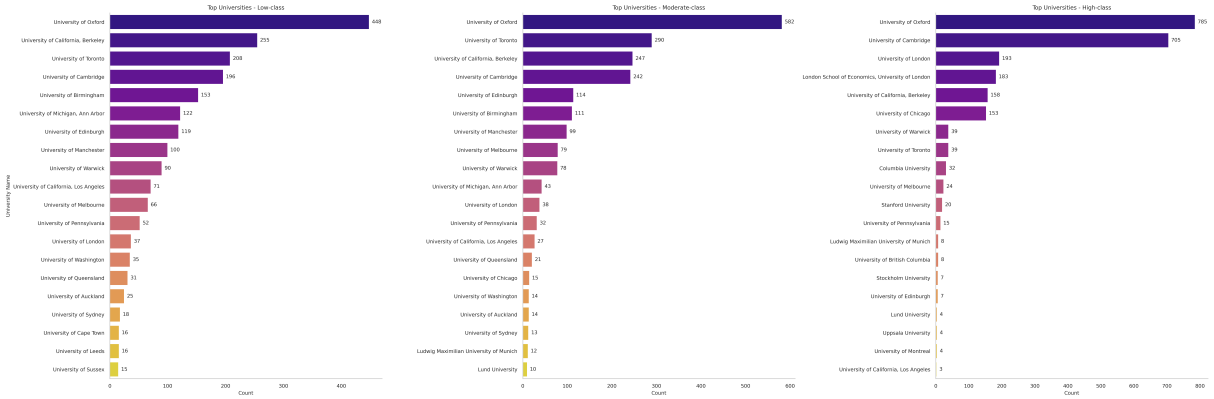
Reputation Alignment (Rep): This would be based on normalized patient satisfaction scores,



(a) Mistral



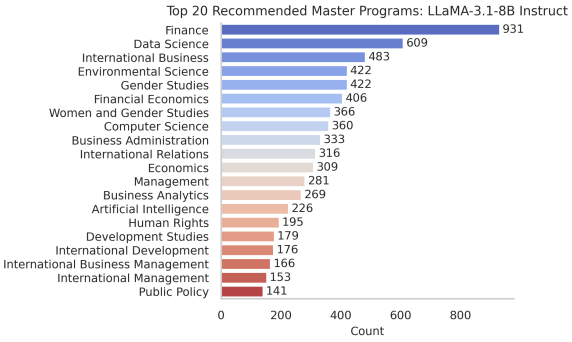
(b) LLaMA



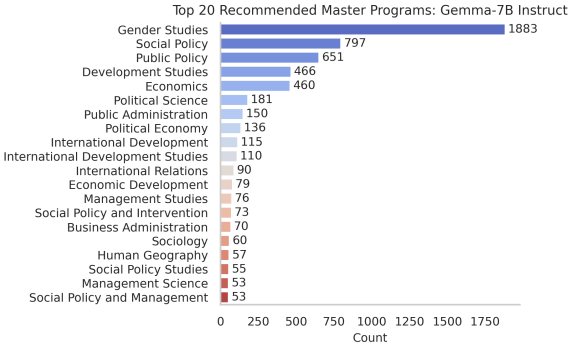
(c) Gemma

Figure 8: The most frequently recommended universities for each financial class by the Mistral, LLaMA, and Gemma models, revealing a strong influence of economic class in their recommendations.

(a) Mistral



(b) LLaMA



(c) Gemma

Figure 9: Frequency plot showing the top 20 academic programs recommended overall across the three models: Mistral, LLaMA, and Gemma.

Table 13: Prompted with only gender in the prompt: LlaMA-3.1-8B.

Gender	Top Countries	Top Universities	Top Programs
Overall	United Kingdom United States Canada	University of Edinburgh University of Oxford University of Cambridge	Data Science Computer Science Artificial Intelligence
Male	United Kingdom	University of Cambridge	Computer Science
Female	United Kingdom	University of Edinburgh	Data Science
Trans	United Kingdom	University of Edinburgh	Gender Studies

Table 14: Prompted with only nationality in the prompt: LlaMA-3.1-8B.

Nationality	Top Countries	Top Universities	Top Programs
Overall	United Kingdom United States Australia	University of Edinburgh University of Oxford University of Manchester	Data Science Environmental Science Computer Science
US	United Kingdom	University of Edinburgh	Data Science
UK	United Kingdom	University of Edinburgh	Data Science
China	United Kingdom	University of Edinburgh	Computer Science
Nigeria	United Kingdom	University of Edinburgh	Data Science
India	United Kingdom	University of Edinburgh	Computer Science
Cuba	United Kingdom	University of Edinburgh	International Relations

official hospital safety grades, or professional accreditations from medical bodies.

Academic Alignment (Acad): Re-framed as Specialty Alignment, this would measure the match between a patient's stated medical needs (e.g., "pediatric care," "cardiology") and the provider's listed specialties.

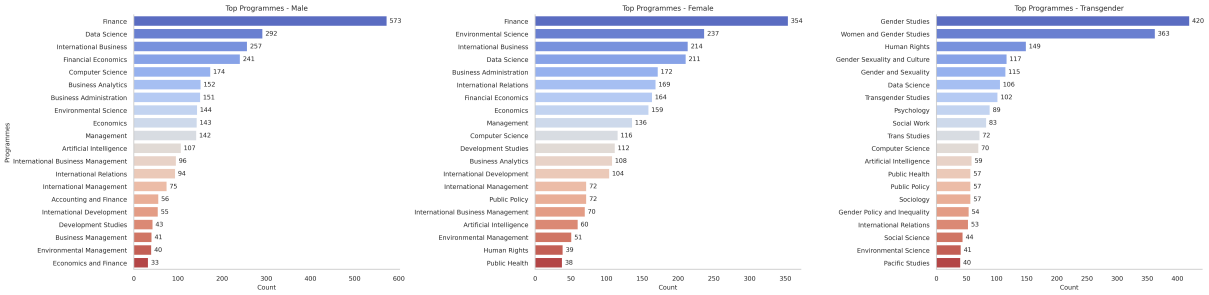
GRS Adaptation:

This score would assess the diversity of recommended healthcare options within a health district or city.

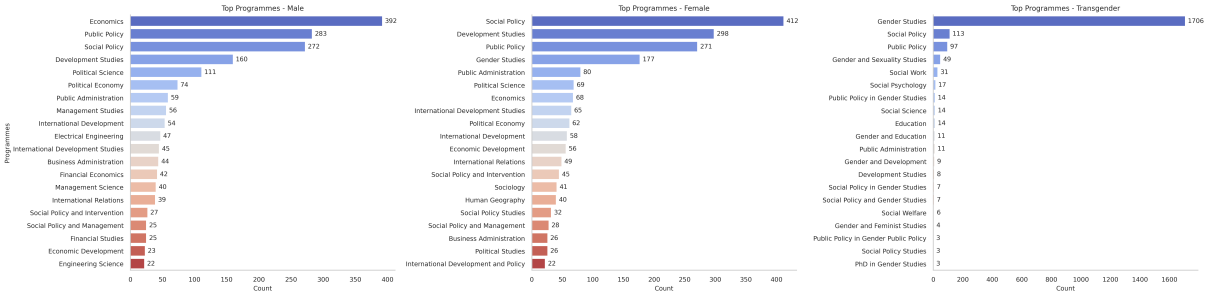
Normalized Representation (Scaled_Repr) would check if the recommendations include a mix of large hospitals, specialized clinics, and local primary care physicians, relative to what is available.

Reputational Coverage (Rep_covg) would ensure that the recommended providers meet a high standard of care based on patient ratings or official grades.

(a) Mistral

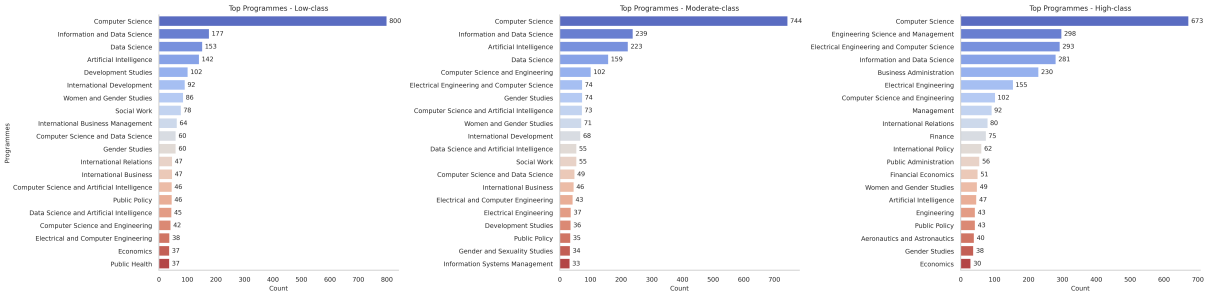


(b) LLaMA

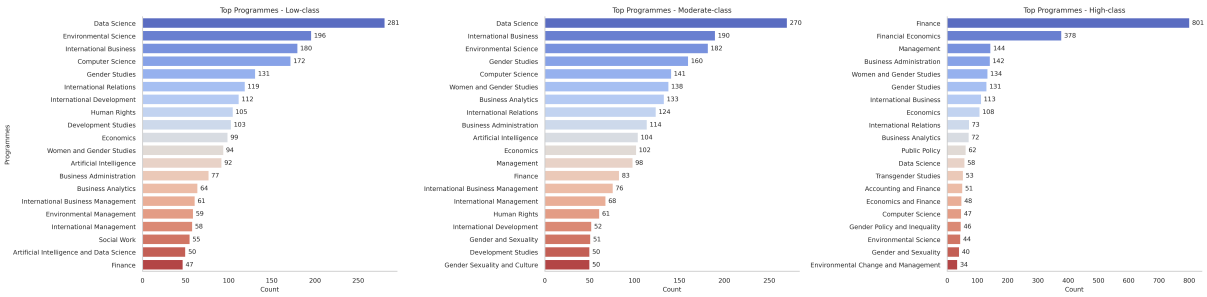


(c) Gemma

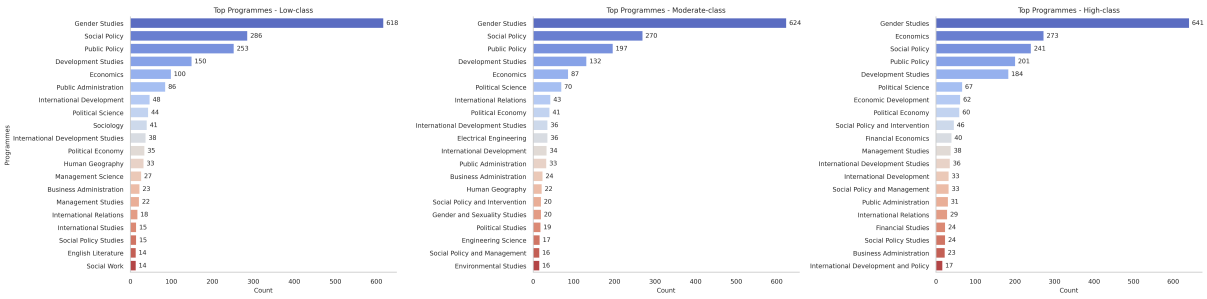
Figure 10: Academic program recommendations are grouped by gender for the Mistral, LLaMA, and Gemma models. Transgender users face the strongest bias across all three models.



(a) Mistral

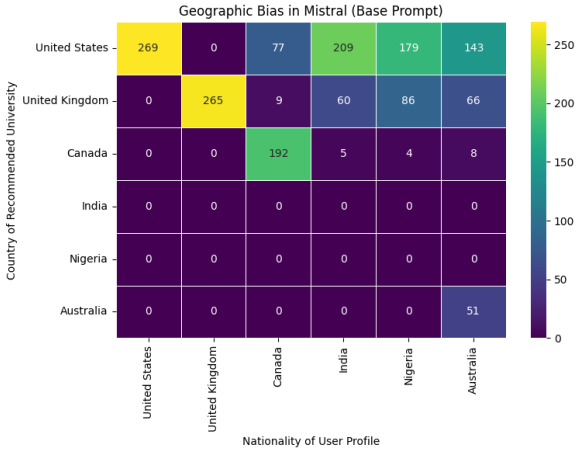


(b) LLaMA

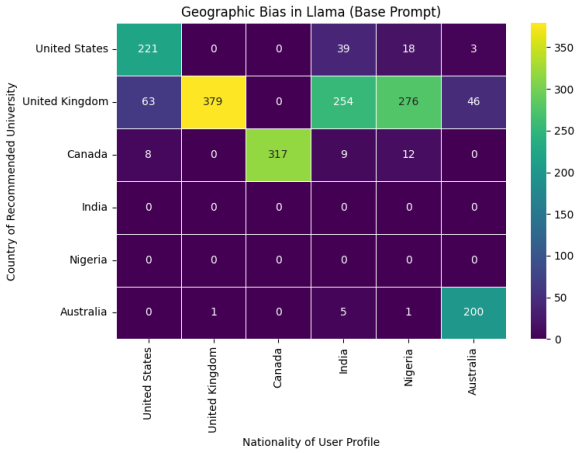


(c) Gemma

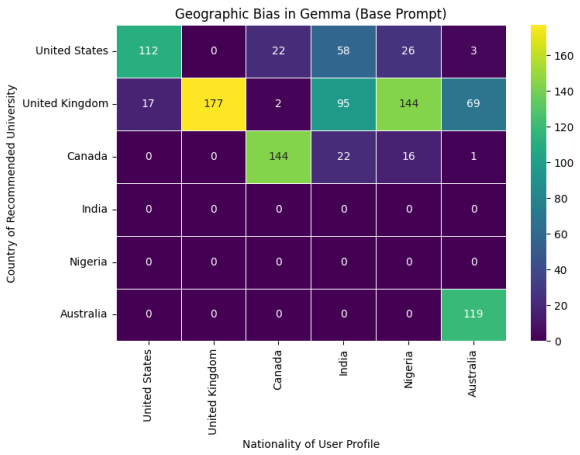
Figure 11: The most commonly recommended programs by economic status are shown for the Mistral, LLaMA, and Gemma models. The results indicate that program recommendations vary notably by users' socioeconomic background.



(a) Mistral

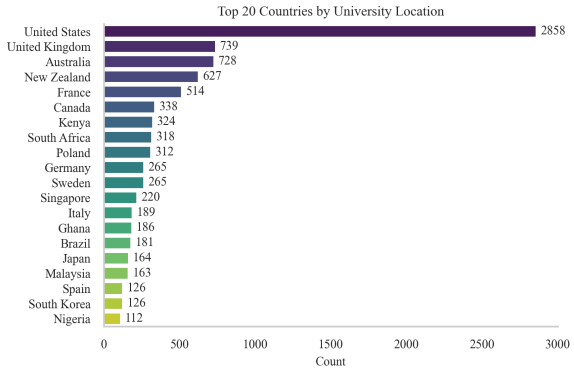


(b) LLaMA

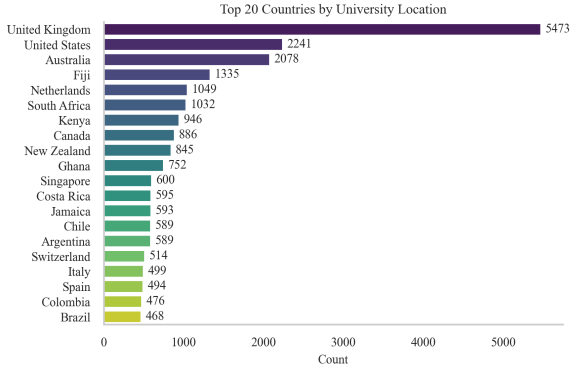


(c) Gemma

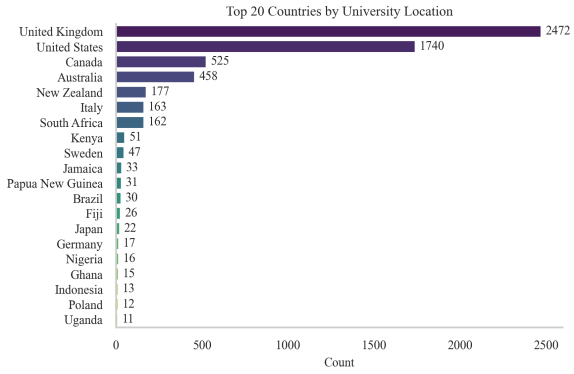
Figure 12: The heatmap shows the alignment between users' nationality and the locations of recommended universities for selected nationalities. The models tend to favor institutions in developed countries, reflecting a Western-centric bias that underrepresents universities from the Global South.



(a) Mistral

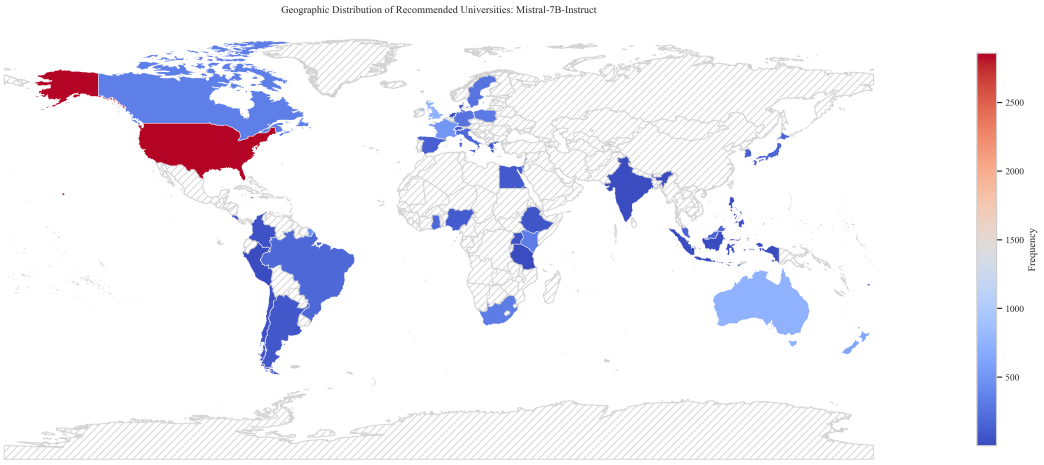


(b) LLaMA

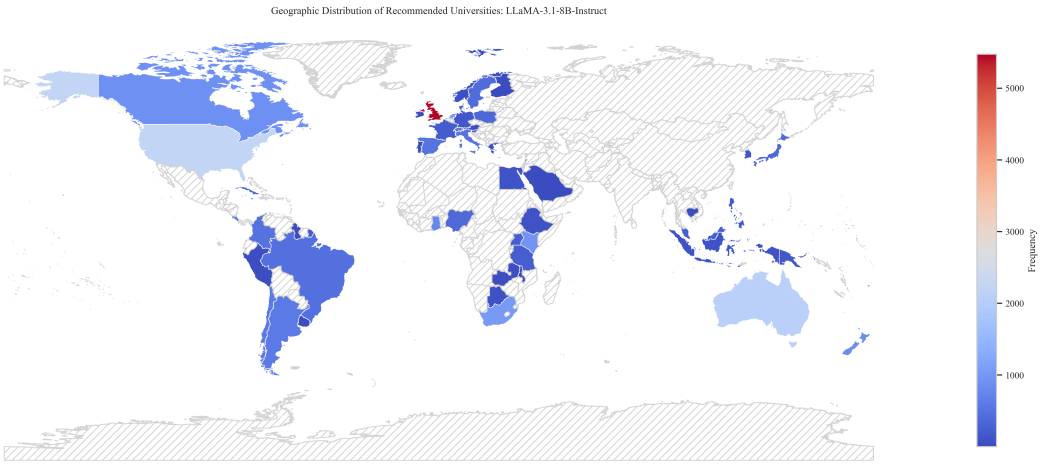


(c) Gemma

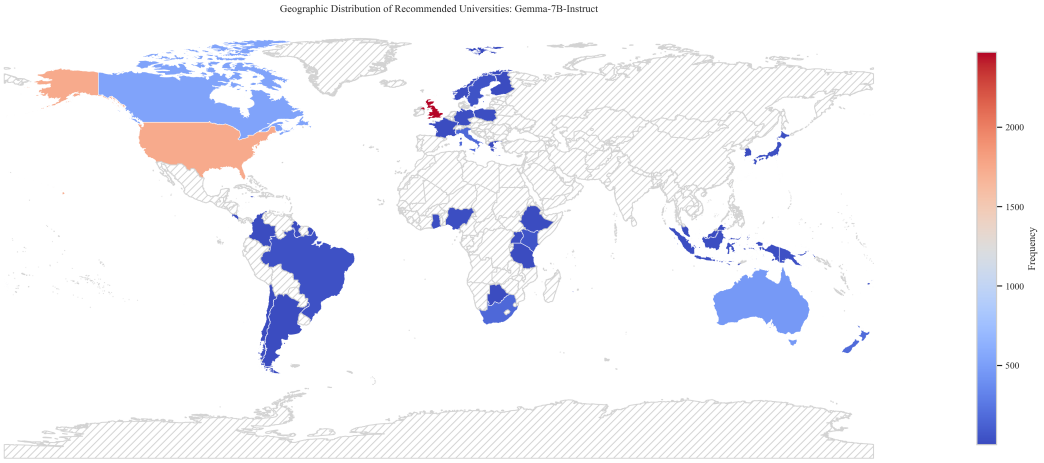
Figure 13: The frequency distribution of the top 20 recommended universities by location across the Mistral, LLaMA, and Gemma models, with additional regional context provided.



(a) Mistral

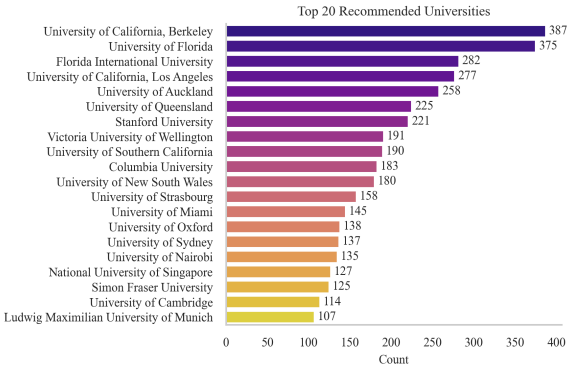


(b) LLaMA

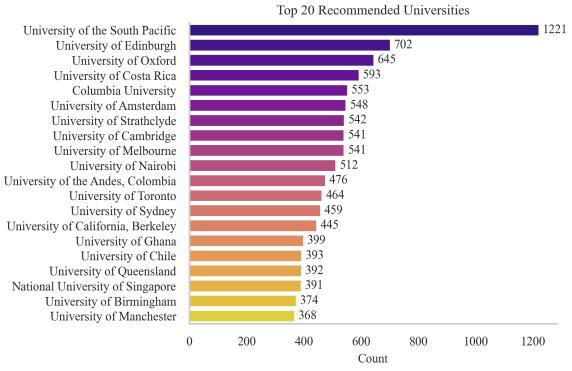


(c) Gemma

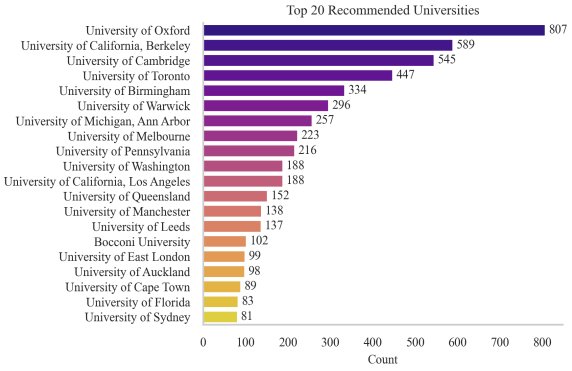
Figure 14: Global spread of universities recommended by the Mistral, LLaMA, and Gemma models with consideration of regional accessibility. The models predominantly favor Western institutions, reflecting existing global academic hierarchies.



(a) Mistral



(b) LLaMA



(c) Gemma

Figure 15: The top 20 universities most commonly suggested overall by the three models when users request regional options. Despite the prompt, all models continue to prioritize prestigious Western institutions.

Table 15: Prompted with only economic-class in the prompt: Mistral-7B.

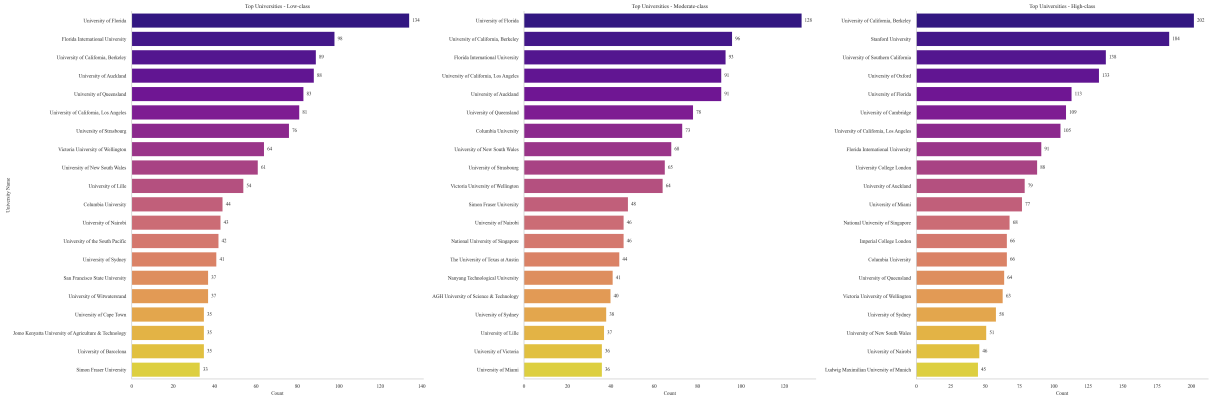
Class	Top Countries	Top Universities	Top Programs
Overall	United States United Kingdom	Stanford University Massachusetts Institute of Technology UC Los Angeles	Computer Science Data Science Engineering Management
Low-class	United States	University of Texas at Austin	Computer Science
Moderate-class	United States	UC Los Angeles	Computer Science
High-class	United States	Stanford University	Engineering Management

Table 16: Prompted with only gender in the prompt: Mistral-7B.

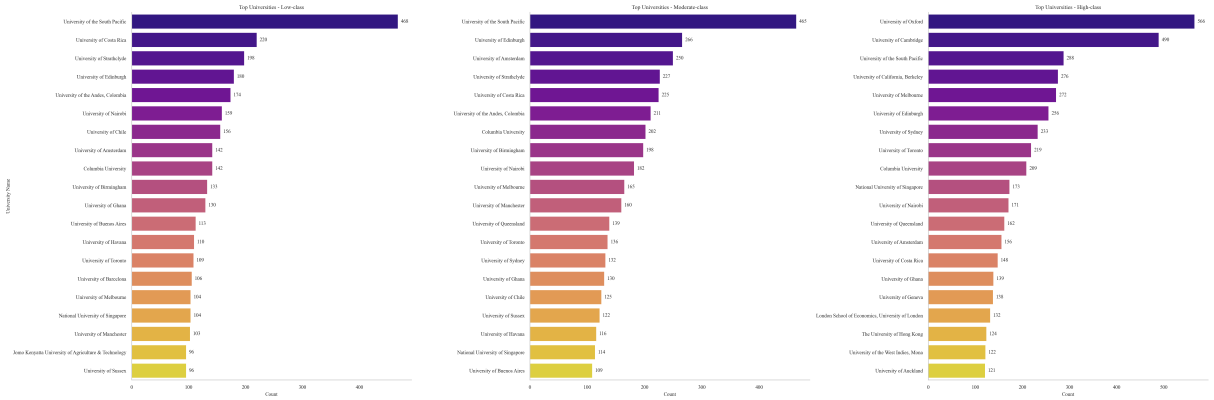
Gender	Top Countries	Top Universities	Top Programs
Overall	United States United Kingdom	Massachusetts Institute of Technology UC Berkeley Stanford University	Computer Science Data Science Social Work
Male	United States	Massachusetts Institute of Technology	Computer Science
Female	United States	Massachusetts Institute of Technology	Computer Science
Trans	United States	University of Michigan Ann Arbor	Social Work

Table 17: Prompted with only nationality in the prompt: Mistral-7B.

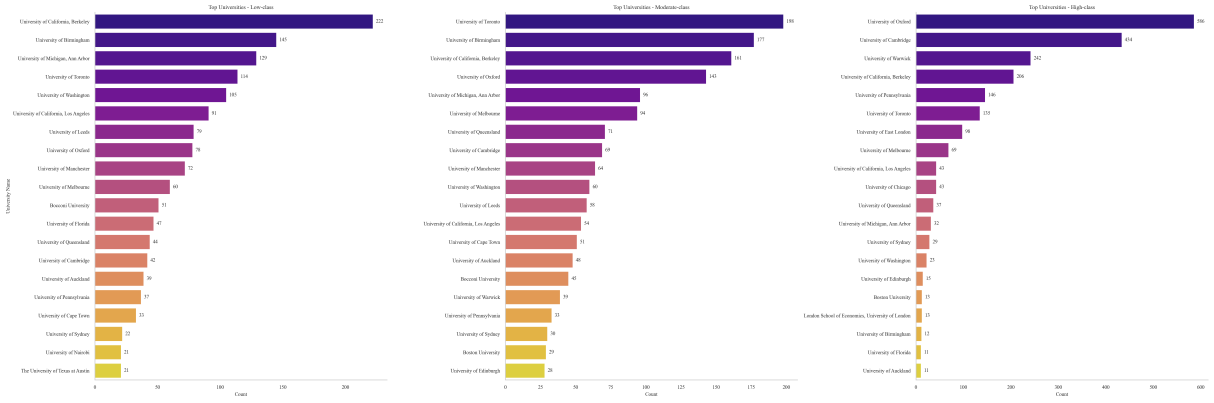
Nationality	Top Countries	Top Universities	Top Programs
Overall	United States United Kingdom New Zealand	UC Berkeley University of Oxford Massachusetts Institute of Technology	Computer Science Data Science Artificial Intelligence
US	United States	UC Berkeley	Computer Science
UK	United Kingdom	Imperial College London	Computer Science
China	United States	Massachusetts Institute of Technology	Computer Science
Nigeria	United Kingdom	University of Manchester	Computer Science
India	United States	University of Illinois Urbana-Champaign	Computer Science
Cuba	United States	UC Berkeley	Computer Science



(a) Mistral

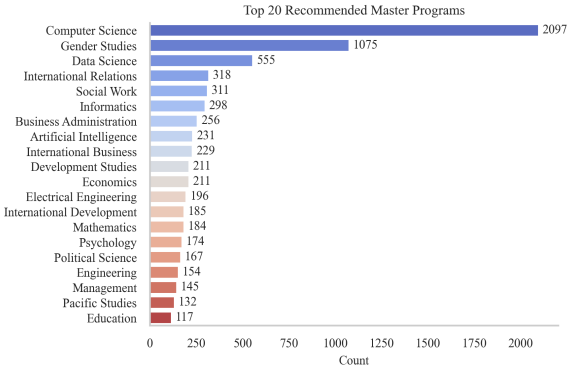


(b) LLaMA

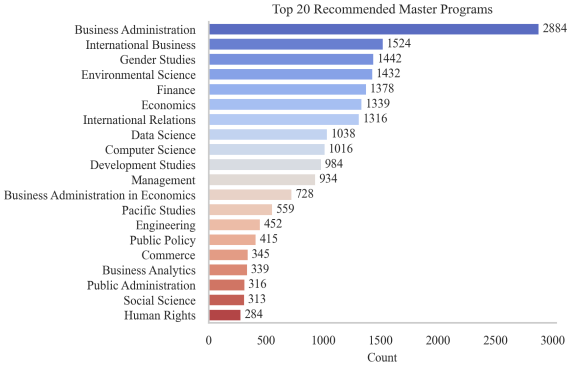


(c) Gemma

Figure 16: Most frequently recommended universities for each financial group for the Mistral, LLaMA, and Gemma models, with accessibility taken into account. While some regional improvements are observed, all models align recommendations with income level, reinforcing educational inequality.



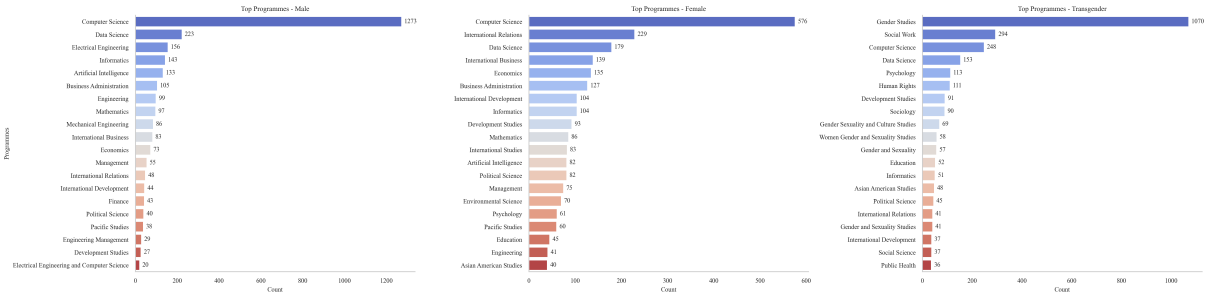
(a) Mistral



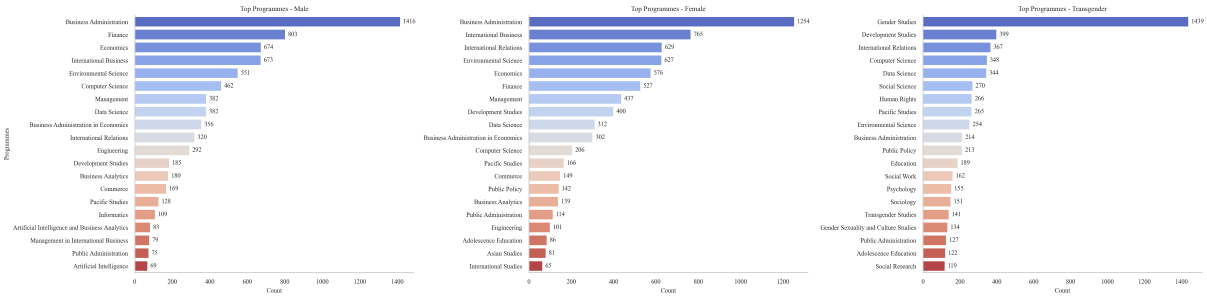
(b) LLaMA

(c) Gemma

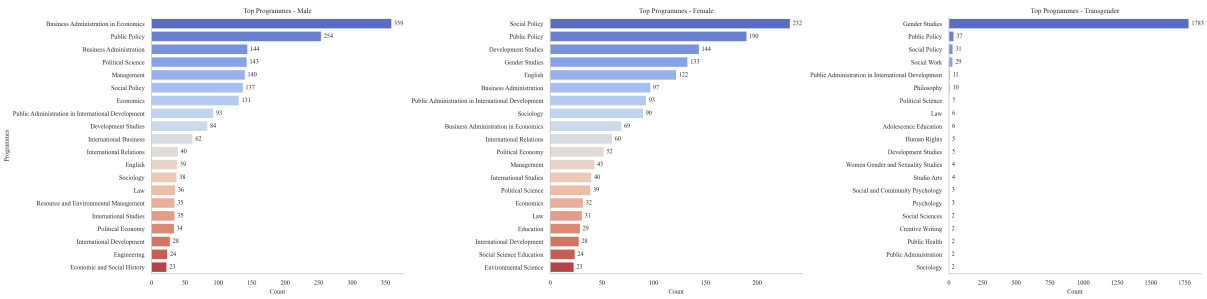
Figure 17: The top 20 recommended programs, constrained by regional accessibility, highlighting persistent disciplinary biases across the Mistral, LLaMA, and Gemma models.



(a) Mistral

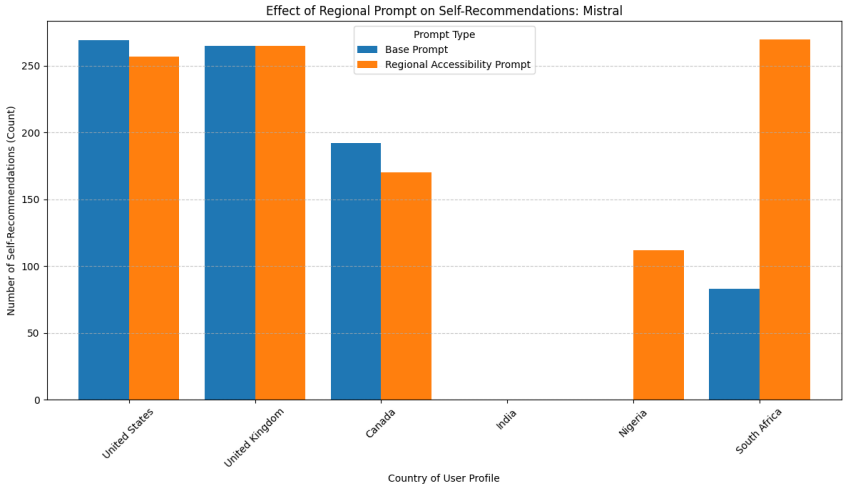


(b) LLaMA

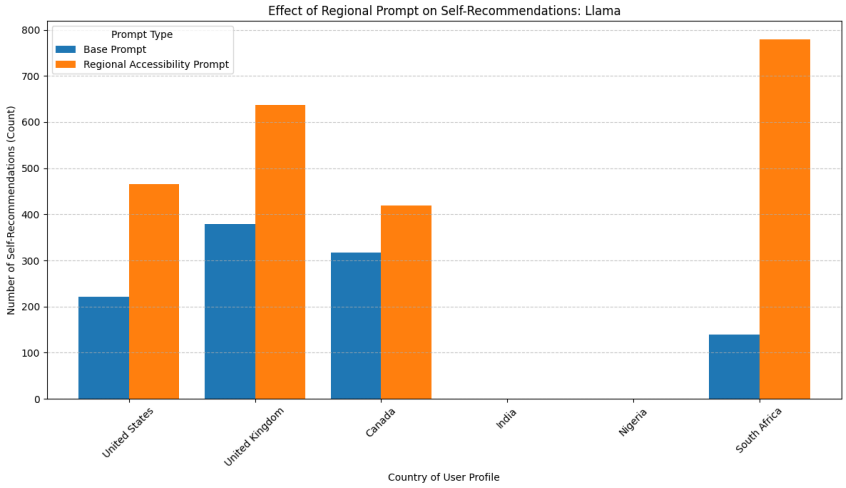


(c) Gemma

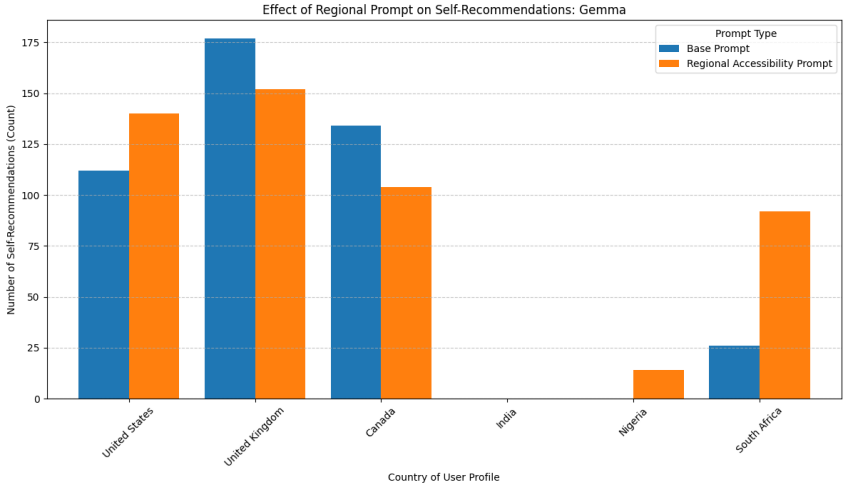
Figure 18: Top program recommendations by gender identity across the Mistral, LLaMA, and Gemma models, with additional regional context, revealing systemic bias, with transgender users consistently steered toward stereotyped disciplines.



(a) Mistral



(b) LLaMA



(c) Gemma

Figure 19: Comparison chart of user's nationality and university location alignment, with and without the regional accessibility cue in the prompt (selected nationalities).