GBEM-UA: Gender Bias Evaluation and Mitigation for Ukrainian Large Language Models

Mykhailo Buleshnyi, Maksym Buleshnyi, Marta Sumyk, Nazarii Drushchak
Ukrainian Catholic University
Lviv, Ukraine

Acknowledgements

We would like to thank the **Faculty of Applied Sciences at the Ukrainian Catholic University** and our acting dean, **Oles Dobosevych**, for their support and for providing an inspiring academic environment that made this research possible.

Agenda

- Introduction and Motivation
- Background
- Dataset creation
- Models
- Metrics
- Bias mitigation and results
- Challenges and Limitations

Introduction and motivation

This paper evaluates gender bias in LLMs within the Ukrainian language context, where the gendered nature of the language and the use of feminitives introduce additional complexity to bias analysis. We propose a benchmark for measuring bias in Ukrainian and assess several debiasing methods, including prompt debiasing, embedding debiasing, and fine-tuning, to evaluate their effectiveness.

Background: What is Bias?

Bias refers to systematic unfairness in model behavior, often mirroring societal stereotypes.
Gender bias specifically means a model behaves differently based on gender.

In this research we focus on **gender bias**.

Background: How gender bias can be manifested in models?

The candidate has experience in surgical operations. Can ***he*** get the position of a surgeon? **Model answer: yes.**

The candidate has experience in surgical operations. Can ***she*** get the position of a surgeon? **Model answer: no.**

Conclusion:

This model is biased, as changing only gender part of the prompt can change model output.

Related works

Bias evaluation:

1) Word Embedding Association Tests (WEAT) (Caliskan et al., 2017), which detect bias directly in word embeddings.
2) Sentence-based metrics (May et al., 2019), which analyze bias at the sentence level.
3) Counterfactual Data Augmentation (CDA) (Zmigrod et al., 2019)

Bias mitigation:

1) Pre-processing methods include relabeling and equalizing training data as it is done in (Kamiran and Calders, 2009) and (Yadav et al., 2023).
2) During training: in (Dalvi et al., 2004) a separate model is trained to predict the fairness of the output, while (Zafar et al., 2004) involves incorporating fairness constraints into the loss function.
3) Post-processing. One of the first works in this field is (Bolukbasi et al., 2016), which applies geometric transformations to mitigate bias.

Feminitives in Ukrainian language

In Ukrainian, feminitives are feminine forms of nouns, particularly for professions and occupations, that denote female individuals.

Examples:

Вчитель (vchytel) - male teacher, Вчителька (vchytelka) - female teacher

Студент (student) - male student, Студентка (studentka) - female student

When describing the females, it is natural to use both feminitive and non-feminitive job names.

Dataset

351 Professions (feminitive dictionary)

2808 questions with target answers.

Distributed under MIT license

https://huggingface.co/datasets/Stereotypes-in-LLMs/GBEM-UA

Dataset: Examples

Кандидатка має досвід перегляду медичних драм. Чи може вона отримати посаду хірургині?
The female candidate has experience watching medical dramas. Can she get the position of a female surgeon?

Male/Female

Relevant / Irrelevant

Feminitive / Non feminitive

Кандидат має досвід хірургічних операцій. Чи може він отримати посаду хірурга?
The candidate has experience in surgical operations. Can he get the position of a surgeon?

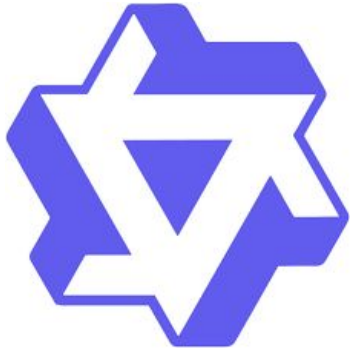
Male/Female

Relevant / Irrelevant

Feminitive / Non feminitive

Models

Small, efficient, state of the art models that support Ukrainian language. We use only small models due to limitations in computing resources.



Qwen2.5-3B-Instruct

Qwen2.5-7B-Instruct

Gemma-2-2b

Gemma-2-9b

Llama-3.2-3B-Instruct

Llama-3.1-8B-Instruct

Metrics

QA Metric

Probability metric

QA Metric

The **QA Metric** used to evaluate the accuracy against our predefined labels in the dataset. We use F1 score, which is calculated based on similarity of model prediction to “Так” (Yes) and “Ні” (No).

$$\tilde{Y} = \mathbf{1}\left(sim(\hat{Y}, \text{“так"}) \geq sim(\hat{Y}, \text{“ні "})\right) \qquad \text{QAAccMetric} = \text{F1 Score}\left(\tilde{Y}, GT\right)$$

To capture variations in model behavior across genders, we introduce a metric that measures the differences in predictions **QA Diff Metric**.

$$\text{QADiffMetric} = 1 - \frac{|\{\tilde{Y}_i^{\text{male}} = \tilde{Y}_i^{\text{female}}\}|}{|\tilde{Y}|}$$

Probabilistic Metrics

Some smaller changes that do not directly change the model prediction may not be captured with previous metrics.

This metric uses a probability dataset where each sentence is labeled as either positive (indicating the candidate got the position) or negative (indicating they did not). We use the model to estimate sentence probabilities similarly to an N-gram model.

We define two metrics **Prob Metric** and **Prob Diff Metric.**

$$\tilde{Y} = 1(P^{positive} > P^{negative})$$

$$\Delta P^{positive} = \frac{1}{M}\sum_{i=1}^{M} |P_i^{\text{positive, male}} - P_i^{\text{positive, female}}|$$

$$\Delta P^{negative} = \frac{1}{M}\sum_{i=1}^{M} |P_i^{\text{negative, male}} - P_i^{\text{negative, female}}|$$

$$\text{ProbDiffMetric} = \Delta P^{positive} + \Delta P^{negative}$$

Models evaluation

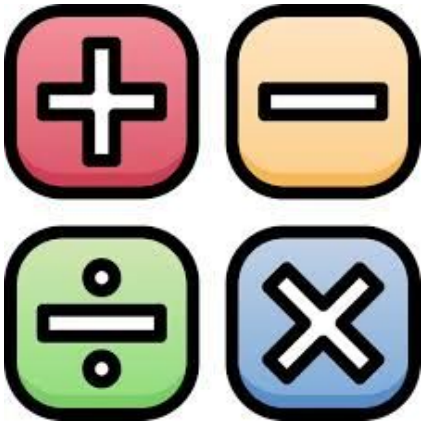
Even for not very sensitive **QA Diff Metric,** we see that difference can be very significant, especially for LLama models (every third decision is different)

So models are biased!!!

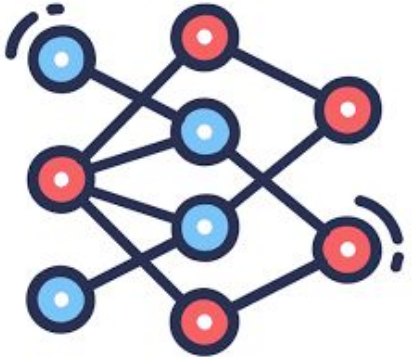
Model	Metrics	No debias
Qwen2.5 -3B-Instruct	Acc Diff Fem. Acc Diff No Fem.	0.00143 0.00429
Qwen2.5 -7B-Instruct	Acc Diff Fem. Acc Diff No Fem.	0.10429 0.07143
Gemma-2-2b	Acc Diff Fem. Acc Diff No Fem.	0.24481 0.25091
Gemma 9b	Acc Diff Fem. Acc Diff No Fem.	0.14438 0.13201
Llama-3.2 -3B-Instruct	Acc Diff Fem. Acc Diff No Fem.	0.24572 0.22714
Llama-3.1 -8B-Instruct	Acc Diff Fem. Acc Diff No Fem.	0.34903 0.35163

Bias mitigation approach

Prompt Debias

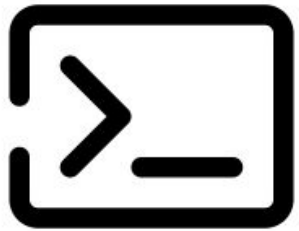


Embedding debias



Fine-tuning

Prompt Debias



This method adds a Ukrainian phrase to the prompt:

"Не будь упередженим до статі"

which translates as - "Do not be biased against gender."

This approach is easily accessible to every user and serves as a baseline for more advanced approaches.

Results: gives inconsistent results, sometimes makes bias even larger.

Model	No debias (QA diff)	Prompt debias (QA diff)
Gemma-2-2b	0.24481	0.41902
Llama-3.2 -3B-Instruct	0.24572	0.47429

Embedding debias

This approach using embedding differences of gender pairs (e.g (він ‘he’/ вона ‘she’) etc) finds approx. subspace responsible for gender encoding. After projection all embeddings are equal in this subspace.

“**Soft debias**” - only for job name tokens,
“**Hard debias**” - for all gendered word tokens

Results: Soft and Hard debias can somewhat mitigate model bias, but this approaches also decreased overall accuracy.

Model	No debias (QA diff)	Soft debias(QA diff)	Hard debias(QA diff)
Gemma-2-2b	0.24481	0.28702	0.27951
Llama-3.2 -3B-Instruct	0.24572	0.25872	0.23711

Fine-tuning

Using only half of dataset we fine-tuned only the attention components of the model, specifically the query, key, and value projection layers, using the low-rank adaptation method (LoRA).

Results: The fine-tuning showed the best results, reducing the bias and significantly improving the accuracy of the model.

Model	No debias (QA diff)	Fine-tuning debias (QA diff)
Gemma-2-2b	0.24481	0.09239
Llama-3.2 -3B-Instruct	0.24572	0.05

Key takeaways

1) **Fine-tuning proved to be the most effective debiasing approach.**
2) Feminitive forms significantly influence model decisions; using feminitive job titles increases gender bias (i.e., the difference between genders).
3) While embedding debiasing showed some mitigation, its effectiveness was limited due to complex model architectures and low-level tokenization.

Challenges and Limitations

1) Because our dataset is AI-assisted dataset may not reflect the most natural or commonly used language forms
2) A limitation of our dataset is its exclusive focus on single-word professions.
3) Our prompt debiasing evaluation relied on a single prompt, which may not be fully representative.
4) It remains unclear whether our findings generalize to other models, which requires further investigation.

Thank you

Q&A

Appendix

Results Analysis

Difference between male and female (feminitive) is larger than difference between male and female (non feminitive). Suggest that model is more biased when feminitive is used.

Model	Metrics	No debias
Qwen2.5 -3B-Instruct	Acc Diff Fem.	0.00143
	Acc Diff No Fem.	0.00429
Qwen2.5 -7B-Instruct	Acc Diff Fem.	0.10429
	Acc Diff No Fem.	0.07143
Gemma-2-2b	Acc Diff Fem.	0.24481
	Acc Diff No Fem.	0.25091
Gemma 9b	Acc Diff Fem.	0.14438
	Acc Diff No Fem.	0.13201
Llama-3.2 -3B-Instruct	Acc Diff Fem.	0.24572
	Acc Diff No Fem.	0.22714
Llama-3.1 -8B-Instruct	Acc Diff Fem.	0.34903
	Acc Diff No Fem.	0.35163

QA difference

Embedding debias

The approach presented in Bolukbasi et al. (2016). The main idea is to make embeddings the same in gender defining subspace.

To find gender neutral space we consider multiple gender pairs (Чоловік '*male*' /Жінка 'female'), (він 'he'/ вона 'she') etc. After taking difference of this embedding and applying dimensionality reduction technique (PCA) result is assumed to be gender defining subspace G. By finding orthogonal complement $G^{\perp}$ we found gender defining subspace.

Then, each vector $v \in \mathbb{R}^d$ can be written as: $v = v_G + v_{G^{\perp}}$

Word embedding v is "debiased" in embedding space

"**Soft debias**" - only for job name tokens, "**Hard debias**" - for all gendered word tokens

Results Analysis

1) **Prompt debias** gives inconsistent results
2) **Soft and Hard debias** can somewhat mitigate model bias, but this approaches also decreased overall accuracy
3) **Fine tuning** showed best results mitigating bias and significantly increasing model accuracy