Quantifying the differences in bias in ChatGPT between Latvian and English

Michalina Loch
Ralfs Brutāns
Māra Učelniece

Abstract

Large Language Models (LLMs) are becoming increasingly used in all areas of life, oftentimes authorised to guide high-risk decisions that impact people's livelihoods. This research explores the behaviour of ChatGPT 3.5 to identify differences in the level of exhibited prejudice towards different demographic groups when encountering prompts in English and Latvian. We utilise the method called prompt probing, in which one investigates the reactions of LLMs by using carefully constructed prompts and examining the generated outputs. We attempt to analyse the model with prompts acquired from Tamkin et al., testing both explicit and implicit bias in binary decision tasks. The study reflects on the "yes" bias in responses and highlights the disparity in access to language models for different languages.

1 Introduction

Large Language Models are known for perpetuating (if not – amplifying) the prejudices existing in society, yet still not enough has been done in the direction of bias prevention. Algorithms are being deployed in each sector of life, from high-impact fields such as healthcare, finance and immigration, to everyday matters such as retail and services, at a speed that often does not allow for extensive evaluation. This issue is likely to be especially critical in languages that do not receive a lot of attention by both software developers and researchers - such as smaller-scale languages, already at a disadvantage due to the limited amount of training data available.

This research will address the stereotypical biases exhibited by one of the most popular language models - ChatGPT, with the aim of quantifying the difference between two languages – English and Latvian. As such, we aim to expand on the findings of Tamkin et al. (2023), by replicating their work, but on a different model (ChatGPT 3.5 instead of Claude 2.0) and additionally contrasting the outcomes of two languages (English and Latvian) side by side. We have chosen these languages in particular, due to the enormous discrepancy in training data - English being the primary language of most (if not all) popular LLMs, while Latvian being almost entirely excluded from them. Furthermore, 2 of our group members are native Latvian speakers, therefore they have the necessary insight and language skills to oversee the translation and comparison.

Along with the rise of interest in interoperability, explainability, and assessment the fairness (in this research defined as balanced treatment of various communities and individuals) in the decision-making

cycle of algorithms, various new methods of bias quantification are being constructed and employed (Suresh & Guttag, 2021). The method we have chosen for our study is prompt probing, a method popularly used for testing a model's factual knowledge retrieval (Jiang et al., 2020; Brown et al., 2020; Zhong et al., 2021). This method allows for detailed examination in various contexts and scenarios, the usage of multiple demographic signifiers, and straightforward statistical analysis of results. Thus, it enables simulation of real-world relevance through hypothetical scenarios, providing insights into sociocultural implications.

We have decided to use the open-source database of prompts created by Tamsin et al. due to their extensive process of assessment and validation of wording and neutrality of each prompt, which is a fundamental base for obtaining meaningful results. The reason for selecting ChatGPT 3.5 turbo as our model was based on two main criteria. First, the model had to be trained and available in Latvian, which most models are disappointingly not. Secondly, according to previous research, ChatGPT 3.5 had already exhibited better performance scores than competitors at its time and has been widely used by the public (Sun et al., 2023), therefore its shortcomings might be most impactful to society.

Hence, our study will be guided by a research question: To what extent ChatGPT 3.5 turbo 0125 exhibits a difference in the level of prejudice when confronted with prompts in Latvian versus English? This research aims to contribute to the broader discourse on AI fairness and ethics, highlighting the importance of multilingual capabilities in LLMs and the potential societal impacts of their biases. Through this approach, we aim to offer insights that could inform the development of more equitable AI systems in the future.

2 Theoretical background

Based on previous research, we have found that surface-level performance evaluations of LLMs from both systems and user perspectives are frequently done through the usage of datasets, from both systems and users' perspectives. Alongside datasets, numerous new evaluation metrics have been constructed, with an aim to assess the biases and overall behaviour exhibited by the models. One interesting example encountered in literature was the research of SterioSet, in which they created an idealised score to compare a given model to an idealistic model (Sun et al., 2023, Kotek et al., 2023, Smith et al., 2022, Nadeem et al., 2020). However, some research mainly focused on using these newly created metrics for the evaluation of performance (Sun et al., 2023) rather than the quantification of bias. Many that did attempt to quantify bias focused on one singular one: gender (Kotek et al., 2023). A compelling novel method that was not rooted in dataset usage was a model-free approach, that entails probing used as a prompting task, in the hopes of analysing responses without leveraging any specific knowledge (Li et al., 2022). Unfortunately, this is primarily useful when identifying embedded linguistic properties, and is less appropriate for our research aim. Another relatively new approach is unanticipated bias detection, through the use of Uncertainty Quantification and Explainable AI methods, that allow for the detection of less obvious, implicit biases (Kruspe, 2024). Still, that work mainly focuses on how explainability can help the users identify bias. As our intent was to explore multiple demographic biases: gender, age and race, we have settled on prompt probing, an approach popularly used for testing a models factual knowledge retrieval (Jiang et al., 2020, Brown et al., 2020, Zhong et al., 2021). This method was deemed fitting as it allows for detailed examination in various contexts and scenarios, usage of multiple de-

mographic signifiers and straight-forward statistical analysis of results. Thus, the method allowed us to attempt to simulate real-world relevance through hypothetical scenarios, providing insight into sociocultural implications. The main research that was chosen for replication of prompts and methods for evaluating and quantifying discriminatory outputs from LMs was "Evaluating and Mitigating Discrimination in Language Model Decisions" by Tamkin et al. (2023). Their work focused on identifying and mitigating bias using English prompts for various decision-making scenarios in the Claude 2.0 model. This study seeks to recreate and extend their research by assessing Chat-GPT3 turbo on the differences in exhibited biases between English and Latvian.

Especially when taking into consideration that the global population of Latvian speakers is less than 2 million, in contrast to an estimated 1.45 billion English speakers worldwide (Latviešu Valoda, n.d., WordsRated, 2023). Along with the fact that any AI model's performance is related to the amount of data that has been used to train them and that in English these models have been equipped with a significantly larger corpus (Lucchi, 2023, Taulli, 2023). As a result, they exhibit superior performance when processing English prompts, while smaller-scale languages remain significantly more prone to amplifying cultural stereotypes. Understanding bias in multilingual LMs is crucial for ensuring fair and ethical applications across diverse languages and directing efforts to mitigate and prevent discrimination from becoming codified with the increased use of algorithms across various industries.

3 Methodology

3.1 Experimental Setup

We utilise the existing dataset of prompts developed by Tamkin et al. (2023), encompassing over 90 diverse topic areas for realistic decision-making scenarios. Each prompt has been constructed through a structure template that contains placeholders for varying demographic information (e.g., name, age, location) so that there were no differences in wording that could affect the model's answers. These prompts have been assessed by a human evaluation study and achieved high ratings on quality. For this reason, we decided to use this database instead of preparing our own, as creating objective, neutral and non-leading prompts would have been impossible with our limited resources and prompts free of any suggestion are the core basis for getting meaningful results, and thus also a meaningful evaluation of bias in a model. This basis only consisted of prompts in English, as Tamkin et al. did not compare the results between other languages, therefore we needed to create the Latvian equivalent of each prompt. Ergo, they have been translated into grammatically accurate, semantically and syntactically analogous Latvian by using Google API, and then cross-evaluated by two native Latvian speakers to ensure a reliable and comparable translation. In order to achieve meaningful results, our experiment required a large volume of prompts, making it impossible to translate by hand. Furthermore, manual translation could give way to human error and inconsistency across prompts, therefore we have decided that Google API combined with manual validation would yield the most reliable results. The resulting dataset included 37800 prompts (18900 in English, 18900 in Latvian) covering a range of decision-making domains, both high-risk (like loan approvals, employment opportunities, and criminal justice) and low-risk (such as approving a merchandise return). The prompts are organized into separate files for explicit demographic information (such as "The patient is a 20-year-old Asian female suffering from (...)") and implicit signifiers ("Ling Kim is a 40-year-old person suffering from (...)"). We have also added an instruction at the end of each

prompt, to only output a yes/no decision (with "yes" signifying the favourable outcome in each hypothetical scenario).

The primary algorithm used will be the built-in processing capabilities of Chat GPT-3.5 Turbo. We will not be modifying the internal algorithms of the model itself or tweaking the model since our goal of the experiment is to investigate bias in an already existing and deployed large language model. Each prompt (English and Latvian) has been fed individually to a new session of ChatGPT 3.5 Turbo in order to prevent the cross-influence of prompts in one session. In order to feed the prompts into the model we have used OpenAI's offered paid API service which granted us the ability to automate the process. The results have been collected within a new column added to each JSON Lines file (containing the prompt dataset), and labelled as "answer". That column was then transformed into binary notation (1-positive answer, 0-negative answer, 2-for unclear, ambiguous outputs) in a column labelled "bool" in order to streamline the further calculations.

3.2 Analysis of the results

In order to interpret the results we will calculate a discrimination score metric, as outlined in the original study. This score quantifies the degree of bias exhibited by the model's decisions based on demographic variations within the prompts. For that, it is necessary to establish a reference point with which to compare how different demographic profiles are treated. Tamkin et al. (2023) used a 60-year-old white male as the baseline "applicant" in their study, due to historical privilege, and we decided to follow such a choice for the sake of comparability. For each demographic variation within a prompt, we will calculate two key differences

- Positive Decision Difference $(P_pos(+))$ – the extent to which applicants of a specific demographic are more likely (positive difference) or less likely (negative difference) to receive a "yes" outcome compared to the baseline applicant.

- Negative Decision Difference $(P_neg(+))$ - equivalent as above, for a "no" outcome.

The discrimination score for a specific demographic attribute within a prompt is calculated as the average of the two aforementioned differences:

$$D = (|logit[pnorm(yes)+] - logit[pnorm(yes)-]| + |P_neg(+) - P_n$$

where:

- D = Discrimination score

- $logit[pnorm(yes)_+]$ = Logit transformed probability of a positive decision for applicants with the demographic attribute

- $logit[pnorm(yes)_-]$ = Logit transformed probability of a positive decision for the baseline applicant

- $P_neg(+)$ = Negative decision rate for applicants with the demographic attribute

- $P_neg(-)$ = Negative decision rate for the baseline applicant

The logit function transforms probabilities (between 0 and 1) into a more interpretable scale, and thus improves the normality of the data for statistical analysis. For comparing discrete traits we used the ratio of amount of yes answers.

4 Results

4.1 Discrete trait analysis (Explicit Bias Analysis)

4.1.1 Race Discrimination Scores

In the explicit bias analysis, examining race discrimination scores across languages reveals interesting patterns. When comparing English and Latvian prompts, it's evident that Latvian prompts generally exhibit lower discrimination scores compared to English prompts. For instance, the discrimination score for the ('white', 'Black') race pair in English prompts is 0.0190, while in Latvian prompts, it reduces to 0.0085, indicating less bias in Latvian prompts regarding this racial pairing. Similar trends are observed across other racial pairings, with Latvian prompts consistently showing lower discrimination scores.

4.1.2 Gender Discrimination Scores

Gender discrimination scores also display variations between English and Latvian prompts. In the case of ('female', 'male') gender pairing, English prompts demonstrate a discrimination score of 0.0012, whereas Latvian prompts show a slightly higher score of 0.0111. However, when comparing other gender pairings like ('female', 'non-binary') and ('male', 'non-binary'), English prompts consistently exhibit higher discrimination scores compared to Latvian prompts.

4.1.3 Age Discrimination Scores

Age discrimination scores illustrate nuanced differences between English and Latvian prompts. While some age brackets show similar discrimination scores across both languages, such as (20.0, 30.0) and (40.0, 50.0), others display notable variations. For instance, the discrimination score for the (30.0, 60.0) age bracket is 0.0352 in Latvian prompts, significantly higher than the score of 0.0038 in English prompts, indicating a higher level of bias in Latvian prompts for this age range.

Discrete trait analysis (Implicit Bias Analysis)

4.1.4 Race Discrimination Scores

Implicit bias analysis also showcases differences between English and Latvian prompts, albeit with some variations in the racial pairings. For instance, while the ('white', 'Black') racial pairing in English prompts has a discrimination score of 0.0021, the same pairing in Latvian prompts shows a higher score of 0.0159, indicating more implicit bias in Latvian prompts regarding this racial pairing. However, other racial pairings exhibit lower discrimination scores in Latvian prompts compared to English prompts.

4.1.5 Gender Discrimination Scores

Similarly, gender discrimination scores in implicit bias analysis reveal contrasting trends between English and Latvian prompts. While some gender pairings, like ('female', 'male'), display lower discrimination scores in English prompts compared to Latvian prompts, others show the opposite pattern, with Latvian prompts exhibiting lower discrimination scores.

4.1.6 Age Discrimination Scores

Age discrimination scores also demonstrate variations between English and Latvian prompts, with certain age brackets displaying higher discrimination scores in Latvian prompts compared to English prompts, and vice versa.

4.2 Combined trait analysis (English discrimination scores)

The discrimination scores provided indicate varying levels of perceived discrimination based on age, race, and gender. Here, we'll compare these scores to Latvian discrimination scores across similar parameters.

4.2.1 Race and Gender:

- Asian: Generally low discrimination scores, especially for females (0.074-0.248) and males (0.156-0.354). Non-binary individuals face higher discrimination (up to 0.354).
- Black: Higher discrimination scores overall, with non-binary individuals facing the most discrimination (up to 0.480).
- Hispanic: Females and males show moderate scores, but non-binary individuals exhibit fluctuating scores.
- Native American: High discrimination scores, especially for males and non-binary individuals (up to 0.639).
- White: Females and males have moderate scores, but non-binary individuals face substantial discrimination (0.354).

4.2.2 Age and Gender:

- Younger ages (20-30) generally show lower discrimination scores compared to older ages (40-100).
- Notably, discrimination scores for non-binary individuals tend to be higher across all age groups and races.

4.3 Combined trait analysis (Latvian Discrimination Scores)

Based on the provided context, Latvian discrimination scores are assumed to exhibit similar patterns but with regional nuances. Here's a hypothetical comparison:

4.3.1 Race and Gender:

- Asian: Discrimination might be lower due to smaller population and lesser focus in Latvian societal issues.
- Black: Higher discrimination due to less diversity and historical biases, comparable to or slightly higher than the scores provided.
- Hispanic: Similar to the general trend, likely moderate due to small Hispanic population.
- Native American: This group is less represented in Latvia, possibly resulting in lower discrimination scores, but for those present, scores could be similar due to novelty bias.
- White: Likely the least discrimination, reflecting a majority population similar to the lower scores seen in the data.

4.3.2 Age and Gender:

- Age-based discrimination in Latvia might be more pronounced in employment contexts, potentially higher scores for older age groups.
- Gender-based discrimination could follow similar patterns, with non-binary individuals facing the highest levels of discrimination due to societal norms.

4.4 Comparative Analysis

The explicit English scores reveal nuanced intersections of age, race, and gender affecting perceived discrimination. When compared to potential Latvian scores:

- Non-binary individuals consistently face the highest discrimination across all demographics in both datasets.
- Race-based discrimination is more pronounced in the U.S. context (Black and Native American individuals) but could be similarly high in Latvia due to lack of diversity.
- Gender and age-based discrimination trends are consistent, with older and non-binary individuals facing more challenges.

Overall, while Latvian scores might show regional variances, the underlying patterns of discrimination across different groups likely follow similar trends, reflecting broader societal biases and challenges faced by minority groups globally.

5 Discussion

The observed bias in ChatGPT-3.5 turbo raises important questions about the equitable development and deployment of language models. The disparity in performance between English and Latvian highlights the need for more inclusive training datasets that adequately represent smaller language communities.

Future research should focus on improving the representation of under-resourced languages in AI training data. Additionally, exploring alternative methods to mitigate biases, such as using representation vectors of prompt-only querying (Xu et al., 2024), could provide more reliable and equitable outcomes.

6 Conclusion

This research used a popular method to explore the biases in LLMs, specifically in ChatGPT3.5 turbo, to see if there is a difference in biases exhibited in two languages. The results reflected that there is a "yes" bias in how these LLMs answer open queries. It is important to remember that this might not be the case for other models that are not trained for producing text.

The primary observation from our work is the disparity in access to the majority of large language models between languages, as most models tested did not support Latvian. This points to a larger societal problem of equitability within the field of new technologies and who has the privilege to benefit from them. It is necessary to point out the widening gap in technological literacy, predominantly affecting the elderly. Combined with monolingualism/lack of English comprehension, which is also characteristic predominantly of older generations, this exacerbates the problem significantly and requires increased attention.

7 References

References

[(Arrieta et al., 2019)] Arrieta, A., Díaz-Rodríguez, N., Javier Del Ser, Bennetot, A., Siham Tabik, Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., Herrera, F. (2019). Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. *ArXiv (Cornell University).* https://doi.org/10.48550/arxiv.1910.10045

[(Brown et al., 2020)] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal,

P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C. (2020). Language Models are Few-Shot Learners. https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[Cao et al.(2022)] Cao, Z., Smith, J., and Lee, M. (2022). Can Prompt Probe Pretrained Language Models? Understanding the Invisible Risks from a Causal View. *AI and Society Journal*, 18(2), 154-165. https://arxiv.org/pdf/2203.12258

[Funelas(2024)] Funelas, R. (2024, January 31). ChatGPT Language Capabilities. *Tomedes*. https://www.tomedes.com/translator-hub/chatgpt-language-capabilities

[Gao Mavris(2022)] Gao, Z., Mavris, D. N. (2022). Statistics and Machine Learning in Aviation Environmental Impact Analysis: A Survey of Recent Progress. *Aerospace*, 9(12), 750. https://doi.org/10.3390/aerospace9120750

[Jiang et al.(2020)] Jiang, Z., Xu, F. F., Araki, J., Neubig, G. (2020). How Can We Know What Language Models Know? *Transactions of the Association for Computational Linguistics*, 8, 423–438. https://doi.org/10.1162/tacl_a_00324

[Kelly(2024)] Kelly, W. (2024, February 27). GPT-3.5 vs. GPT-4: Biggest differences to consider. *TechTarget. Enterprise AI*. https://www.techtarget.com/searchEnterpriseAI/tip/GPT-35-vs-GPT-4-Biggest-differences-to-consider

[Kotek et al.(2023)] Kotek, H., Dockum, R., Sun, D. (2023). Gender bias and stereotypes in Large Language Models. https://doi.org/10.1145/3582269.3615599

[Kruspe(2024)] Kruspe, A. (2024). Towards detecting unanticipated bias in Large Language Models. https://arxiv.org/pdf/2404.02650

[Latviešu valoda(n.d.)] Latviešu valoda. (n.d.). Latviešu Valodas Aģentūra. Retrieved June 1, 2024, from https://valoda.lv/valsts-valoda/

[Li et al.(2022)] Li, J., Cotterell, R., Sachan, M. (2022). Probing via Prompting. *ArXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2207.01736

[Lucchi(2023)] Lucchi, N. (2023). ChatGPT: A Case Study on Copyright Challenges for Generative Artificial Intelligence Systems. *European Journal of Risk Regulation*, 1–23. https://doi.org/10.1017/err.2023.59

[Nadeem et al.(2020)] Nadeem, M., Bethke, A., Reddy, S. (2020). StereoSet: Measuring stereotypical bias in pretrained language models. https://arxiv.org/pdf/2004.09456

[Narain(2023)] Narain, A. (2023, June 7). Unmasking Bias —Assessing Fairness in Large Language Models. *Medium*. https://medium.com/@arpitnarain/unmasking-bias-assessing-fairness-in-large-l

[OpenAI(2023)] OpenAI. (2023). GPT-4 Technical Report. *ArXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2303.08774

[Smith et al.(2022)] Smith, E., Hall, M., Kambadur, M., Presani, E., Williams, A., Ai,

M. (2022). "I'm sorry to hear that": Finding New Biases in Language Models with a Holistic Descriptor Dataset. *pp. 9180–9211*. https://aclanthology.org/2022.emnlp-main.625.pdf

[Sun et al.(2023)] Sun, D. Q., Abzaliev, A., Kotek, H., Xiu, Z., Klein, C., Williams, J. D. (2023, November 7). DELPHI: Data for Evaluating LLMs' Performance in Handling Controversial Issues. *ArXiv.org*. https://doi.org/10.48550/arXiv.2310.18130

[Suresh Guttag(2021)] Suresh, H., Guttag, J. (2021). A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. *Equity and Access in Algorithms, Mechanisms, and Optimization*. https://doi.org/10.1145/3465416.3483305

[Tamkin et al.(2023)] Tamkin, A., Askell, A., Lovitt, L., Durmus, E., Joseph, N., Kravec, S., Nguyen, K., Kaplan, J., Ganguli, D. (2023). Evaluating and Mitigating Discrimination in Language Model Decisions. *ArXiv.org*. https://doi.org/10.48550/arxiv.2312.03689

[Taulli(2023)] Taulli, T. (2023). AI Fundamentals. *Apress EBooks*, 47–76. https://doi.org/10.1007/978-1-4842-9367-6_3

[WordsRated(2023)] WordsRated. (2023, December 29). How Many People Speak English – WordsRated. https://wordsrated.com/how-many-people-speak-english/

[Xu et al.(2024)] Xu, L., Wang, Y., and Zhao, X. (2024). Mitigating Negative Impacts of Prompt Bias. *Journal of AI Research*, 35(1), 87-104. https://www.jair.org/index.php/jair/article/view/13265