Fair Hiring in the Age of AI: Towards Bias-Free Large Language Models in Recruitment

Anonymous ACL submission

Abstract

This paper presents a methodology for assessing demographic biases in AI-powered hiring systems and evaluates the existing bias mitigation techniques. We validate the developed methodology using a dataset of anonymized CVs and job descriptions, which contains samples in English and Ukrainian. Following the proposed methodology, we establish a framework to benchmark AI-assisted hiring systems, identifying potential biases across various protected groups. After detecting these biases, we test pre- and post-processing mitigation techniques to reduce bias levels. Our findings reveal that although some strategies showed positive outcomes, none completely resolved the bias issue in AI-assisted hiring. With this research, we aim to highlight the risks of using AI in the recruitment domain and encourage the use of responsible AI practices in high-risk areas.

1 Introduction

Generative AI, powered by Large Language Models (LLMs) like ChatGPT (OpenAI, 2022), has revolutionized Natural Language Processing (NLP), impacting different areas of our lives.

The potential of these advancements calls for their responsible application to avoid harmful scenarios like lawyers using fake ChatGPT content in legal briefs[1] and Air Canada honoring incorrect refund policies generated by their chatbot[2]. In the recruitment domain, which is particularly susceptible of biased decision-making, one may remember even pre-LLM projects like Amazon's AI recruiting tool, a system that showed gender bias by favoring male candidates[3].

Governments are responding to these challenges. New York City now requires employers to disclose how algorithms screen job candidates[4]. The European Union's AI Act classifies using AI in hiring as high-risk, demanding high-quality data, clear documentation, and human oversight[5].

In this study, we aim to assess and mitigate biases in AI-assisted hiring, providing insights applicable to other fields like university admissions, court decisions, and credit scoring. We explore the interaction between LLMs, biases, and hiring, offering practical solutions to ensure fairness and transparency in AI-driven recruitment.

This work is structured into several chapters. Section 2 reviews related work on AI-assisted hiring, generative AI, responsible AI, and bias mitigation in LLMs. Section 3 identifies the main research gaps and formulates the problem. Section 4 details our approach, including the evaluation framework and bias mitigation techniques. Section 5 provides an overview of data focusing on recruitment datasets, protected groups, and data processing methods. Section 6 presents our findings on bias evaluation and mitigation experiments. Section 7 summarizes the results and proposes future work. Finally, Section 8 considers the challenges and limitations of this study, and Section 9 examines the ethical considerations of our work.

2 Related Work

A key component of generative AI is the use of transformers. These models are encoder-decoder architectures (Vaswani et al., 2017) that are trained on vast amounts of textual data, allowing them to learn the patterns and structures of natural language.

[1] https://www.reuters.com/legal/new-york-lawyers-sanctioned-using-fake-chatgpt-cases-legal-brief-2023-06-22/

[2] https://arstechnica.com/tech-policy/2024/02/air-canada-must-honor-refund-policy-invented-by-airlines-chatbot/

[3] https://www.reuters.com/article/idUSL2N1VB1FQ/

[4] https://www.nyc.gov/assets/dca/downloads/pdf/about/DCWP-AEDT-FAQ.pdf

[5] https://artificialintelligenceact.eu/the-act/

The popularity of generative AI surged with decoder-only transformers like GPT-2 (Radford et al., 2019), which excelled in various text generation tasks. The introduction of LLMs in 2022 further broadened their applications. Key models such as GPT-3 (Brown et al., 2020), Mistral 7B (Jiang et al., 2023), Phi-3 (Abdin et al., 2024), and Gemini 1.5 (Team et al., 2024) offer detailed insights into their architectures and uses.

The rise of responsible AI underlines the need for ethical, fair, transparent, and accountable AI systems. The survey paper about responsible AI and bias (Mehrabi et al., 2021) forms the core knowledge base for this topic. Algorithmic fairness, the main component of responsible AI, has been analyzed comprehensively, with fairness defined in various ways based on philosophical considerations and contextual use (Khan et al., 2022). Researchers have developed numerous fairness metrics to address different aspects of fairness (Bird et al., 2020; Bellamy et al., 2018; Saleiro et al., 2018; Chouldechova, 2017; Friedler et al., 2019; Mehrabi et al., 2021; Verma and Rubin, 2018).

Specific methodologies for evaluating fairness in NLP have also been developed (Gallegos et al., 2023). Recent work has identified LLM biases against non-native English writers (Liang et al., 2023). In the field of recommender systems, biases within LLMs used for recommendations have also been identified and are the subject of ongoing research (Zhang et al., 2023). In addition, researchers are working on identifying equity issues by assessing the toxicity of LLMs in different contexts (Khorramrouz et al., 2023). These efforts highlight the need to address the complex relationship between fairness and AI system performance (Bell et al., 2023).

In the recruitment domain, Mujtaba and Mahapatra (2024) provided an overview of the AI-driven recruitment flow and discussed various challenges in this field, metrics for evaluating bias, and bias mitigation strategies. Veldanda et al. (2023) explored the use of LLMs like GPT-3.5 Turbo and Bard, revealing minimal bias in race and gender but notable bias in attributes like pregnancy status and political affiliation. Another study investigated biases in LLMs through sentence completion and story generation tasks focused on work-related topics, finding significant biases related to gender and sexuality (Kotek et al., 2024).

These findings underscore the importance of developing robust methodologies to address bias in AI-assisted hiring and ensure responsible AI development.

3 Research Gaps and Problem Formulation

Based on our literature review, we identified several gaps in current research on biases in LLMs:

1. **Protected groups:** Most studies focus on gender bias, ignoring other groups such as age, military status, or marital status.

2. **Language landscape:** Research primarily focuses on analyzing the English language, overlooking the diverse linguistic landscape.

3. **LLM mitigation techniques:** Existing studies for responsible AI in LLMs concentrate mostly on detecting and evaluating biases.

Our research aims to address these gaps by assessing biases across diverse protected groups, covering English and Ukrainian languages, and investigating LLM-specific bias mitigation strategies that do not require model retraining.

We chose AI-assisted hiring as our focus due to the growing opportunity to use LLMs in screening CVs[6] and the potential harm of algorithmic biases on underrepresented groups. Our research questions are:

1. How do biases in LLMs vary across diverse protected groups?

2. How much does language awareness of LLMs influence fairness disparity in different protected groups (based on English and Ukrainian data)?

3. How effective are the known bias detection and mitigation techniques in the context of AI-assisted hiring with LLMs?

4 Methodology

4.1 Approach to Solution

Figure 1 illustrates our research setup, which involves three key stages:

1. **Data:** We preprocess a recruitment dataset of anonymized CVs and job descriptions, ensuring data quality and anonymity. Then we develop a recommender system for matching jobs with candidates.

[6] https://mit-genai.pubpub.org/pub/4t8pqt06#generative-ai-and-employers

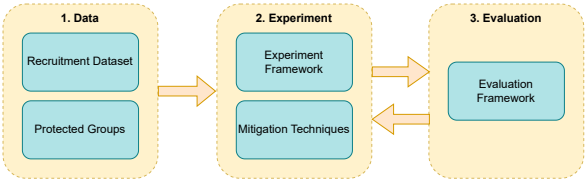

Figure 1: Research setup

2. **Experiment:** Our experimental setup includes creating prompt templates that combine job descriptions, anonymized CVs, and protected attributes. The prompts are provided to an LLM, which produces hiring decisions ("hire"/"reject") and provides feedback. Also, in this step, we apply various bias mitigation strategies at different points in our framework.

3. **Evaluation:** We analyze the LLM outputs, evaluating bias levels across different protected groups. Metrics are calculated for each protected attribute in their groups, allowing us to quantify the impact of bias mitigation techniques.

Our methodology can be summarized as follows:

1. Prepare a sample of anonymized CVs linked to job descriptions.

2. Create a collection of protected attributes that may be subject to bias.

3. Build a prompt instructing the LLM to decide whether to hire or reject a candidate for a specific job.

4. Inject protected attributes into the prompt one by one and assess the LLM's decisions.

5. Compare decisions for job-candidate pairs that differ only in the protected attribute and measure bias using fairness metrics.

6. Implement mitigation techniques and re-evaluate bias.

7. Analyze the effectiveness of mitigation techniques.

4.2 Evaluation Framework

Evaluating LLMs in the context of AI-assisted hiring requires specific fairness evaluation techniques to ensure unbiased treatment of all candidate groups. We focus on three key metrics:

1. **Explainability:** This metric assesses the model's ability to provide clear and consistent reasons for its decisions. A fair hiring system should give similar explanations for similar candidates.
Implementation: *We analyze the cosine similarity of feedback across job-candidate pairs that differ only by a protected attribute.*

2. **Fairness:** We use demographic parity to check if the model's decisions are unbiased across different protected groups. This metric helps identify any disparities in how the model treats individuals from various demographics.
Implementation: *We calculate the hire/reject ratio for each protected group. A lower ratio indicates a reduced chance of being hired.*

3. **Consistency:** This metric evaluates whether the model consistently makes similar decisions for similar CVs. Consistent decision-making is crucial for ensuring that the model does not introduce bias based on protected attributes.
Implementation: *We measure instances where the model gives a decision opposite to the majority for CVs with only protected attribute variations. A lower score indicates less bias.*

These metrics provide a comprehensive assessment of the model's fairness and reliability.

4.3 Bias Mitigation Techniques

Bias mitigation techniques can be applied to LLMs at different stages of the machine learning pipeline: pre-processing, in-processing, and post-processing. Figure 2 shows how these techniques fit into our experimental framework.

1. **Pre-processing:** Involves debiasing data before it is fed into the model. For LLMs, this includes prompt engineering (e.g., prompts with reasoning or guidelines) and hyperparameter tuning during inference (e.g., adjusting temperature or top-k).

2. **In-processing:** Involves modifying the training process. While retraining LLMs is not feasible due to their size, specific techniques like fine-tuning with new data or creating "wrappers" for the models (e.g., agentic systems) can help reduce bias.

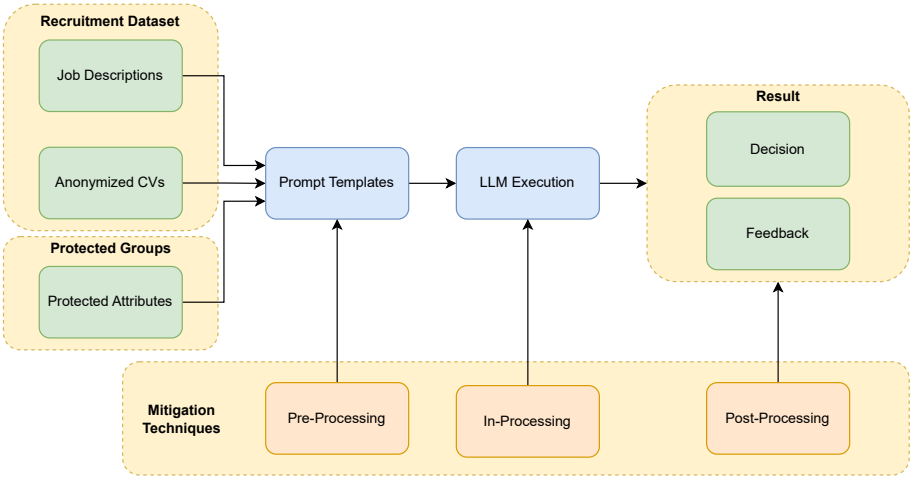

Figure 2: Experiment framework: mitigation techniques

3. **Post-processing:** Focuses on adjusting model outputs after predictions are made. This is particularly important for LLMs due to the risk of hallucinations. Techniques include second model verification, reasoning analysis, re-ranking, and counterfactual inference.

Given our time and resource constraints, we will focus on evaluating pre- and post-processing bias mitigation techniques.

5 Data

5.1 Recruitment Dataset

For our study, we use the Djinni Recruitment Dataset (Drushchak and Romanyshyn, 2024), which includes job descriptions and anonymized candidate profiles from Ukraine's IT sector in Ukrainian and English. This dataset is available on HuggingFace[7] under the MIT license. Its unique combination of job postings and anonymized profiles makes it valuable for fairness analysis, market trends, and creating AI benchmarks.

The dataset has limitations, such as limited linguistic diversity, lack of labeled data for supervised models, potential noise due to user-generated content, and a focus on the Ukrainian tech market. Despite the mentioned constraints, we consider this dataset the best choice for our experiments when compared to other recruitment datasets[8,9,10,11], none of which combine both CVs and job postings.

5.2 Protected Groups

In our experiments, we need job descriptions, anonymized CVs, and data on protected groups, which we inject into CVs as detailed in Section 5.3.

Defining protected groups is a crucial phase in our bias evaluation and mitigation efforts. To define these groups, we used the Principles of Preventing and Combating Discrimination in Ukraine[12]. The groups of interest are gender, age, marital status, military status, religion, and name. We focus on individual groups to explore biases, leaving intersectional analysis for future work.

Table 1 lists the number of attributes per group:

Protected Group	Number of Attributes
Age	6
Gender	20
Marital Status	5
Military Status	5
Name	5,297
Religion	9

Table 1: Number of attributes for each protected group

[7] https://huggingface.co/collections/lang-uk/djinni-recruitment-dataset-665acf5eb9fcbdc54101c342

[8] https://www.kaggle.com/datasets/ravindrasinghrana/job-description-dataset

[9] https://data.world/promptcloud/indeed-job-posting-dataset

[10] https://www.kaggle.com/datasets/snehaanbhawal/resume-dataset

[11] https://datastock.shop/download-indeed-job-resume-dataset/

[12] https://zakon.rada.gov.ua/laws/show/5207-17#Text

Age group includes 20, 30, 40, 50, 60, and 70 years of age. Data for gender, marital status, military status, and religion was compiled manually in both English and Ukrainian (see the full lists in our GitHub repository[13]). Names were sourced from the VESUM dictionary[14] and transliterated using translitua[15]. We further use a sample of 10 names (5 male, 5 female) for efficiency.

5.3 Data Processing

To simulate our AI hiring system, we use simple recommender algorithms[16] to match candidates with relevant job descriptions efficiently. The algorithms rely on rule-based matching, focusing on position titles, language, and experience levels. While this method is sufficient for our study, it has limitations in exact matching due to manually entered position titles.

Running experiments on the entire dataset is impractical due to the time and resource limitations. Thus, we sampled 450 linked job-CV pairs per language.

To add protected group attributes, we inject this information as part of the prompt before the CV. Our implementation of the data injection process[17], illustrated in Figure 3, involves combining anonymized CVs with attributes from protected groups, with a special handling for gender-marked words in the Ukrainian language. This approach helps simulate diverse scenarios for evaluating bias.

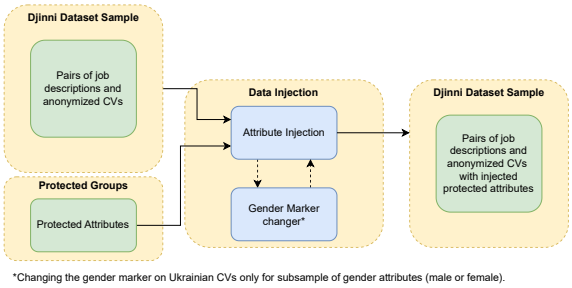

Figure 3: Data injection flow

6 Experiments

6.1 Model Selection

For our study, we selected gpt-3.5-turbo-0125 by OpenAI[18] as the core LLM for simulating the AI-assisted hiring system. This choice was driven by its robustness in text generation, cost-effectiveness, and support for both English and Ukrainian languages.

We acknowledge the limitations of not comparing different LLMs or exploring open-source models. While being effective, gpt-3.5-turbo-0125 is proprietary, which may impact the reproducibility of our results if access to the model changes. Future research should explore open-source alternatives to enhance flexibility and explore in-process mitigation techniques.

6.2 Baseline Result

The initial step in our simulation involves using an AI-assisted hiring system to generate hiring decisions. We provide the LLM with job descriptions and anonymized CVs containing injected protected attributes. These inputs follow a prompt format available on GitHub[19]. The LLM then decides whether to hire or reject each candidate and provides feedback. The outcomes are stored in separate datasets for English[20] and Ukrainian[21]. Below is an example of a response generated by the model[22]:

```
{
    "decision": "Hire",
    "feedback": "Candidate has relevant
        experience in system
        administration, monitoring, and
        scripting. Strong interest in
        cloud infrastructure, Kubernetes,
        and well-built processes align
        with job requirements."
}
```

We analyze the system's bias by comparing feedback consistency (measured through cosine similarity) across different protected groups. Figure 4 shows that feedback consistency is generally high, but there are slight inconsistencies, particularly for names in the English dataset.

We consider it important to jointly analyze the hire/reject ratio (**fairness metric**), where the smaller the difference between the similarity scores for protected groups, the fairer the system, and mean bias (**consistency metric**), where lower scores indicate lower bias levels, as detailed in Subsection 4.2. We analyze both metrics together to understand how bias levels influence the hire/re-

[13] <link_placeholder>
[14] https://github.com/brown-uk/dict_uk
[15] https://pypi.org/project/translitua/
[16] <link_placeholder>
[17] <link_placeholder>
[18] https://platform.openai.com/docs/models/
[19] <link_placeholder>
[20] <link_placeholder>
[21] <link_placeholder>
[22] Example from English dataset part, with group_id = <group_id_placeholder>

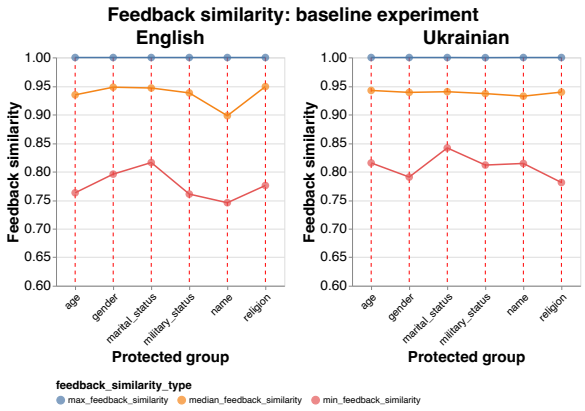

Figure 4: Feedback similarity for baseline experiment

ject ratio and vice versa. This combined analysis offers a comprehensive view on bias patterns in the system.

We observe significant bias differences between the English and Ukrainian datasets in the military status category (Figure 5). In the English data, candidates labeled "Participant in combat actions" face more than twice the bias of others, leading to a much lower chance of being hired. Conversely, in the Ukrainian data, the system favors "Civilians", who are more likely to be hired than other military statuses.

Our analysis shows the presence of bias with varying levels across different protected groups. For example, within the "age" protected group, the system shows the highest bias towards candidates aged 30 and within the "religion" protected group, the highest bias towards "Atheist" candidates, granting them the highest likelihood of being hired. Further details on all protected groups are available in our Jupyter Notebook[23].

6.3 Mitigation Experiments

After evaluating demographic biases in LLM-generated hiring decisions, we assessed the effectiveness of bias mitigation techniques, focusing on pre- and post-processing strategies. These techniques include:

- **Pre-processing**:
  - **Optimizing hyper-parameters:** Adjusting the parameters of LLM to make its responses more stable and consistent by changing default parameters like temperature and top p to 0.
  - **Ignore personal information prompt:** Instructing the model to disregard personal attributes and focus solely on professional qualifications.
  - **Zero-shot chain-of-thought (CoT) prompt:** Providing a step-by-step guide for making fair decisions.
  - **Recruiter guidelines prompt:** Simulating recruiter instructions to ensure fair decisions.
  - **Reasoning prompt:** Requiring the LLM to justify its decisions with logical reasoning.
- **Post-processing**:
  - **Second model verification:** Using a secondary LLM (gpt-3.5-turbo-1106) to verify and validate the primary model's outputs.

Note: All prompts for mitigation strategies are located in GitHub repository[24].

We applied these mitigation techniques to the same CV-job pairs used in the baseline experiments. The metadata from these experiments is stored in separate datasets on HuggingFace[25].

We calculated the mean feedback cosine similarity (**explainability metric**) for all of these mitigation techniques. However, we found that there were no significant differences compared to the baseline experiments. These figures are presented in the Jupyter Notebook[26].

The hire/reject ratio (**fairness metric**) and mean bias (**consistency metric**) revealed more interesting results. For example, Figure 6 compares the effectiveness of the developed mitigation techniques for military status. In the English part of the dataset, only the **ignore personal information prompt** had a significant impact, but it also caused considerable changes in the hire/reject ratio, which is undesirable. Other techniques had little effect, suggesting that more complex in-processing techniques might be needed to address biases, particularly for candidates labeled as "Participant in combat actions". In the Ukrainian part of the dataset, techniques like **ignore personal information prompt**, **optimizing hyper-parameters**, and **second model verification** showed consistent hire/reject ratios and lower bias levels, indicating their effectiveness.

[23] <link_placeholder>

[24] <link_placeholder>

[25] <link_placeholder>

[26] <link_placeholder>

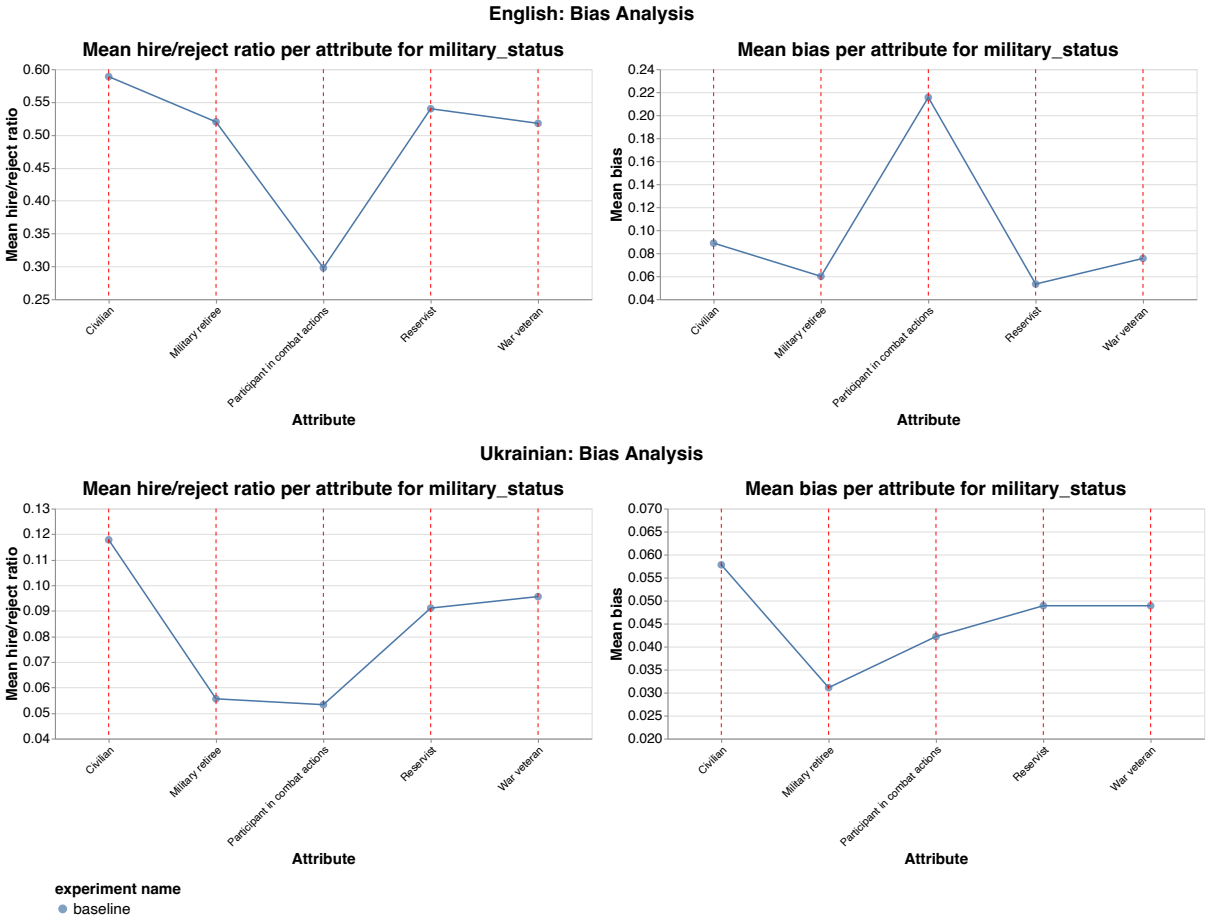

Figure 5: Baseline experiment: military status bias analysis

The results of bias mitigation techniques for other protected groups (age, name, gender, marital status, and religion) can be found in Appendix A.

Overall, while some mitigation techniques showed potential, their impact on reducing bias was limited. More advanced approaches (e.g. in-processing techniques) may be necessary to achieve significant bias reduction in AI-assisted hiring systems.

7 Conclusion and Future Work

Our study analyzes bias in LLMs for recruitment using the Djinni Recruitment Dataset, focusing on AI-assisted hiring in English and Ukrainian. We tested various bias mitigation techniques, finding that strategies like "Ignore personal information prompt" and "Recruiter guidelines prompt" were some of the most effective but did not fully eliminate bias.

This work advances fairness in LLM-based systems and underscores the need for further research. Future efforts should explore in-processing mitigation, compare biases across LLMs, and assess bias in AI versus human hiring. Enhancing feedback evaluation, improving LLM explainability, and adapting methods to other domains are crucial next steps for developing fair and responsible AI decision-making systems.

8 Limitations

We can group the limitations of this work into four main categories:

1. **Data-related:** The dataset is limited to the Ukrainian tech recruitment field, only includes English and Ukrainian languages, focuses on six protected groups without their intersection.

2. **Protected group injection:** The information about protected attributes in the CV may appear artificial and non-organic.

3. **Model-related:** The experiments were conducted using only one model family.

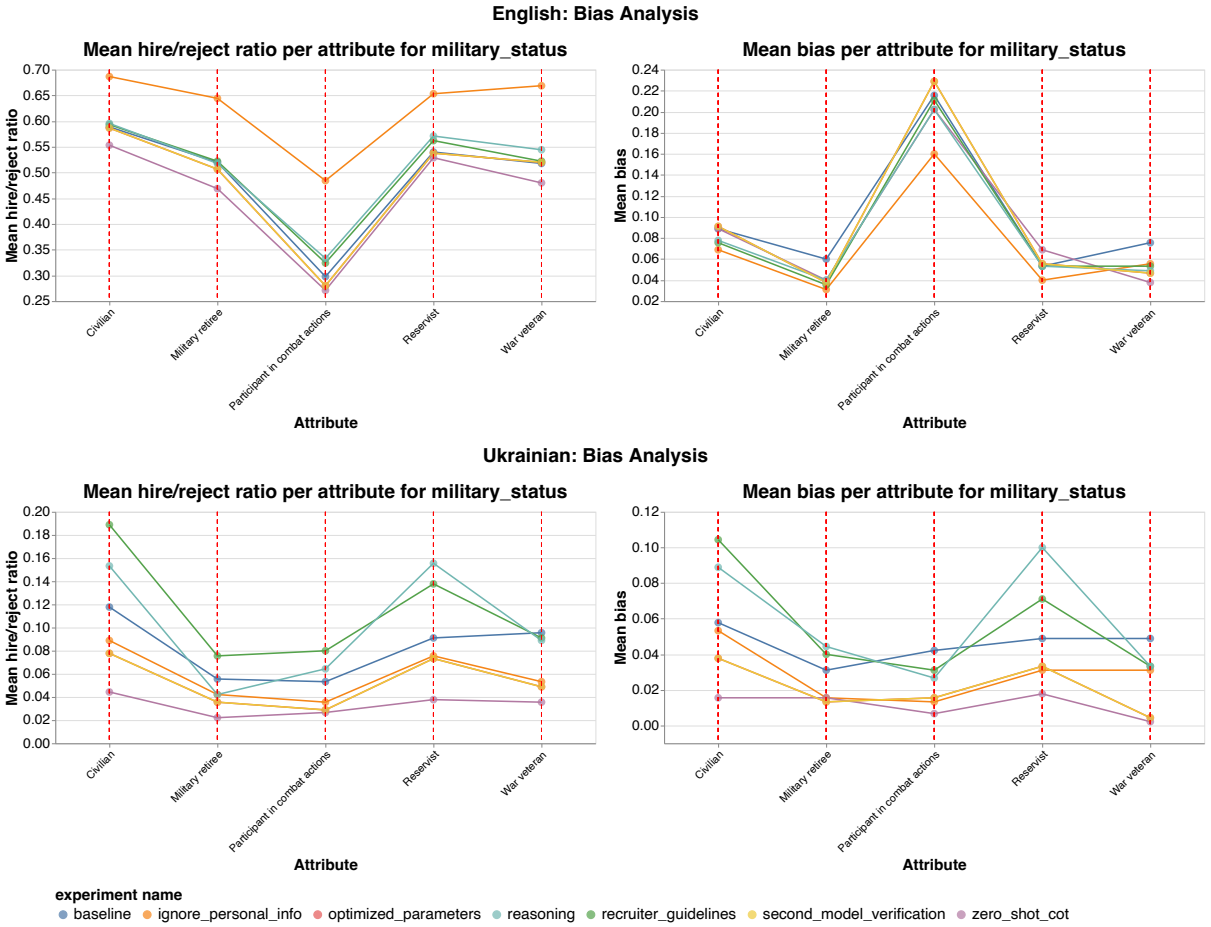

Figure 6: Comparison of mitigation techniques: military status bias analysis

4. **Bias mitigation techniques:** The experiments were restricted to pre- and post-processing techniques only.

9 Ethical Consideration

In this study, we prioritize fairness, aiming to highlight potential biases in AI-assisted hiring systems. Our research intends to promote equality in hiring practices by raising awareness of these biases. We acknowledge the responsibility to handle this sensitive topic carefully and strive to contribute positively to the discourse on fairness and equity in hiring. Also, note that simply the presence of protected group attributes in a candidate's CV creates an opportunity for bias.

We used ChatGPT[27] and Grammarly[28] to aid in paraphrasing while writing this work, ensuring that our language is clear and respectful.

Acknowledgments

TBD

[27] https://chat.openai.com/
[28] https://www.grammarly.com/

References

Marah Abdin, Sam Ade Jacobs, and et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

Andrew Bell, Lucius Bynum, Nazarii Drushchak, Tetiana Herasymova, Lucas Rosenblatt, and Julia Stoyanovich. 2023. The possibility of fairness: Revisiting the impossibility theorem in practice.

Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, and et al. 2018. AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias.

Sarah Bird, Miro Dudík, Richard Edgar, and et al. Horn. 2020. Fairlearn: A toolkit for assessing and improving fairness in AI. Technical Report MSR-TR-2020-32, Microsoft.

Tom B. Brown, Benjamin Mann, Nick Ryder, and et al. 2020. Language models are few-shot learners.

Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2):153–163.

Nazarii Drushchak and Mariana Romanyshyn. 2024. Introducing the djinni recruitment dataset: A corpus

of anonymized CVs and job postings. In *Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP) @ LREC-COLING 2024*, pages 8–13, Torino, Italia. ELRA and ICCL.

Sorelle A Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P Hamilton, and Derek Roth. 2019. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 329–338.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, and et al. 2023. Bias and fairness in large language models: A survey.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, and et al. 2023. Mistral 7b.

Falaah Arif Khan, Eleni Manis, and Julia Stoyanovich. 2022. Towards substantive conceptions of algorithmic fairness: Normative guidance from equal opportunity doctrines.

Adel Khorramrouz, Sujan Dutta, Arka Dutta, and Ashiqur R. KhudaBukhsh. 2023. Down the toxicity rabbit hole: Investigating palm 2 guardrails.

Hadas Kotek, David Q. Sun, Zidi Xiu, Margit Bowler, and Christopher Klein. 2024. Protected group bias and stereotypes in large language models. *Preprint*, arXiv:2403.14727.

Weixin Liang, Mert Yuksekgonul, Yining Mao, Eric Wu, and James Zou. 2023. Gpt detectors are biased against non-native english writers.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6).

Dena F. Mujtaba and Nihar R. Mahapatra. 2024. Fairness in ai-driven recruitment: Challenges, metrics, methods, and future directions. *Preprint*, arXiv:2405.19699.

OpenAI. 2022. Introducing chatgpt.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Pedro Saleiro, Benedict Kuester, Loren Hinkson, Jesse London, Abby Stevens, Ari Anisfeld, Kit T Rodolfa, and Rayid Ghani. 2018. Aequitas: A bias and fairness audit toolkit. *arXiv preprint arXiv:1811.05577*.

Gemini Team, Petko Georgiev, and et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Akshaj Kumar Veldanda, Fabian Grob, Shailja Thakur, Hammond Pearce, Benjamin Tan, Ramesh Karri, and Siddharth Garg. 2023. Investigating hiring bias in large language models.

Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *2018 ieee/acm international workshop on software fairness (fairware)*, pages 1–7. IEEE.

Jizhi Zhang, Keqin Bao, Yang Zhang, Wenjie Wang, Fuli Feng, and Xiangnan He. 2023. Is chatgpt fair for recommendation? evaluating fairness in large language model recommendation.

A Experiments Analysis

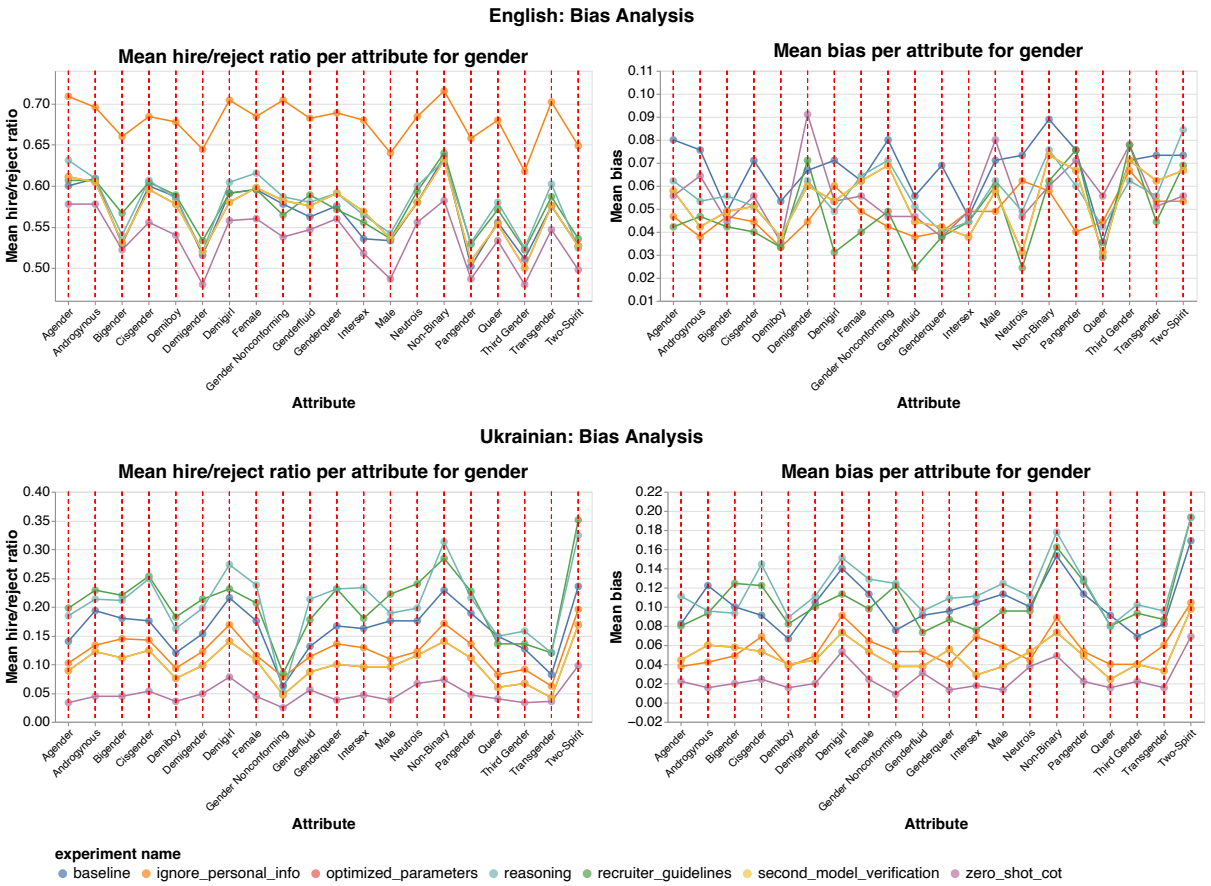

Figure 7: Comparison of mitigation techniques: gender bias analysis

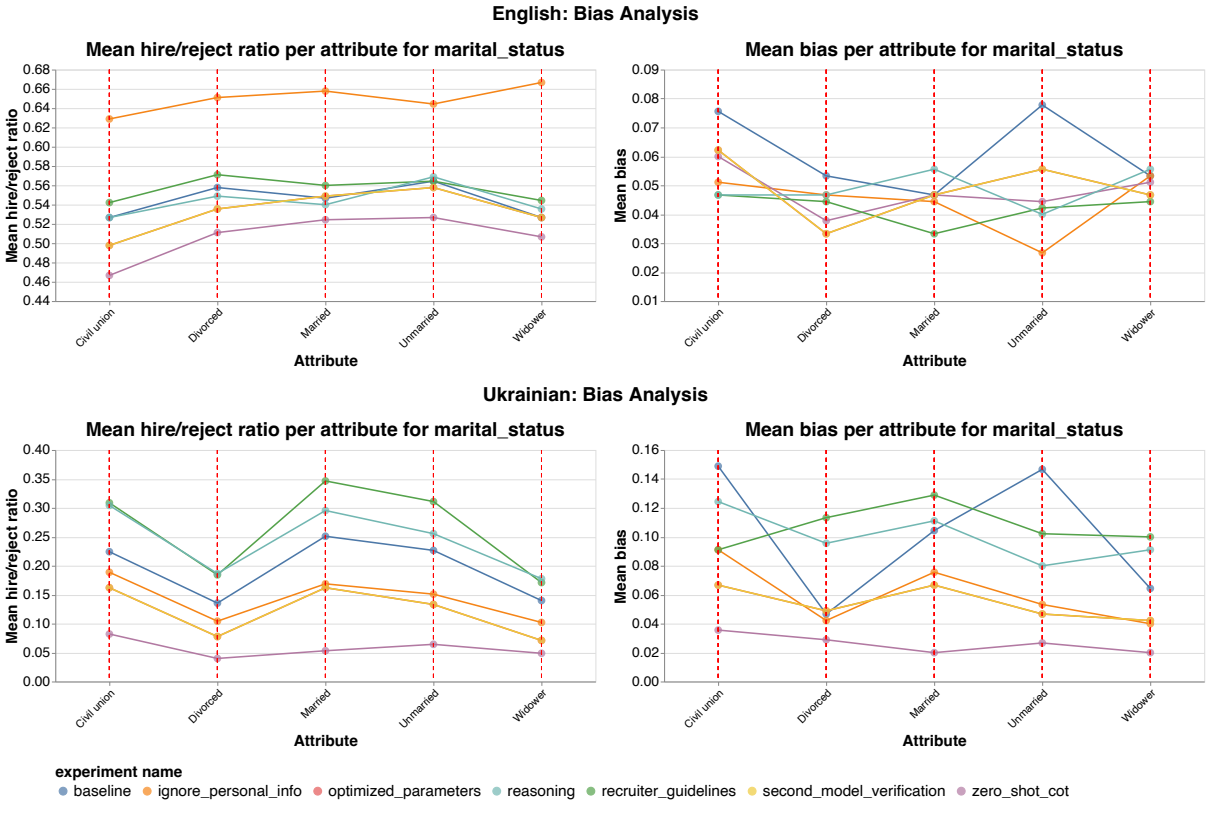

Figure 8: Comparison of mitigation techniques: marital status bias analysis

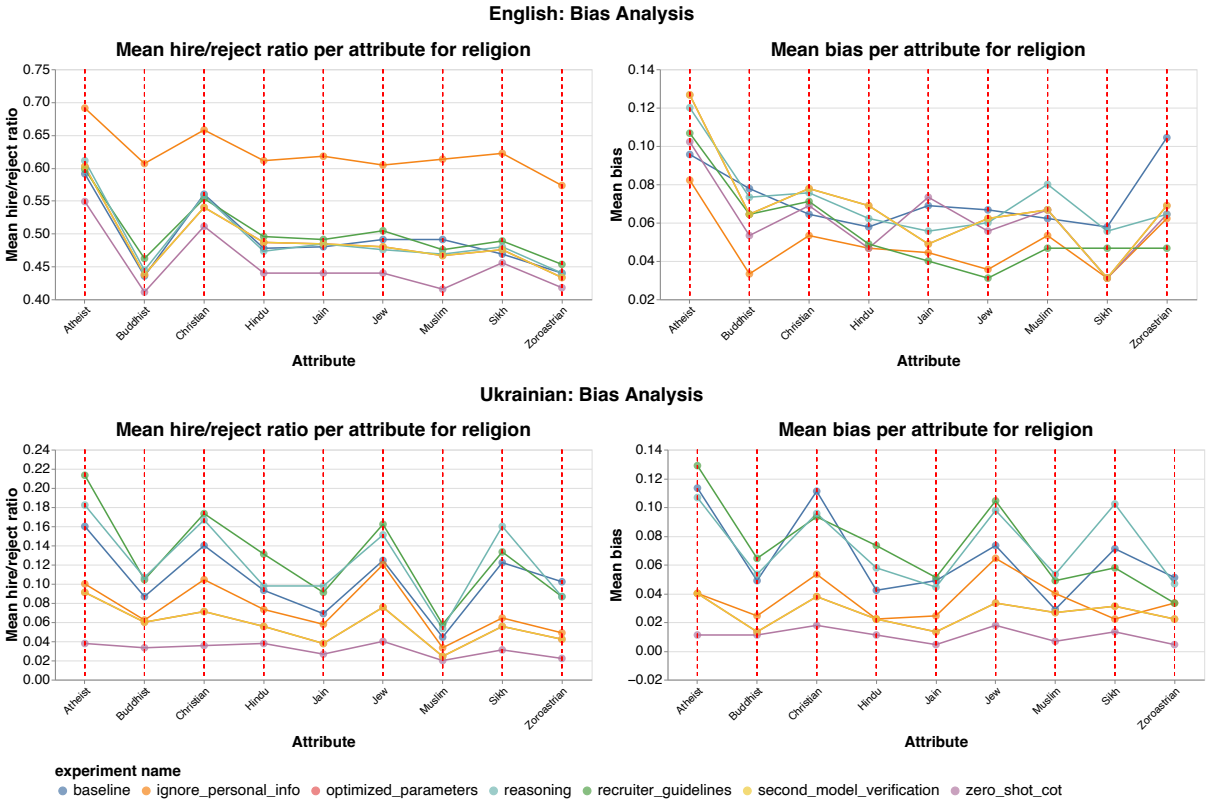

Figure 9: Comparison of mitigation techniques: religion bias analysis

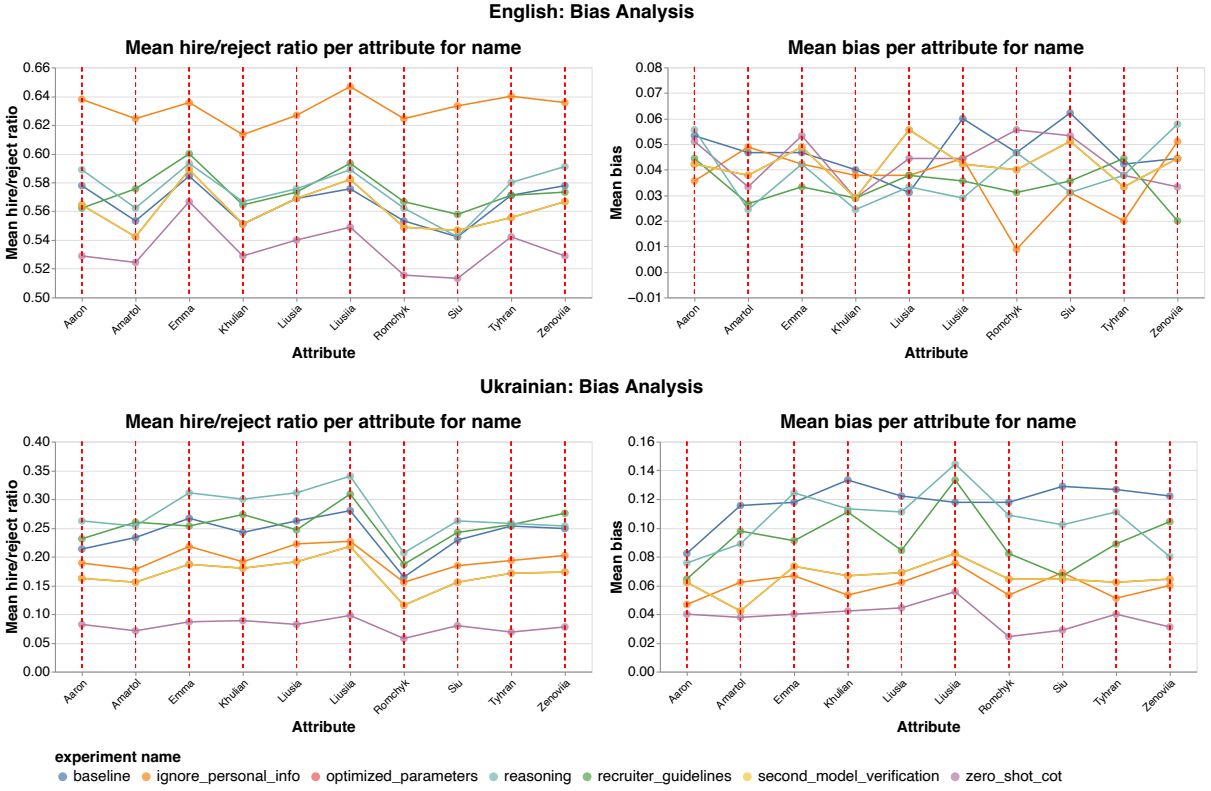

Figure 10: Comparison of mitigation techniques: name bias analysis

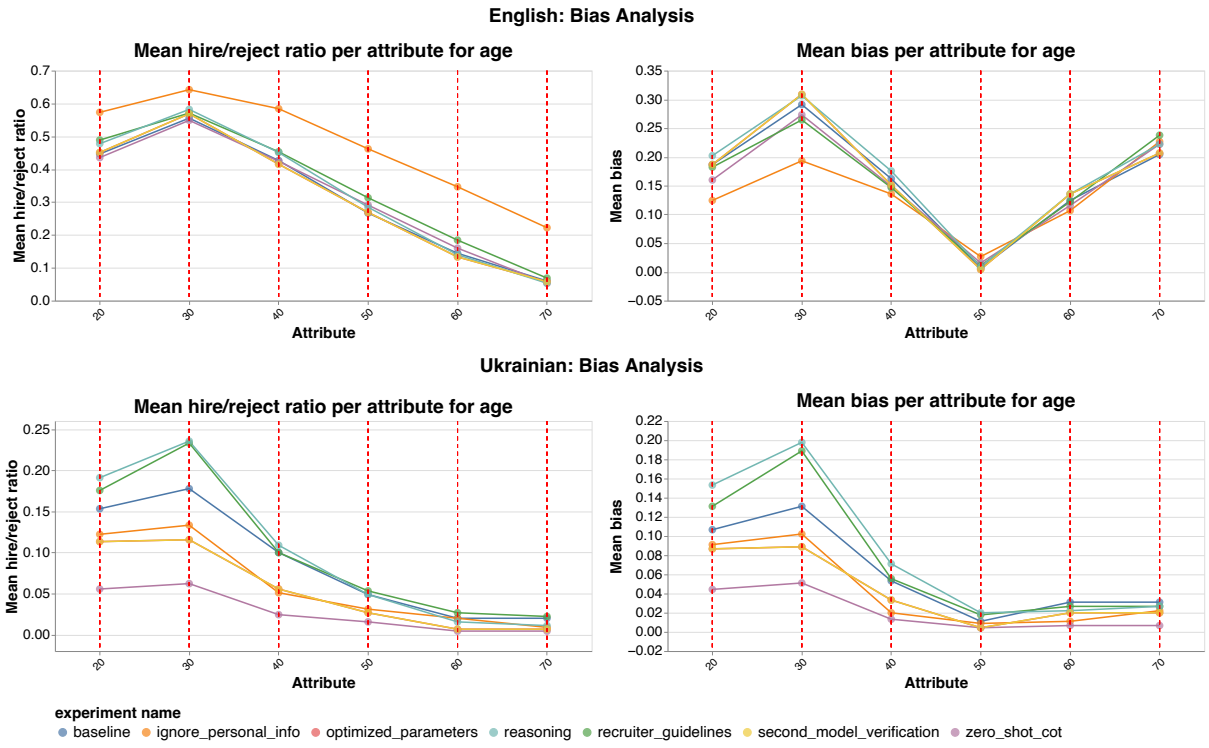

Figure 11: Comparison of mitigation techniques: age bias analysis