

Building Job Seekers' Profiles: Can LLMs Level the Playing Field?

Susana Lavado^{*1}, Leid Zejnilovic¹

¹*Nova School of Business and Economics, Lisbon, Portugal*

Abstract

This study investigates the impact of language complexity on the performance of an NLP-based recommender system that assists job seekers in adding relevant occupation labels and skills to their profiles. The system, deployed by Job Market Finland (JMF), was evaluated to determine whether it biases its recommendations towards more complex language inputs, potentially disadvantaging users who employ simpler language. Additionally, the study explores the effectiveness of using large language models (LLMs) to enhance simpler descriptions and mitigate potential biases. By utilizing a stratified sample of occupations and crafting varied descriptions (original, simple, complex, and LLM-improved), we analyzed the system's recommendations against a ground truth. Results indicate that the system favored more complex language, improving occupation label suggestions (but not skill recommendations). This bias is not mitigated by the use of an LLM, suggesting potential unintended consequences for users who employ simpler language and highlighting the opacity in optimizing such systems.

Keywords

Job matching, Large language models, Natural language processing, Algorithmic bias, Human-machine interaction

1. Introduction

The usage of systems relying on natural language processing (NLP) techniques, especially those systems using large language models (LLMs), is swiftly increasing across diverse tasks [1]. However, due to their black-box nature [2], understanding how to maximize these models' utility poses considerable challenges to users [3]. The complexity of optimizing the performance of NLP systems has sparked discussions regarding the importance of developing domain-specific prompt engineering skills [4, 5, 6]. When opaque NLP systems are available to casual users without clear information on how to optimally use them, biased outputs can emerge [3].

As NLP-based recommender systems (including LLMs) are more often an external face of organizations interacting with citizens, decision-makers are confronted with a trade-off between scalability and efficiency of services on one side and potential biases on the other. The question of biases is complex, as they emerge from the interaction between the technological artifact (a model) and a human, and are context dependent. In this study, we explore a type of bias that may emerge due to the form (rather than the content) of the interaction between a model and a

EWF'24: European Workshop on Algorithmic Fairness, July 01–03, 2024, Mainz, Germany


*Corresponding author.

✉ susana.lavado@novasbe.pt (S. Lavado*); leid.zejnilovic@novasbe.pt (L. Zejnilovic)

🆔 0000-0002-1088-6357 (S. Lavado*); 0000-0002-4209-4637 (L. Zejnilovic)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

human. More precisely, we examined the potential bias present in an NLP-based recommender system designed to assist job seekers in adding skills and occupation labels to their profiles, which are subsequently used to match them to relevant job offers.

Humans may have different skills to craft the prompts with which they interact with the system and the system may hypothetically produce more relevant output to better crafted prompts. As the quality of the prompts is related to individual skills, or education, the system may introduce unintended bias that favor more skilled people. A way to address such bias may lie within the NLP tools themselves. Recent evidence suggests that LLMs might be more effective when used by individuals with lower skills, hinting at a potential for these systems to "level the playing field" [7]. We investigated whether feeding a simpler description to an LLM before submitting it to the recommender system would reduce unintended bias. Hence, we hypothesized that:

1. An NLP system that recommends skills based on candidates' job descriptions will more accurate when more complex vs. simpler language is used.
2. Improving of the candidates' profile using an LLM, before inputting it to recommender system would eliminate the complex language bias.

1.1. Context

This study investigated potential language bias of an NLP tool deployed at Job Market Finland (JMF), the front-page of the Finnish PES e-services¹. JMF offers job seekers the possibility of creating a profile to be shown to potential employers, which can then contact the job seekers. Besides adding a description of their job applicant profile in natural language, job seekers may select the relevant job occupation label and skills. These occupations labels and skills are defined the multilingual classification of European Skills, Competences, Qualifications and Occupations (ESCO)². However, this is arguably a daunting task, considering that ESCO contains more than 3 thousand occupations and 13 thousand skills. To assist job seekers, JMF offers a NLP tool, based on Word2Vec, that recommend job occupation labels and skills to be added to their profile. These job occupation labels and skills will then be used to match the job seeker to the available job opportunities, which are also tagged with occupation(s) label(s) and relevant skills.

2. Method

To create employees' profiles, we randomly selected 63 occupations (corresponding to 2% of the 3008 ESCO occupations), stratified by the highest hierarchy level of ESCO (the 10 job groups). Because ESCO descriptions are associated to occupations' labels and a skill set, using them as employees' profiles gave us a ground truth to which we could compare the performance of the system. From the ESCO occupations' descriptions, we removed the occupation tag and shifted the language from the third person singular to the first person singular (e.g., we replaced "Dietitians create dietary plans" by "I create dietary plans"). Then, we crafted two alternative descriptions by keeping the original content but either using simpler or more complex language

¹<https://tyomarkkinatori.fi/en>

²<https://esco.ec.europa.eu/en/about-esco/what-esco>

(i.e. using vocabulary and sentence structure that is easily understandable to a wide audience vs. using advanced vocabulary and nuanced sentence constructions). This process involved leveraging a language model (ChatGPT 3.5). To simplify the descriptions, we opened a new chat for each occupation, where we entered the following prompt "Can you please help me simplify some sentences?" and then proceeded to introduce each sentence of the occupation's description, one by one. We revised each of the simplified sentence to guarantee they kept the same content, and asked the LLM for further simplification if needed. For the more complex descriptions, we followed the same procedure, but replaced the prompt for "Can you please help me rewrite some sentences using more polished/complex language?" To assert the validity of our created materials, we computed embedding vectors of the descriptions using sentence-BERT, a Transformers model, and computed the Cosine similarity between the vectors. To verify whether LLM-improved descriptions would achieve better results than the original, we crafted an additional occupation description by inputting the simplified description into another LLM (Gemini), preceded by the following prompt: "I am creating my candidate profile in a website where employers can find my profile and contact me to offer me a job interview. I need help to polish my description. Can you help me improve my paragraph? Here is my original description:." We then verbatim copied the paragraph improved by Gemini, but removed any suggested occupation by Gemini by either deleting it or replacing it with the word "professional", like we had done in the remaining descriptions. The job descriptions used can be obtained from the authors upon request.

We then inserted the original, simple, complex, and LLM-improved descriptions into the JMF profile tool. The profile tool recommends seven occupation labels and 20 skills to be added to any description inputted by the user. We compared the tool's output with the occupation labels and skills associated with the ESCO occupation descriptions to compute the following dependent variables for each occupation:

1. *Occupation position*: The position in which the tool recommended the correct occupation label (8 if not recommended);
2. *Skills hit*: The number of matches between the set of skills suggested by the tool and by ESCO.

To assess the impact of the different descriptions on the performance of the JMF tool across our two dependent variables, we conducted a Repeated-measures ANOVA with pairwise comparisons with a Bonferroni correction. We considered a significance level of $\alpha = 0.05$. Figure 1 presents an overview of the research methodology.

3. Results

3.1. Similarity of the occupation descriptions

Table 1 presents a comparison of the average number of sentences, words and characters in the original, simple, complex and LLM-improved descriptions.

We computed the cosine similarity between embedding vectors of the different descriptions to validate that, despite using different words, the original, simple, and complex descriptions

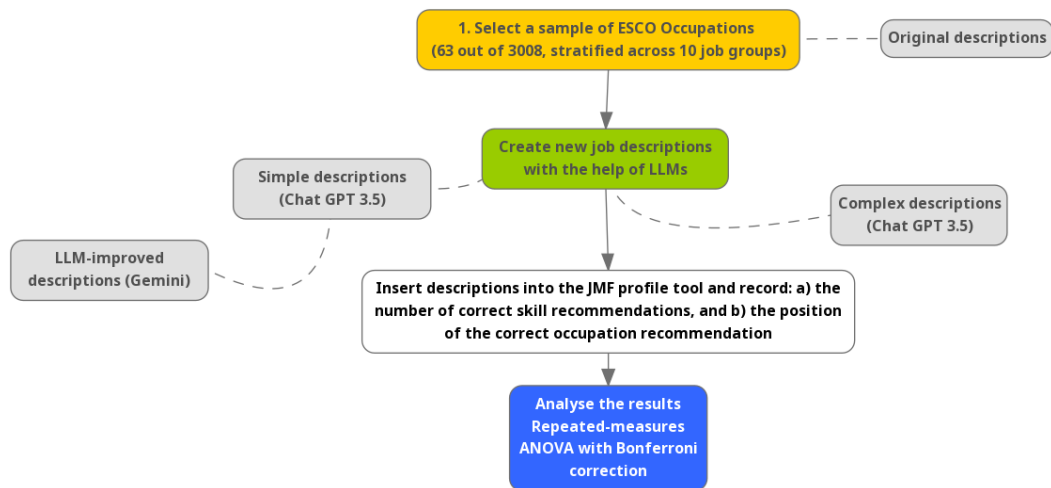


Figure 1: Overview of the research methodology.

Description	Average character count	Average word count	Average sentence count
Original	266	41	2
Simple	227	39	2
Complex	333	46	2
LLM-improved	358	48	3

Table 1

Average number of sentences, words and characters in the original, simple, complex and LLM-improved descriptions

maintained the same content. There were no differences between the similarity of the original and the simple descriptions ($M = 0.84$, $SD = 0.06$) and the similarity of the original and the complex descriptions ($M = 0.85$, $SD = 0.06$). The simple and complex descriptions had, on average, a distance of 0.75 ($SD = 0.08$). There were similar distances between the LLM-improved description and the original ($M = 0.71$, $SD = 0.10$), the complex ($M = 0.71$, $SD = 0.09$) and the simple ($M = 0.68$, $SD = 0.09$) descriptions.

3.2. Occupation position

The analysis revealed significant differences among the descriptions levels for the position the correct occupation label was suggested, $F(3, 186) = 25.26$, $p < 0.001$. Pairwise comparisons with a Bonferroni correction revealed that the description improved by the LLM performed the worst ($M = 4.52$, $SD = 2.82$), suggesting the correct occupation significantly later than the original description ($M = 1.79$, $SD = 1.76$, corrected- $p < 0.001$) and the complex description ($M = 2.70$, $SD = 2.28$, corrected- $p < 0.001$). No differences were found between the occupation position for the simple ($M = 3.49$, $SD = 2.51$) and the LLM-improved description (corrected- $p = 0.100$). The complex description performed better than the simple description (corrected- $p = 0.038$), but

worse than the original description ($p = 0.001$).

3.3. Skills hit

The pattern of the means for each of the levels generally mimic the results obtained for the occupation position variable: LLM-improved description: $M = 4.92$, $SD = 4.41$; Simple description: $M = 5.02$, $SD = 4.12$; Original description: $M = 5.84$, $SD = 4.25$; Complex description: $M = 5.67$, $SD = 4.24$. However, pairwise comparisons with a Bonferroni correction revealed no significant differences in the number of skills correctly recommended by JMF using the different description, $F(3, 186) = 2.30$, $p = 0.079$.

4. Discussion

Results partially supported hypothesis 1. The NLP system exhibited unintended biases that favored more complex language over simpler language in recommending occupation labels, but this bias did not extend to skills recommendations. However, we found no support for hypothesis 2. Results showed that improving the inputted description using an LLM did not improve the performance of the system. While better results may have been found if more time was invested in crafting the prompt asking the LLM to improve the description, we were interested in the behavior of an average user, who may stop interacting with the LLM after it provides a (seemingly) successful response [8].

This study has implications in the context of ensuring fair access to employment opportunities through digital platforms. Since the occupations and skills associated with a profile are crucial for matching job seekers to relevant job opportunities, biases in NLP-based recommender systems, specifically those favoring complex language, may inadvertently disadvantage job seekers who use simpler language. This unequal performance could lead to less relevant job matches for these individuals, potentially limiting their employment prospects. Additionally, while LLMs have been presented as tools to potentially level the playing field [7], our findings suggest that their use alone does not eliminate such biases. In fact, LLM-enhanced inputs did not improve the recommender system's performance, underscoring the need for more transparent and carefully designed interventions. These findings highlight the opacity surrounding the optimization of NLP-based systems like this one, which may pose challenges for users seeking to maximize their utility.

We acknowledge that our operationalization of language complexity in this study remains underdeveloped. This was an exploratory study, and as such, we did not adopt a formal framework for characterizing language complexity. Future work should address this limitation by providing a more standardized measure of language complexity.

The NLP-based recommender system analysed in this work was based on word2vec algorithm. Future research should investigate whether systems leveraging advanced LLMs exhibit similar biases, and examine the boundary conditions of these findings. Future research should also explore the real-world impact of using LLMs in hiring processes, specifically by examining whether LLM-enhanced descriptions result in increased engagement from potential employers.

In conclusion, this study highlights important considerations for policymakers and developers in designing AI-powered employment services that ensure all job seekers are equally empowered.

References

- [1] Y. Liu, T. Han, S. Ma, J. Zhang, Y. Yang, J. Tian, H. He, A. Li, M. He, Z. Liu, Z. Wu, L. Zhao, D. Zhu, X. Li, N. Qiang, D. Shen, T. Liu, B. Ge, Summary of ChatGPT-Related research and perspective towards the future of large language models, *Meta-Radiology* 1 (2023) 100017. doi:<https://doi.org/10.1016/j.metrad.2023.100017>.
- [2] Y. Wang, Deciphering the Enigma: A Deep Dive into Understanding and Interpreting LLM Outputs, *TechRxiv* (2023). doi:10.36227/techrxiv.24085833.v1.
- [3] J. Schneider, S. Haag, L. C. Kruse, Negotiating with llms: Prompt hacks, skill gaps, and reasoning deficits, *arXiv preprint arXiv:2312.03720* (2023).
- [4] T. F. Heston, C. Khun, Prompt Engineering in Medical Education, *International Medical Education* 2 (2023) 198–205. doi:10.3390/ime2030019.
- [5] L. Giray, Prompt Engineering with ChatGPT: A Guide for Academic Writers, *Annals of Biomedical Engineering* 51 (2023) 2629–2633. doi:10.1007/s10439-023-03272-4.
- [6] M. Wang, M. Wang, X. Xu, L. Yang, D. Cai, M. Yin, Unleashing chatgpt’s power: A case study on optimizing information retrieval in flipped classrooms via prompt engineering, *IEEE Transactions on Learning Technologies* 17 (2024) 629–641. doi:10.1109/TLT.2023.3324714.
- [7] S. Noy, W. Zhang, Experimental evidence on the productivity effects of generative artificial intelligence, *Science* 381 (2023) 187–192. doi:10.1126/science.adh2586.
- [8] J. Zamfirescu-Pereira, R. Y. Wong, B. Hartmann, Q. Yang, Why johnny can’t prompt: How non-ai experts try (and fail) to design llm prompts, in: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI ’23*, Association for Computing Machinery, New York, NY, USA, 2023. doi:10.1145/3544548.3581388.