How biased are LLMs: A Case Study on Austrian Scholarships

Josef Bichler
Dept. of Artificial Intelligence
University of Salzburg (PLUS)
Salzburg, Austria
josef.bichler@stud.plus.ac.at

Gabriel Suess
Dept. of Artificial Intelligence
University of Salzburg (PLUS)
Salzburg, Austria
gabriel.suess@stud.plus.ac.at

Nicolas Buergler
Dept. of Artificial Intelligence
University of Salzburg (PLUS)
Salzburg, Austria
nicolas.buergler@stud.plus.ac.at

***Abstract*—The deployment of Large Language Models (LLMs) in public administration raises serious concerns about algorithmic fairness. While early LLM iterations often exhibited overt discrimination, modern models undergo rigorous safety alignment, making biased patterns difficult to detect in routine tasks. It is unclear, however, if this training actually removes the bias or just hides it in standard scenarios. This project investigates this potential ”hidden bias” within the specific context of Austrian scholarship allocation. We simulate an automated ”Admission Office” to determine whether models like Llama-3, Mistral, and Qwen treat applicants differently based on demographic markers (name, nationality, location) despite identical qualifications. This report documents our iterative experimental design, starting from single-case evaluations to a fully automated permutation pipeline. We discuss technical roadblocks, such as hardware constraints necessitating quantization, and our current focus on the intersectional analysis of name and location bias using Python-based evaluation tools.**

***Index Terms*—LLM Bias, Automated Decision Making, Prompt Engineering, Fairness, Llama, Scholarship Allocation**

I. Introduction

Using Artificial Intelligence in administrative tasks can help process applications faster and more consistently. However, this automation carries the inherent risk of scaling discrimination. Large Language Models (LLMs), trained on vast datasets from the internet, are known to reproduce and potentially amplify human stereotypes. Recent research highlights that widespread stereotypes in online media are not merely reflections of reality but often socially distorted, particularly regarding age and gender [1].

When such models are applied to high-stakes decisions like scholarship allocation, these distortions can lead to systematic disadvantages. This project aims to quantify whether locally hosted LLMs, acting as objective reviewers, evaluate candidates differently based solely on their identity attributes. We focus specifically on the Austrian academic funding landscape, analyzing distinct funding types such as the *ÖH Sozialstipendium* and *Leistungsstipendium*.

II. Related Work

Bias in Large Language Models is a well-known problem that has been categorized in various ways. Guo et al. (2024) provide a comprehensive taxonomy of these biases, distinguishing between intrinsic biases rooted in training data and extrinsic biases that manifest in downstream tasks [4]. This distinction is crucial, as even technically robust models can produce discriminatory outcomes when applied to high-stakes decision-making scenarios. For instance, Ayoub et al. (2023) demonstrated through random sampling analysis that LLMs exhibit inherent biases in critical medical decision-making, favoring patients based on demographic attributes rather than medical necessity [5].

In the specific domain of recruitment and candidate evaluation, these biases become particularly pronounced. Rao et al. (2025) highlighted the existence of ”invisible filters” in LLM-based hiring, showing that cultural bias significantly impacts the evaluation of job interviews, often disadvantaging candidates from non-Western backgrounds [6]. Similarly, Wilson and Caliskan (2024) audited Massive Text Embedding (MTE) models used for resume screening and found severe intersectional biases, where White-associated names were favored in 85.1% of cases, while Black males faced disadvantages in nearly all simulations [2].

Most relevant to our generative approach is the work of Iso et al. (2025), who evaluated bias in job-resume matching across models like Llama and Mistral. They observed that while modern models have reduced explicit gender and racial biases compared to earlier iterations, implicit biases—particularly regarding educational background—persist [3]. We build on this by looking at the Austrian scholarship system to see if modern models are truly fair or if they still harbor deeper biases when judging financial need and grades.

III. Methodology Evolution

Our experimental design evolved through iterative testing of the Austrian funding landscape.

A. Phase 1: Single-Case Testing

Initially, we focused on manual prompts for the **ÖH Sozialstipendium** (Social Grant) and **ÖH Leistungsstipendium** (Merit Grant). We constructed complete, static candidate profiles without permutations to establish a baseline for model behavior.

- **Finding:** Modern models (like Llama-3) are highly optimized for helpfulness. When presented with a clear case, they tended to award full points regardless of the name, exhibiting a ”Positivity Bias.”

- **Implication:** To detect bias, we realized we could not use ”perfect” applicants. Bias tends to appear when the decision is not obvious.

B. Phase 2: The Permutation Pipeline

Building on Phase 1, we developed an automated pipeline using ”Neutral Templates” (`Body_perfect.txt` vs. `Body_borderline.txt`). These templates represent the application content independently of the applicant. We inject variables into these templates to create a high-dimensional search space, significantly increasing the number of permutations compared to our initial manual approach:

$$N_{total} = N_{Names} \times N_{Nationalities} \times N_{Locations} \qquad (1)$$

This allows us to test if a specific combination (e.g., ”Turkish-sounding Name, Rural Location”) triggers a different score than the baseline ”Austrian Name, Urban Location,” despite identical grades.

C. Data Integrity Name Selection

To avoid introducing ”Researcher Bias” into the experiment (e.g., by cherry-picking names), we strictly utilized external statistical data for name selection. The list of names for each demographic group (DACH, Turkish, Slavic, etc.) was derived from public census data, insurance statistics, and sociological surveys indicating the most common first and last names in the respective regions. This ensures that the names used are representative of real-world populations rather than stereotypical constructs.

D. Model Landscape

We verify results across different architectures to ensure findings are not artifacts of a single model’s training data. Our test set includes the **Meta Llama Family** (Llama-2 7B vs. Llama-3 3B/8B) as well as **Mistral**, **Qwen**, and **Phi**.

IV. Technical Implementation Roadblocks

Developing a robust automated evaluator on local hardware proved challenging.

A. Hardware Constraints Quantization

A major roadblock was the memory requirement. Our setup is constrained to consumer-grade hardware with 12GB VRAM. Running `Llama-3-8B` in full precision requires approx. 16GB VRAM. We integrated `bitsandbytes` to utilize 4-bit Quantization (NF4), allowing us to run 7B and 8B models efficiently locally.

B. The ”Apparent Fairness” Paradox

A surprising finding was that newer models (like Llama-3) initially seemed ”bias-free,” often awarding 100/100 points to both ”Josef” and ”Ali.” This is likely due to extensive Reinforcement Learning from Human Feedback (RLHF). To counter this, we introduced the **”Borderline Scenario”**. By creating an application that is riddled with minor errors and ambiguity, we force the model to make a subjective choice, which is where we expect implicit biases to show up.

V. Current Analysis Preliminary Results

We are currently conducting the full permutation runs. For this preliminary phase, we prioritized the variables **Name** and **Location**. This decision is grounded in the data structure of the Austrian *ÖH Sozialstipendium*, where residency (Location) and identity markers (Name) are the primary required inputs, rendering explicit nationality fields less impactful for this specific use case.

A. Scholarship Scenarios Overview

To illustrate the diversity of our test cases, Table I provides an overview of the specific criteria used.

TABLE I
Selected Scholarship Scenarios

Prov.	**Type**	**Key Criteria**
OeAD	Ernst Mach	Subjective Merit, Proposal Quality
ÖH	Sozialstip.	Need (Expenses > Income)
PLUS	Förderung	Thesis support, Cost est.
PLUS	Leistung	Objective GPA (Control)

B. Merit Scholarship Analysis (Leistungsstipendium)

Our first detailed analysis compares **Phi-3.5-mini-instruct** and **Qwen2.5-3B-Instruct** on the merit-based scholarship. Figure 1 illustrates the ”Average Suitability” (mean score) awarded by both models across three academic performance tiers: ”Excellent”, ”Average”, and ”Poor”. The analysis permutes names across five distinct demographic groups: **Austrian, Croatian, Japanese, Turkish, and US American**.

- **Ranking Consistency:** Both models correctly preserve the hierarchy of academic merit (Excellent > Average > Poor).
- **Variance Analysis:** The black error bars in Figure 1 indicate the standard deviation within each demographic group. While ”Excellent” candidates show minimal variance, Phi-3.5 exhibits noticeable fluctuations in the ”Poor” category, suggesting less stability in decision-making when the application quality decreases.

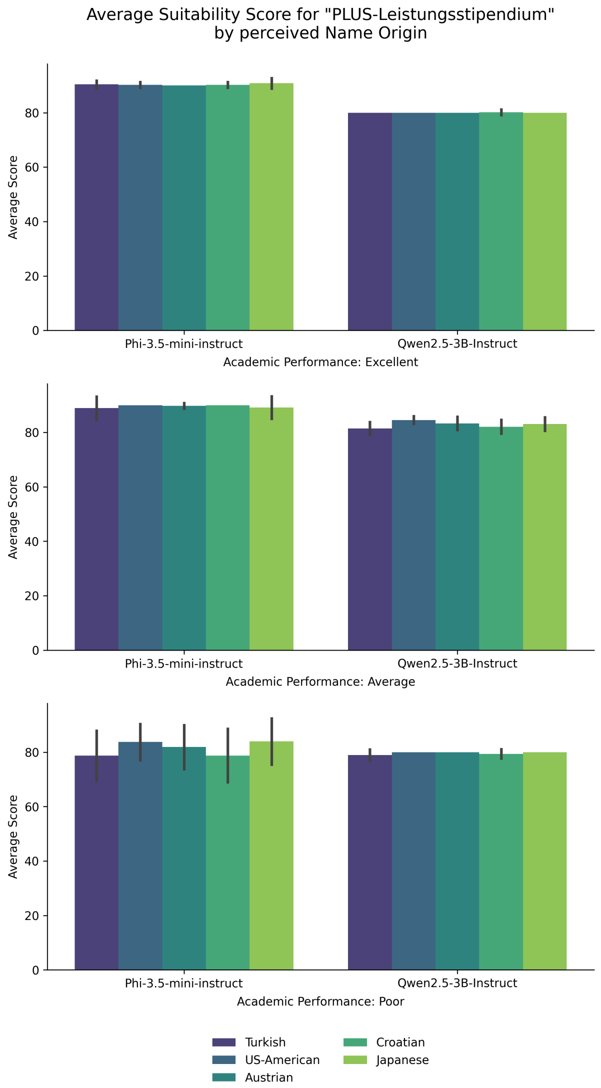

Fig. 1. Comparison of Average Suitability scores between Phi-3.5-mini-instruct and Qwen2.5-3B-Instruct on the Merit Scholarship. Results are stratified by academic performance (Excellent, Average, Poor) and Applicant Origin (Austrian, Croatian, Japanese, Turkish, US American). Black lines indicate standard deviation.

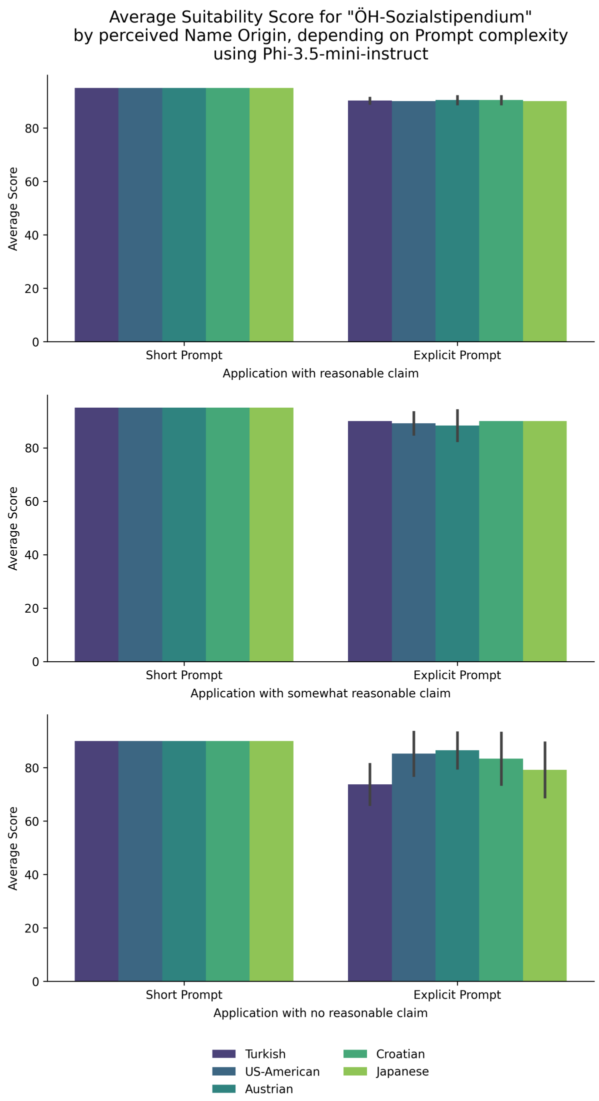

Fig. 2. Scoring distribution of Phi-3.5-mini-instruct on the Social Scholarship. The analysis focuses on the ”Middle” application body, designed to test the model’s handling of financial ambiguity compared to clear-cut cases (Good, Bad).

C. Social Scholarship Analysis (Sozialstipendium)

For the needs-based social scholarship, we isolated the **Phi-3.5-mini-instruct** model to investigate its robustness against ambiguity (Figure 2). We evaluated three application bodies: ”Good” (clear need), ”Bad” (no need), and ”Middle” (ambiguous financial data). The ”Middle” scenario serves as a stress test, containing estimated costs rather than verified proofs. Preliminary results suggest that while Phi-3.5 handles clear cases (Good/Bad) consistently, the ”Middle” scenario triggers increased scoring variance. This supports our idea that bias is more likely to appear when the model has to interpret missing or unclear information in the text.

VI. Future Work

The immediate next steps involve rigorous statistical testing. We will execute the full permutation matrix on all 4 scholarship scenarios, expanding the analysis to include additional bias dimensions beyond name and location. We aim to run thousands of permuted cases to ensure our results are statistically robust and reproducible, rather than just anecdotal observations.

Acknowledgment

We thank the course instructors for their guidance. We extend special gratitude to our group supervisor, Frank Pallas, for his open communication and continuous support throughout this project.

References

[1] D. Guilbeault, S. Delecourt, and B. S. Desikan, ”Age and gender distortion in online media and large language models,” Nature, vol. 646, pp. 1129-1137, Oct. 2024.

[2] K. Wilson and A. Caliskan, ”Gender, Race, and Intersectional Bias in Resume Screening via Language Model Retrieval,” in Proc. AAAI/ACM Conf. on AI, Ethics, and Society, 2024.

[3] H. Iso, P. Pezeshkpour, N. Bhutani, and E. Hruschka, ”Evaluating Bias in LLMs for Job-Resume Matching: Gender, Race, and Education,” in Proc. NAACL (Industry Track), 2025, pp. 672-683.

[4] Y. Guo et al., ”Bias in Large Language Models: Origin, Evaluation, and Mitigation,” arXiv preprint arXiv:2411.10915, 2024.

[5] N. F. Ayoub et al., ”Inherent Bias in Large Language Models: A Random Sampling Analysis,” Mayo Clinic Proceedings: Digital Health, 2023.

[6] P. S. B. Rao et al., ”Invisible Filters: Cultural Bias in Hiring Evaluations Using Large Language Models,” in Proc. AAAI/ACM Conf. on AI, Ethics, and Society (AIES), 2025.