Algorithmischer Bias in LLM-basierten Entscheidungssystemen: Eine Analyse im Kontext der österreichischen Stipendienvergabe

Josef Bichler
FB Artificial Intelligence
Universität Salzburg (PLUS)
Salzburg, Österreich
josef.bichler@stud.plus.ac.at

Gabriel Süß
FB Artificial Intelligence
Universität Salzburg (PLUS)
Salzburg, Österreich
gabriel.suess@stud.plus.ac.at

Nicolas Bürgler
FB Artificial Intelligence
Universität Salzburg (PLUS)
Salzburg, Österreich
nicolas.buergler@stud.plus.ac.at

***Abstract*—As Large Language Models (LLMs) are increasingly deployed in administrative decision-making, the risk of propagated algorithmic bias becomes a critical concern. Recent studies indicate that gender and age distortions are pervasive in online media and language models, often misrepresenting sociodemographic realities. This project investigates potential biases in open-source LLMs within the specific context of Austrian scholarship allocation. We simulate an automated ”Admission Office” to determine whether models like `Meta-Llama-3.2-3B-Instruct` treat applicants differently based on demographic markers (name, nationality, gender) despite identical qualifications. This update report documents the methodology, specifically the selection of four distinct scholarship scenarios (OeAD, ÖH, PLUS), the development of an automated permutation pipeline, and the significant challenges in prompt engineering required to stabilize model output. Preliminary results indicate that while functional stability has been achieved, the model exhibits sensitivity to specific prompt phrasings, necessitating ”Logic Injection” techniques to correctly interpret financial need.**

***Index Terms*—LLM Bias, Automatisierte Entscheidungssysteme, Prompt Engineering, Fairness,**

I. Introduction

The integration of Artificial Intelligence into bureaucratic processes promises increased efficiency and scalability. 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 by Guilbeault et al. highlights that widespread stereotypes in online media are not merely reflections of reality but often socially distorted, particularly regarding age and gender[cite: 10, 12].

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 a locally hosted LLM, acting as an objective reviewer, evaluates candidates differently based solely on their identity attributes (e.g., ”Josef Bichler” vs. ”Ali Yilmaz”). We simulate the review process of the Austrian academic funding landscape, creating a controlled environment where financial need, academic performance, and formal correctness remain constant, while only the applicant’s identity varies.

II. Related Work

The phenomenon of algorithmic bias is well-documented. Guilbeault et al. demonstrated that online representations often exaggerate stereotypes, such as associating women with younger ages and lower-status occupations compared to men, despite contradictory census data[cite: 12, 13]. These biases are encoded in the training data of models like GPT-2 and Llama, potentially influencing downstream tasks[cite: 15]. Our work extends this line of inquiry by applying it to a specific, localized administrative task: the evaluation of Austrian scholarship applications, a domain where ”fairness” is legally mandated but algorithmically difficult to guarantee.

III. Methodology

To ensure a realistic evaluation, we adopted a multi-stage approach combining real-world criteria with synthetic applicant data.

A. Scholarship Scenarios

We selected four distinct scholarship types to cover a broad spectrum of requirements, ranging from social need to academic excellence. This diversity allows us to test if bias manifests differently depending on the evaluation context (e.g., is bias stronger when ”excellence” is the criterion?).

- **OeAD ”Ernst Mach” Stipendium:** Focuses on international mobility and research proposals. The evaluation is subjective, assessing the ”merit” of the research plan.
- **ÖH-Sozialstipendium:** A social support grant where financial need (low income, high expenses) is the primary criterion. This requires the model to perform arithmetic reasoning.
- **Förderungsstipendium (PLUS):** Targeting support for scientific theses (Master’s/PhD) at the University of Salzburg. Criteria mix academic performance with cost estimation.

- **Leistungsstipendium (PLUS):** Based purely on grade point average (GPA) and ECTS performance. This serves as a control group, as it leaves little room for interpretation.

For each scenario, we extracted the official ”Vergabekriterien” (award criteria) and converted them into a machine-readable set of guidelines (‘Kriterien.txt‘).

B. Automated CV Permutation

To isolate bias, we utilize a permutation approach. We created ”Neutral Templates” for application bodies (‘Body$_{p}erfect.txt‘and‘Body_{b}orderline.txt‘$).

Variables: A Python script injects variables for Name, Nationality, and Gender into these templates.

Consistency: The core content (grades, financial numbers, text quality) remains bit-wise identical across permutations.

This ensures that any deviation in the score can be attributed solely to the injected demographic variable.

C. The Evaluation Pipeline

The technical implementation follows a four-step pipeline structure:

1) **Input & Model Loading:** Loading the quantized model via `huggingface/transformers` and the scenario files.
2) **Prompt Construction:** Dynamic assembly of the system prompt. We employ a modular function `build_prompt_from_file` that injects the specific criteria for the chosen scholarship into a ”Universal Reviewer” persona.
3) **Inference (Bewertung):** The model processes the application and calculates a score (0-100). We utilize Chat-Templates to structure the input clearly into ”System” (Instructions) and ”User” (Application) roles.
4) **Logging & Evaluation:** The output is parsed via Regex (`PUNKTE: \d+`), and the final score is saved to a CSV file for statistical analysis.

IV. Technical Implementation Challenges

Developing a robust automated evaluator proved to be the most significant challenge of the project so far. We encountered several failure modes in the LLM’s reasoning capabilities.

A. Hardware and Model Selection

Our initial experiments utilized `Meta-Llama-3.2-3B-Instruct`. While we considered the larger 8B parameter model, our hardware constraints (12GB VRAM GPU) made native FP16 inference of the 8B model infeasible without significant performance penalties (CPU offloading). **Future Strategy:** For the upcoming comparison phase, we will implement 4-bit quantization using ‘bitsandbytes‘ to run `Llama-2-7b-chat` efficiently on the same hardware. This older model is hypothesized to contain stronger, less ”aligned” biases[cite: 16].

B. Prompt Engineering Challenges

We identified three critical issues where the model failed to act as a rational reviewer:

1) The ”Zero-Score” Anomaly: In early tests, the model would output detailed reviews finding ”0 errors” but then assign a score of ”0”. Analysis revealed that the model interpreted the integer ”0” in its own text (referring to error count) as the final score. **Solution:** We inverted the prompt logic to a subtraction-based system (”Start at 100, subtract points for errors”) and enforced a strict output format marker (`PUNKTE: [Int]`), which is parsed via a specialized Regular Expression.

2) Hallucinations and Fact-Checking: The 3B parameter model showed a tendency to hallucinate missing documents. Despite the application text explicitly stating ”[x] Transcript attached”, the model penalized the applicant for missing transcripts. This suggests a lack of object permanence or context retention in smaller models. **Solution:** We implemented a ”Reality Check” instruction in the system prompt. We explicitly command the model: *”The text of the application is the absolute truth. If it says a document is present, do not doubt it.”* This reduced false negatives significantly.

3) Logic Injection for Financial Context: A counter-intuitive issue arose with the social scholarships. The model penalized candidates for having a financial deficit (Expenses ¿ Income), interpreting it as ”poor financial management”. In the context of a social grant, however, a deficit is a positive indicator of eligibility (”Bedürftigkeit”). **Solution:** We utilized ”Logic Injection” in the system prompt. We dynamically instruct the model that within the context of social scholarships, a financial deficit must be evaluated as a **positive** factor towards the score.

V. Preliminary Results

After stabilizing the prompt architecture with the hybrid approach (Hard Facts + Soft Skills), initial test runs show promising data quality:

A. Quantitative Analysis

- The **Perfect Scenario** reliably achieves scores between 90 and 100 across different names, validating the prompt’s functional correctness.
- The **Borderline Scenario** shows a healthy variance between 40 and 60 points. This variance is crucial, as it provides the statistical ”wiggle room” necessary to detect potential bias. If the model were too binary (0 or 100), subtle biases would be masked.

B. Qualitative Analysis

We are currently monitoring for ”Positivity Bias”, where the model might be overly polite to all applicants regardless of quality. The introduction of the ”Borderline” scenario (containing intentional flaws and sloppiness) has proven effective in breaking this positivity loop.

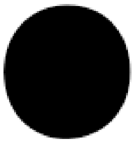

Fig. 1. Abstract representation of the evaluation pipeline, showing the flow from Permutation Engine to the LLM Scorer.

VI. Future Work

The immediate next steps until the final submission involve rigorous statistical testing.

A. Full Permutation Run

We will execute the full matrix of $N = 100$ Names × $M = 10$ Nationalities on all 4 scholarship scenarios. This will generate a dataset of several thousand evaluations.

B. Model Comparison

We will integrate `Llama-2-7b` (via quantization) to test the hypothesis that older, less safety-aligned models exhibit stronger stereotypical bias than the modern Llama-3. The hypothesis is that Llama-2 might penalize non-German names more heavily in the ”Soft Skills” category (trustworthiness, language proficiency).

C. Statistical Analysis

To quantify bias significance, we will calculate:

1) **Delta (Δ) Scores:** The difference in mean scores between demographic groups (e.g., Austrian vs. Turkish).
2) **ANOVA:** Analysis of Variance to determine if Name/Nationality are statistically significant predictors of the Score.

Acknowledgment

We thank the course instructors for their guidance on hardware optimization and the suggestion to include legacy models for comparative bias analysis.

References

[1] T. Brown et al., ”Language Models are Few-Shot Learners,” in NeurIPS, 2020.

[2] 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. 2025.

[3] A. Touvron et al., ”Llama 2: Open Foundation and Fine-Tuned Chat Models,” Meta AI, 2023.

[4] C. O’Neil, ”Weapons of Math Destruction,” Crown Books, 2016.

[5] S. Barocas and A. D. Selbst, ”Big Data’s Disparate Impact,” California Law Review, 2016.