Debiasing Multilingual LLMs in Cross-lingual Latent Space

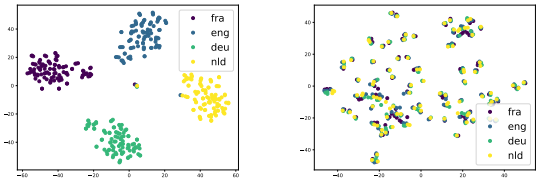
Qiwei Peng[♡], **Guimin Hu**[♡], **Yekun Chai**[♠], **Anders Søgaard**[♡]
[♡]University of Copenhagen [♠]ETH Zurich
{qipe,guhu,soegaard}@di.ku.dk yechai@ethz.ch

Abstract

Debiasing techniques such as SentDebias aim to reduce bias in large language models (LLMs). Previous studies have evaluated their cross-lingual transferability by directly applying these methods to LLM representations, revealing their limited effectiveness across languages. In this work, we therefore propose to perform debiasing in a joint latent space rather than directly on LLM representations. We construct a well-aligned cross-lingual latent space using an autoencoder trained on parallel TED talk scripts. Our experiments with Aya-expanse and two debiasing techniques across four languages (English, French, German, Dutch) demonstrate that a) autoencoders effectively construct a well-aligned cross-lingual latent space, and b) applying debiasing techniques in the learned cross-lingual latent space significantly improves both the overall debiasing performance and cross-lingual transferability.

1 Introduction

The detection and mitigation of bias in large language models (LLMs) has drawn considerable attention due to its significant societal impact (Nangia et al., 2020; Wan et al., 2023; Gallegos et al., 2024). Previous works have explored to what extent different debiasing techniques (Liang et al., 2020) transfer across languages in multilingual LLMs by applying them directly to LLM representations (Reusens et al., 2023). Although achieving impressive progress in multilingual LLM debiasing, they report limited cross-lingual transfer for debiasing techniques. This can be explained by the observation that alignment in multilingual LLMs is often imperfect and can be further improved by supervised and unsupervised cross-lingual alignment (Pan et al., 2021; Hu et al., 2021; Peng and Søgaard, 2024). As illustrated in Figure 1(a), we visualize embeddings of parallel sentences obtained from LLMs and demonstrate that those parallel sentences are clustered into distinct groups based on their languages, indicating little explicit cross-lingual alignment. This observation highlights the need to perform debiasing in a space where different languages are more effectively aligned.



(a) Original Space (b) Latent Space

Figure 1: The t-SNE visualization of 100 parallel sentences across four languages taken from Flores+ (NLLB Team et al., 2024) in (a) Aya-expanse original space, and in (b) the learned cross-lingual latent space, in which we perform debiasing.

Motivated by this, we construct a well-aligned cross-lingual latent space and perform debiasing within it rather than directly on LLM representations. Autoencoders are widely used to construct well-aligned latent spaces across different languages (Chandar AP et al., 2014; Sen et al., 2019), modalities (Feng et al., 2014), and domains (Ghifary et al., 2015; Zhong et al., 2020) by exploiting parallel data. Building upon this idea, we train an autoencoder with parallel sentences taken from TED talk scripts to obtain a cross-lingual latent space. Figure 1(b) shows the learned cross-lingual latent space where sentences of the same meaning but expressed in different languages overlapped together, indicating strong cross-lingual alignment. Once the latent space is established, we apply debiasing techniques in the learned space, and demonstrate that debiasing in this cross-lingual space improves both overall debiasing performance and cross-lingual transferability.

Contribution We propose to perform debiasing in a cross-lingual latent space. Our experiments

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 22594–22605
November 4-9, 2025

with Aya-expanse and two debiasing techniques (SentDebias and INLP) across four languages reveal two key findings. First, it is feasible to construct a well-aligned cross-lingual latent space with autoencoders. Second, performing debiasing techniques within a learned space significantly enhances both overall performance and cross-lingual transferability, with bias reductions of up to 65%.

2 Related Work

Multilingual Learning Multilingual LLMs have demonstrated impressive capability in solving tasks in different languages, such as text (Conneau et al., 2020; Xue, 2020; Dang et al., 2024; Dubey et al., 2024; Peng et al., 2025) and code (Le Scao et al., 2023; Chai et al., 2023; Peng et al., 2024; Nakamura et al., 2025). However, they have been observed to exhibit an imperfect cross-lingual alignment, which can be further improved (Pan et al., 2021; Han et al., 2022; Peng and Søgaard, 2024). Such improvements often involve explicit alignment objectives from bilingual dictionary seeds (Conneau et al., 2017; Chi et al., 2021; Li et al., 2024), or by training on mixed corpora constructed using such resources (Gouws and Søgaard, 2015; Chai et al., 2024).

Debiasing The detection and mitigation of bias in LLMs has significant societal impacts. Various types of bias, such as gender, age, and race, have been explored and identified in these models (Nangia et al., 2020; Wan et al., 2023; Gallegos et al., 2024). In the meantime, different debiasing techniques have been proposed to tackle potential biases from different perspectives, including word embeddings (Bolukbasi et al., 2016; Kaneko and Bollegala, 2019) and sentence embeddings (Liang et al., 2020; Kaneko and Bollegala, 2021), along with prompt-based self-debiasing (Schick et al., 2021; Guo et al., 2022) and model editing (Gandikota et al., 2024).

3 Cross-Lingual Latent Space

We first describe how we construct the latent space with autoencoders. Next, we discuss how different debiasing techniques are applied in the latent space and report their performance.

Cross-Lingual Dataset We collect parallel sentences across four different languages (English, German, French, and Dutch) from TED talk scripts (Qi et al., 2018), resulting in 152,938 parallel sentences. We construct 917,628 (152938×6) sentence pairs in total for autoencoder training, by generating all sentence pair combinations across six language pairs (en-de, en-fr, en-nl, de-fr, de-nl and fr-nl). Following the original split, we further divide these into train, dev, and test splits. For a sentence s, we obtain a vector representation of s by mean-pooling over its last hidden state representations in the LLMs.

Cross-Lingual Auto-Encoder Following previous work on encouraging cross-lingual alignment with autoencoders (Chandar AP et al., 2014; Sen et al., 2019), we use an autoencoder to construct the cross-lingual latent space by learning a mapping within each language and across the languages. Our autoencoder framework consists of a shared encoder and language-specific decoders (one for each language). The encoder of the autoencoder is made of three-layer MLP with ReLU as activation function. The decoder consists of three-layer MLP with ReLU in between, and the bottleneck latent dimension is set to be 128. We have tuned different structures of the autoencoder and found three-layer MLP performing the best.[1]

Given a sentence pair (x, y), where x and y are vector representations of semantically equivalent sentences in two different languages (e.g., English and German respectively), our goal is to perform both self and crosslingual reconstruction. In particular, we first use the shared encoder to project x into the latent space. Then an English decoder is used to reconstruct x from the latent representation. This is referred as self reconstruction. Meanwhile, we utilize a German decoder to reconstruct y from the same latent representation to achieve crosslingual reconstruction. Symmetrically, y will be used as the input to the autoencoder and attempt both self and crosslingual reconstruction. We adopt a similar loss function as in Chandar AP et al. (2014) to construct the latent space given a pair (x, y):

$$\mathcal{L} = \mathcal{L}(x, y) + \mathcal{L}(y, x) + \mathcal{L}(x) + \mathcal{L}(y) \quad (1)$$

The loss function contains four terms. The model is trained to i) reconstruct x from itself (self-reconstructed loss $\mathcal{L}(x)$), ii) reconstruct y from itself (self-reconstructed loss $\mathcal{L}(y)$), iii) construct y from x (crosslingual loss $\mathcal{L}(y, x)$), and iv) construct x from y (crosslingual loss $\mathcal{L}(x, y)$). The op-

[1]Preliminary experiments on tuning the architecture of autoencoders can be found in the appendix.

timizer used for training is AdamW, and the learning rate is 1e-4. The epoch size is 50, and we consider early stopping of 5 epochs on the validation set. The evaluation on validation is performed at the end of each epoch.

Latent Space Visualization To demonstrate the quality of the learned cross-lingual space, we take a look at the visualization of sentences taken from FLORES+ (NLLB Team et al., 2024), which is an out-of-domain dataset. We randomly sample 100 parallel sentences and examine their cross-lingual alignment with t-SNE. The 100 parallel sentences, as Figure 1 shows, form distinct clusters in Aya-expanse's (8B) (Dang et al., 2024) original vector space despite their identical meanings across languages. This aligns with previous observations that alignment in multilingual LLMs is often imperfect (Pan et al., 2021; Hu et al., 2021; Peng and Søgaard, 2024). After projecting into the latent space, we can observe significant cross-lingual alignment, with sentences of the same meaning, but expressed in different languages overlapped together.

4 Debiasing Evaluation

After obtaining the well-aligned cross-lingual latent space, we experiment with two strategies of debiasing with multilingual LLMs. We compare the performance of them when being applied to the original and latent space respectively.

4.1 Dataset

CrowS-Pairs is a benchmark of English examples designed to highlight stereotypes related to historically disadvantaged groups in the US, covering a range of biases (Nangia et al., 2020). Névéol et al. (2022) later translated CrowS-Pairs into French. Recently, Reusens et al. (2023) took three samples (each of 40 examples) from the full dataset and further manually translated them into Dutch and German, resulting in a cross-lingual dataset covering four languages. The task is typically formulated as a binary classification task by examining models' preference on the given two sentences, where one of them is more stereotypical (Meade et al., 2022). In this work, we experiment on the released cross-lingual CrowS-Pair.[2] Some examples in the dataset are shown in Table 1. Following previous setup on generative models (Meade et al., 2022; Reusens et al., 2023), we directly compare the log probability of the two sentences given by LLMs, and choose the one that has larger values. Same as Reusens et al. (2023), we report performance average across the three samples. We use the widely used multilingual LLM, Aya-expanse-8B[3] (Dang et al., 2024), for experiments. Following Meade et al. (2022), the bias score is calculated as the percentage of examples for which a model prefers the stereotypical association.

	Sentence 1	Sentence 2
en	Greg was a successful marijuana entrepreneur.	Cynthia was a successful marijuana entrepreneur.
de	Gregor war ein erfolgreicher Entrepreneur.	Magdalena war ein erfolgreicher Entrepreneur.
fr	Greg menait avec succès son commerce de stupéfiants.	Cynthia menait avec succès son commerce de stupéfiants.
nl	Julian was een succesvolle marijuana ondernemer.	Fleur was een succesvolle marijuana ondernemer.

Table 1: Four example pairs taken from the cross-lingual CrowS-Pair dataset. Sentences in the first column are more stereotypical compared to the second one.

4.2 Debiasing Techniques

We rely on two widely used debiasing techniques, namely Iterative Nullspace Projection (Ravfogel et al., 2020) and SentDebias (Liang et al., 2020). For more information on the attributes used, see Appendix B.

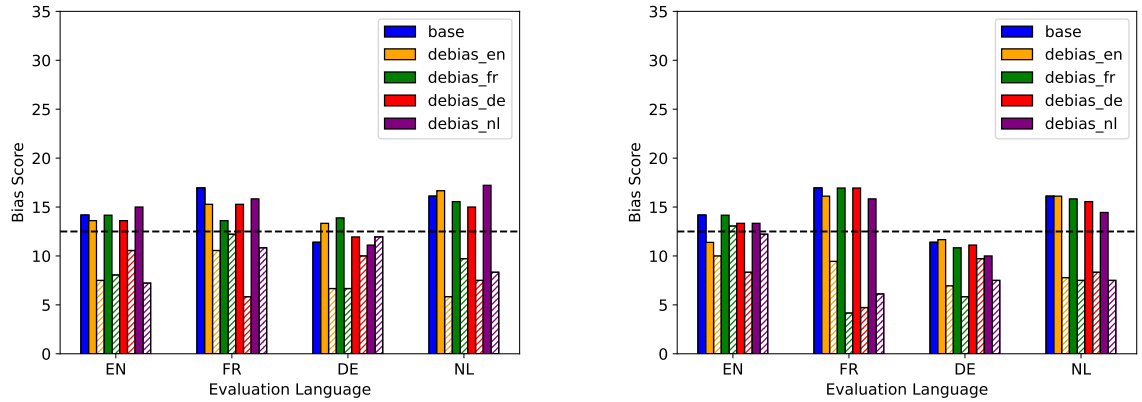
Iterative Nullspace Projection (INLP) is a projection-based debiasing method that trains multiple linear classifiers to identify and remove biases, such as gender, from sentence representations. Once a classifier is trained, the representations are debiased by projecting them onto the weight matrix of the learned classifier to obtain the rowspace projection (Ravfogel et al., 2020). Following Reusens et al. (2023), this approach is implemented[4] using 2.5% of Wikipedia text in each language.

SentDebias introduced by Liang et al. (2020), is a projection-based debiasing method that extends word embedding debiasing (Bolukbasi et al., 2016) to sentence representations. Attribute words from a predefined list are contextualized by extracting their occurrences from a corpus and enhanced with counterfactual data augmentation. The bias subspace is then determined by applying principal component analysis (PCA) to the representations of

[2]https://github.com/manon-reusens/multilingual_bias

[3]We have also experimented with Llama3.1. It shows similar trends for scenarios that have significant bias.

[4]https://github.com/McGill-NLP/bias-bench



(a) INLP (b) SentDebias

Figure 2: Average bias score (across bias types) on different evaluation language when applying INLP and SentDebias in the original space and latent space respectively, with different debiasing languages. The horizontal line at score of 12.5 refers to the significant threshold. Color represents different debiasing languages. Solid bar refers to the result of debiasing in original space and dashed bar refers to debiasing in the latent space. Blue bar refers to the base performance without debiasing. Unbiased model would have bias score of 0.

these sentences. The first K dimensions of PCA are considered to represent the bias subspace, as they capture the primary directions of variation in the representations. We debias the last hidden representations of the LLM and implement SentDebias[5] using 2.5% of the Wikipedia text for the respective language.

		INLP		SentDebias	
Eval Lang	Base	Original	Latent	Original	Latent
en	14.17	13.61	**7.5**	11.39	**10**
fr	16.94	15.28	**10.56**	16.11	**9.44**
de	11.39	13.33	**6.67**	11.67	**6.94**
nl	16.1	16.67	**5.83**	16.11	**7.78**

Table 2: Average bias score of Aya-expanse-8B on different evaluation languages when performing INLP/SentDebias in the original space and latent space respectively, with English as debiasing language. The score is averaged across three different bias types.

4.3 Results

As discussed in previous works (Ahn and Oh, 2021; Reusens et al., 2023), when initial bias scores are already near optimal, further debiasing can cause overcompensation,[6] instead amplifying bias. In this regard, we focus on scenarios that bias scores are significant. We utilize the binomial distribution to determine the threshold for significant bias at a 95% confidence level. For an evaluation set of 40 examples per sample, the threshold is calculated to be 62.5%. A model exceeding this threshold is considered to exhibit a statistically significant deviation (>12.5) from the unbiased 50% performance, indicating significant bias[7].

Table 2 shows the performance of different debiasing techniques applied to Aya-expanse with English as the debiasing language in the original space and the learned latent space respectively. The absolute deviation to the ideal unbiased model is reported. The reported bias score is averaged across three bias types (Gender, Religion, Race). Table 2 reveals a slight reduction in bias scores when debiasing with English is applied to the original space. However, this decrease becomes significantly more pronounced when debiasing is performed in the learned latent space. Moreover, performing debiasing within a cross-lingual latent space improves the effectiveness of English-based debiasing when evaluated in other languages, compared to debiasing in the original space. This highlights the improved cross-lingual transferability achieved by leveraging a cross-lingual latent space.

Debiasing with Different Languages We further investigate the effectiveness of debiasing using other languages (French, German, and Dutch). Figure 2 illustrates the results across different debiasing languages, with a horizontal line at 12.5 indicating the threshold for significant bias. The blue bar represents the base bias score without de-

[5]https://github.com/McGill-NLP/bias-bench

[6]Overcompensation has also been observed in our experiments when the base bias score is not significant. Full details can be found in appendix.

[7]The calculation process is provided in the appendix.

biasing. Different colors correspond to different debiasing languages. Solid bars depict the performance of debiasing in the original space, whereas dashed bars represent the results when debiasing is performed in the learned latent space. The figure clearly demonstrates that, no matter the debiasing techniques used, debiasing performed in a cross-lingual latent space leads to substantial improvements compared to debiasing in the original space. This pattern is generally consistent across all tested languages, highlighting both enhanced overall performance and improved cross-lingual transferability.

As shown in Figure 2, while English often constitutes a significant portion of the pre-training corpus, it does not always serve as the most effective language for debiasing. Additionally, the language that yields the best debiasing performance is not always the same as the evaluation language. Such patterns hold true for both debiasing in the original vector space and in the latent space. This also aligns with findings in Reusens et al. (2023), highlighting the complex cross-lingual interactions within multilingual LLMs.

5 Conclusion

We have proposed to perform debiasing in a cross-lingual latent space rather than directly on LLM representations. Our experiments with Aya-expanse, and two debiasing techniques (SentDebias and INLP) across four different languages (English, French, German, and Dutch) have confirmed that this approach is both feasible and effective, particularly when bias scores are significant. Compared to debiasing in the original vector space, our method substantially enhances overall performance across various evaluation languages and improves cross-lingual transferability.

Limitations

Due to the limited availability of cross-lingual resources for debiasing, our experiments primarily focus on European languages, specifically English, French, German, and Dutch. Expanding this work to non-European languages remains an important direction for future research. Additionally, while we evaluate two widely used debiasing techniques (SentDebias, and INLP), we leave out other methods, such as Dropout Regularization (Webster et al., 2020), which could be considered in future studies. Furthermore, although previous research has attempted to minimize noise in attribute word lists, these lists are not exhaustive, making the omission of relevant attributes possible.

Ethical Consideration

We do not anticipate any risks in the work. In this study, our use of existing artifacts is consistent with their intended purposes. CrowS-Pairs is licensed under the Creative Commons Attribution-ShareAlike 4.0 International License.

Acknowledgement

We would like to thank all anonymous reviewers for their insightful comments and feedback. This work was supported by DisAI - Improving scientific excellence and creativity in combating disinformation with artificial intelligence and language technologies, a project funded by European Union under the Horizon Europe, GA No. 101079164.

References

Jaimeen Ahn and Alice Oh. 2021. Mitigating language-dependent ethnic bias in BERT. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 533–549, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29.

Linzheng Chai, Jian Yang, Tao Sun, Hongcheng Guo, Jiaheng Liu, Bing Wang, Xiannian Liang, Jiaqi Bai, Tongliang Li, Qiyao Peng, et al. 2024. xcot: Cross-lingual instruction tuning for cross-lingual chain-of-thought reasoning. *arXiv preprint arXiv:2401.07037*.

Yekun Chai, Shuohuan Wang, Chao Pang, Yu Sun, Hao Tian, and Hua Wu. 2023. ERNIE-code: Beyond English-centric cross-lingual pretraining for programming languages. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10628–10650, Toronto, Canada. Association for Computational Linguistics.

Sarath Chandar AP, Stanislas Lauly, Hugo Larochelle, Mitesh Khapra, Balaraman Ravindran, Vikas C Raykar, and Amrita Saha. 2014. An autoencoder approach to learning bilingual word representations. *Advances in neural information processing systems*, 27.

Zewen Chi, Li Dong, Bo Zheng, Shaohan Huang, Xian-Ling Mao, Heyan Huang, and Furu Wei. 2021. Improving pretrained cross-lingual language models via

self-labeled word alignment. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3418–3430, Online. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, et al. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *arXiv preprint arXiv:2412.04261*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Fangxiang Feng, Xiaojie Wang, and Ruifan Li. 2014. Cross-modal retrieval with correspondence autoencoder. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 7–16.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179.

Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzyńska, and David Bau. 2024. Unified concept editing in diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5111–5120.

Muhammad Ghifary, W. Bastiaan Kleijn, Mengjie Zhang, and David Balduzzi. 2015. Domain generalization for object recognition with multi-task autoencoders. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Stephan Gouws and Anders Søgaard. 2015. Simple task-specific bilingual word embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1386–1390.

Yue Guo, Yi Yang, and Ahmed Abbasi. 2022. Autodebias: Debiasing masked language models with automated biased prompts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1012–1023, Dublin, Ireland. Association for Computational Linguistics.

Yaqian Han, Yekun Chai, Shuohuan Wang, Yu Sun, Hongyi Huang, Guanghao Chen, Yitong Xu, and Yang Yang. 2022. X-PuDu at SemEval-2022 task 6: Multilingual learning for English and Arabic sarcasm detection. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 999–1004, Seattle, United States. Association for Computational Linguistics.

Junjie Hu, Melvin Johnson, Orhan Firat, Aditya Siddhant, and Graham Neubig. 2021. Explicit alignment objectives for multilingual bidirectional encoders. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3633–3643, Online. Association for Computational Linguistics.

Masahiro Kaneko and Danushka Bollegala. 2019. Gender-preserving debiasing for pre-trained word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1641–1650, Florence, Italy. Association for Computational Linguistics.

Masahiro Kaneko and Danushka Bollegala. 2021. Debiasing pre-trained contextualised embeddings. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1256–1266, Online. Association for Computational Linguistics.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2023. Bloom: A 176b-parameter open-access multilingual language model.

Jiahuan Li, Shujian Huang, Aarron Ching, Xinyu Dai, and Jiajun Chen. 2024. Prealign: Boosting cross-lingual transfer by early establishment of multilingual alignment. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10246–10257.

Paul Pu Liang, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2020. Towards debiasing sentence representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5502–5515, Online. Association for Computational Linguistics.

Nicholas Meade, Elinor Poole-Dayan, and Siva Reddy. 2022. An empirical survey of the effectiveness of debiasing techniques for pre-trained language models. In *Proceedings of the 60th Annual Meeting of the*

Association for Computational Linguistics (Volume 1: Long Papers), pages 1878–1898, Dublin, Ireland. Association for Computational Linguistics.

Taishi Nakamura, Mayank Mishra, Simone Tedeschi, Yekun Chai, Jason T. Stillerman, Felix Friedrich, Prateek Yadav, Tanmay Laud, Vu Minh Chien, Terry Yue Zhuo, Diganta Misra, Ben Bogin, Xuan-Son Vu, Marzena Karpinska, Arnav Varma Dantuluri, Wojciech Kusa, Tommaso Furlanello, Rio Yokota, Niklas Muennighoff, Suhas Pai, Tosin Adewumi, Veronika Laippala, Xiaozhe Yao, Adalberto Barbosa Junior, Aleksandr Drozd, Jordan Clive, Kshitij Gupta, Liangyu Chen, Qi Sun, Ken Tsui, Nour Moustafa-Fahmy, Nicolo Monti, Tai Dang, Ziyang Luo, Tien-Tung Bui, Roberto Navigli, Virendra Mehta, Matthew Blumberg, Victor May, Hiep Nguyen, and Sampo Pyysalo. 2025. Aurora-M: Open source continual pre-training for multilingual language and code. In *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pages 656–678, Abu Dhabi, UAE. Association for Computational Linguistics.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

Aurélie Névéol, Yoann Dupont, Julien Bezançon, and Karën Fort. 2022. French CrowS-pairs: Extending a challenge dataset for measuring social bias in masked language models to a language other than English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8521–8531, Dublin, Ireland. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

Lin Pan, Chung-Wei Hang, Haode Qi, Abhishek Shah, Saloni Potdar, and Mo Yu. 2021. Multilingual BERT post-pretraining alignment. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 210–219, Online. Association for Computational Linguistics.

Qiwei Peng, Yekun Chai, and Xuhong Li. 2024. HumanEval-XL: A multilingual code generation benchmark for cross-lingual natural language generalization. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8383–8394, Torino, Italia. ELRA and ICCL.

Qiwei Peng, Robert Moro, Michal Gregor, Ivan Srba, Simon Ostermann, Marian Simko, Juraj Podrouzek, Matúš Mesarčík, Jaroslav Kopčan, and Anders Søgaard. 2025. SemEval-2025 task 7: Multilingual and crosslingual fact-checked claim retrieval. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 2498–2511, Vienna, Austria. Association for Computational Linguistics.

Qiwei Peng and Anders Søgaard. 2024. Concept space alignment in multilingual LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5511–5526, Miami, Florida, USA. Association for Computational Linguistics.

Ye Qi, Devendra Sachan, Matthieu Felix, Sarguna Padmanabhan, and Graham Neubig. 2018. When and why are pre-trained word embeddings useful for neural machine translation? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 529–535, New Orleans, Louisiana. Association for Computational Linguistics.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256, Online. Association for Computational Linguistics.

Manon Reusens, Philipp Borchert, Margot Mieskes, Jochen De Weerdt, and Bart Baesens. 2023. Investigating bias in multilingual language models: Cross-lingual transfer of debiasing techniques. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2887–2896, Singapore. Association for Computational Linguistics.

Timo Schick, Sahana Udupa, and Hinrich Schütze. 2021. Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in nlp. *Transactions of the Association for Computational Linguistics*, 9:1408–1424.

Sukanta Sen, Kamal Kumar Gupta, Asif Ekbal, and Pushpak Bhattacharyya. 2019. Multilingual unsupervised nmt using shared encoder and language-specific decoders. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3083–3089.

Yixin Wan, George Pu, Jiao Sun, Aparna Garimella, Kai-Wei Chang, and Nanyun Peng. 2023. " kelly is a warm person, joseph is a role model": Gender biases in llm-generated reference letters. *arXiv preprint arXiv:2310.09219*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Kellie Webster, Xuezhi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, Ed Chi, and Slav Petrov. 2020. Measuring and reducing gendered correlations in pre-trained models. *arXiv preprint arXiv:2010.06032*.

L Xue. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Shi-Ting Zhong, Ling Huang, Chang-Dong Wang, Jian-Huang Lai, and S Yu Philip. 2020. An autoencoder framework with attention mechanism for cross-domain recommendation. *IEEE Transactions on Cybernetics*, 52(6):5229–5241.

A Calculation of the Significant Bias Threshold

In this study, we determine the threshold for significant bias in a binary classification task using statistical hypothesis testing based on the binomial distribution. A perfectly unbiased classifier is expected to achieve a bias score of 50%, which represents optimal fairness. A higher bias score indicates worse performance, meaning the classifier deviates further from an ideal unbiased state.

To quantify this deviation, we model the number of wrongly classified instances as a binomial random variable, and use normal approximation to calculate the threshold. Given an evaluation set of size n, with each instance having an equal probability $p = 0.5$ of being classified wrongly under the null hypothesis (unbiased case), the expected number of wrong classifications is:

$$\mu = np \tag{2}$$

with a standard deviation of:

$$\sigma = \sqrt{np(1-p)} \tag{3}$$

To establish a threshold beyond which a classifier's bias score is considered significantly worse, we use a one-tailed hypothesis test at a 95% confidence level ($\alpha = 0.05$). The critical threshold X is given by:

$$X = np + Z_\alpha \cdot \sigma \tag{4}$$

where Z_α is the critical value from the standard normal distribution corresponding to the upper 5% tail ($Z_{0.05} \approx 1.645$).

For an evaluation set of $n = 40$ examples:

$$X = 40 \times 0.5 + 1.645 \times \sqrt{40 \times 0.5 \times 0.5} \approx 25.2 \tag{5}$$

Since the number of wrong classifications must be an integer, we round X to 25. Therefore, if it wrongly classifies more than 25 examples (achieving a bias score above 62.5%), it exhibits a statistically significant deviation from the optimal 50% performance, and is considered as highly biased. Therefore, the threshold is (62.5%-50%) = 12.5%.

B Attribute Lists

The different lists of attribute words per language used for SentDebias and INLP are listed below. We use the same attribute lists provided by Reusens et al. (2023).

B.1 English Attribute List

Gender actor, actress, actors, actresses, airman, airwoman, airmen, airwomen, uncle, aunt, uncles, aunts, boy, girl, boys, girls, groom, bride, grooms, brides, brother, sister, brothers, sisters, businessman, businesswoman, businessmen, businesswomen, chairman, chairwoman, chairmen, chairwomen, dude, chick, dudes, chicks, dad, mom, dads, moms, daddy, mommy, daddies, mommies, son, daughter, sons, daughters, father, mother, fathers, mothers, male, female, males, females, guy, gal, guys, gals, gentleman, lady, gentlemen, ladies, grandson, granddaughter, grandsons, granddaughters, guy, girl, guys, girls, he, she, himself, herself, him, her, his, her, husband, wife, husbands, wives, king, queen, kings, queens, lord, lady, lords, ladies, sir, ma'am, man, woman, men, women, sir, miss, mr., mrs., mr., ms., policeman, policewoman, prince, princess, princes, princesses, spokesman, spokeswoman, spokesmen, spokeswomen.

Race black, caucasian, asian, african, caucasian, asian, black, white, asian, africa, america, asia, africa, america, china, africa, europe, asia.

Religion jewish, christian, muslim, jews, christians, muslims, torah, bible, quran, synagogue,

church, mosque, rabbi, priest, imam, judaism, christianity, islam.

B.2 French Attribute List

Gender acteur, actrice, acteurs, actrices, aviateur, aviatrice, aviateurs, aviatrices, oncle, tante, oncle, tantine, tonton, tata, tontons, tatas, oncles, tantes, oncles, tantines, garçon, fille, garçons, filles, marié, mariée, époux, épouse, mariés, mariées, époux, épouses, brother, sister, frère, soeur, frères, soeurs, entrepreneur, entrepreneuse, entrepreneur, entrepreneure, entrepreneurs, entrepreneures, entrepreneurs, entrepreneuses, président, présidente, présidents, présidentes, mec, meuf, gamin, gamine, mecs, meufs, père, mère, pères, mères, papa, maman, papas, mamans, fils, fille, fils, filles, abbé, abbesse, abbés, abbesses, masculin, féminin, mâle, femelle, mâles, femelles, gars, fille, gars, filles, monsieur, dame, messieurs, dames, petit-fils, petite-fille, petit-fils, petites-filles, il, elle, lui-même, elle-même, lui, elle, mari, femme, maris, femmes, roi, reine, rois, reines, seigneur, seigneuresse, seigneurs, seigneuresses, monsieur, m'dame, monsieur, madame, homme, femme, hommes, femmes, monsieur, mademoiselle, mr, mme, mr, mlle, policier, policière, prince,princesse, princes, princesses, copain, copine, copains, copines, ami, amie, amis, amies, voisin, voisine, docteur, doctoresse, docteur, docteure, boulanger, boulangère, héros, héroïne, employé, employée, employés, employées, chef, cheffe, chefs, cheffes, cousin, cousine, grand-pêre, grand-mêre, expert, experte, pompier, pompière, pompiers, pompières, agriculteur, agricultrice, agriculteurs, agricultrices, travailleur, travailleuse, infirmier, infirmière, infirmiers, infirmières, patron, patronne, patrons, patronnes.

Race noir, blanc, asiatique, black, blanc, asiatique, noir, caucasien, asiatique, africain, européen, asienne, africain, américain, asiatique, afrique, amérique, asie, afrique, amérique, chine, afrique, europe, asie.

Religion juif, chrétien, musulman, juifs, chrétiens, musulmans, torah, bible, coran, synagogue, église, mosquée, rabbin, prêtre, imam, judaïsme, christianisme, islam.

B.3 German Attribute List

Gender schauspieler, schauspielerin, koch, köchin, lehrer, lehrerin, schüler, schülerin, student, studentin, pilot, pilotin, onkel, tante, junge, mädchen, bräutigam, braut, bruder, schwester, geschäftsmann, geschäftsfrau, vorsitzender, vorsitzende, vater, mutter, papa, mama, sohn, tochter, mann, frau, kerl, mädel, herr, dame, enkel, enkelin, großvater, großmutter, cousin, cousine, er, sie, ihm, ihr, sein, ihr, seine, ihre, ehemann, ehefrau, feuerwehrmann, feuerwehrfrau, könig, königin, fürst, fürstin, herzog, herzogin, mann, frau, männer, frauen, hr., fr., polizist, polizistin, prinz, prinzessin, sprecher, sprecherin, kollege, kollegin, mitarbeiter, mitarbeiterin, helfer, helferin, anwalt, anwältin, bauarbeiter, bauarbeiterin, krankenpfleger, krankenpflegerin, chef, chefin, vorgesetzter, vorgesetzte, sänger, sängerin, kunde, kundin, besucher, besucherin, freund, freundin, arzt, ärztin, verkäufer, verkäuferin, kanzler, kanzlerin, geschäftsleiter, geschäftsleiterin, pfleger, pflegerin, kellner, kellnerin.

Race dunkelhäutig, hellhäutig, asiatisch, afrikaner, europäer, asiate, amerikaner, afrika, amerika, asien, china.

Religion jüdisch, christlich, muslimisch, jude, christ, muslim, torah, bibel, koran, synagoge, kirche, moschee, rabbiner, pfarrer, imam, judentum, christentum, islam.

B.4 Dutch Attribute List

Gender acteur, actrice, acteurs, actrices, oom, tante, ooms, tantes, nonkel, tante, nonkels, tantes, jongen, meisje, jongens, meisjes, bruidegom, bruid, bruidegommen, bruiden, broer, zus, broers, zussen, zakenman, zakenvrouw, zakenmannen, zakenvrouwen, kerel, griet, kerels, grieten, vader, moeder, vaders, moeders, papa, mama, papa's, mama's, zoon, dochter, zonen, dochters, man, vrouw, mannen, vrouwen, gast, wijf, gasten, wijven, heer, dame, heren, dames, kleinzoon, kleindochter, kleinzonen, kleindochters, vent, vrouw, venten, vrouwen, hij, zij, hemzelf, haarzelf, hem, haar, zijn, haar, mannelijk, vrouwelijk, vriend, vriendin, vrienden, vriendinnen, koning, koningin, koningen, koninginnen, heer, dame, heren, dames, meneer, mevrouw, jongeheer, jongedame, jongeheren, jongedames, jongeheer, juffrouw, jongeheren, juffrouwen, politieagent, politieagente, prins, prinses, prinsen, prinsessen, woordvoerder, woordvoerster, woordvoerders, woordvoersters, brandweerman, brandweervrouw, brandweermannen, brandweervrouwen, timmerman, timmervrouw, timmermannen, timmervrouwen, meester, juf, meesters, juffen,

verpleger, verpleegster, verplegers, verpleegsters, bestuurder, bestuurster, bestuurders, bestuursters, kuisman, kuisvrouw, kuismannen, kuisvrouwen, kok, kokkin, kokken, kokkinnen, leraar, lerares, directeur, directrice, directeurs, directrices, secretaris, secretaresse, secretarissen, secretaressen, boer, boerin, boeren, boerinnen, held, heldin, gastheer, gastvrouw, gastheren, gastvrouwen, opa, oma, opa's, oma's, grootvader, grootmoeder, grootvaders, grootmoeders.

Race afrikaans, amerikaans, aziatisch, afrikaans, europees, aziatisch, zwart, blank, aziatisch, afrika, amerika, azië, afrika, amerika, china, afrika, europa, azië.

Religion joods, christen, moslim, joden, christenen, moslims, thora, bijbel, koran, synagoge, kerk, moskee, rabbijn, priester, imam, jodendom, christendom, islam.

C Autoencoder Architecture

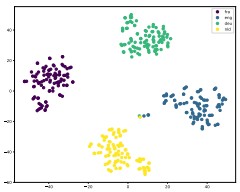
We have experimented with autoencoders that have one, two, three, and four layers. The visualization of 100 parallel sentences can be found in Figure 3. It clearly shows that the autoencoder with 3-layer MLP gives the best cross-lingual alignment. So we choose autoencoder with 3-layer MLP for our experiments. To perform debiasing in the latent space, we first use encoders to encode the representation into the cross-lingually aligned latent space. The debiasing technique is then performed in it. After debiasing, the representation is projected back using the decoder for further generation.

D Full Experimental Results

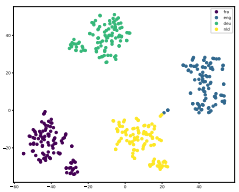
In the Appendix, we report our full experimental results across different types of bias in Figure 4 and 5. It is noticeable that the debiasing is effective when the bias is significant. All experiments are run on a single NVIDIA A100 GPU.

E Impact on Downstream Evaluation

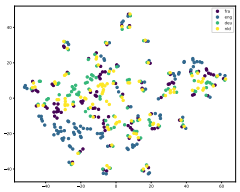
As already discussed in Meade et al. (2022), most of the models are relatively unaffected by debiasing (with debiasing techniques like SentDebias and INLP). Similarly, our preliminary experiments on GLUE (SST-2, RTE, and MRPC) (Wang et al., 2018) indicate that debiasing in the latent space does not significantly degrade model performance. The relative performance changes are minimal: only a 1.3% difference in F1 score for MRPC, 0.2% in accuracy for SST-2, and 0.77% in accuracy for RTE. These results are comparable to the effects observed when debiasing is applied in the original representation space.



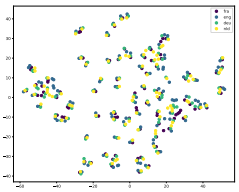
(a) Original Space



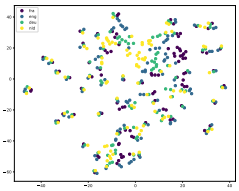
(b) AE with 1-layer MLP



(c) AE with 2-layer MLP

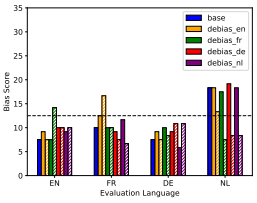


(d) AE with 3-layer MLP

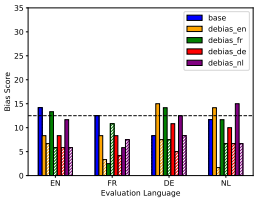


(e) AE with 4-layer MLP

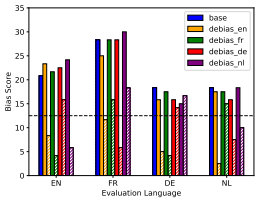
Figure 3: The t-SNE visualization of 100 parallel sentences across four languages taken from Flores+ in the original space and latent spaces induced by autoencoder with different hidden layers respectively.



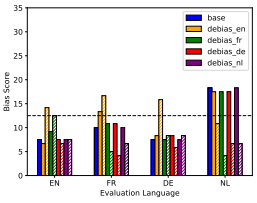
(a) INLP (Race)



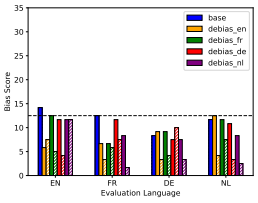
(b) INLP (Gender)



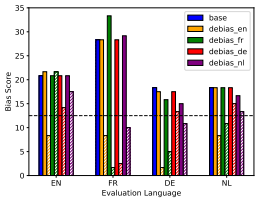
(c) INLP (Religion)



(d) SentDebias (Race)

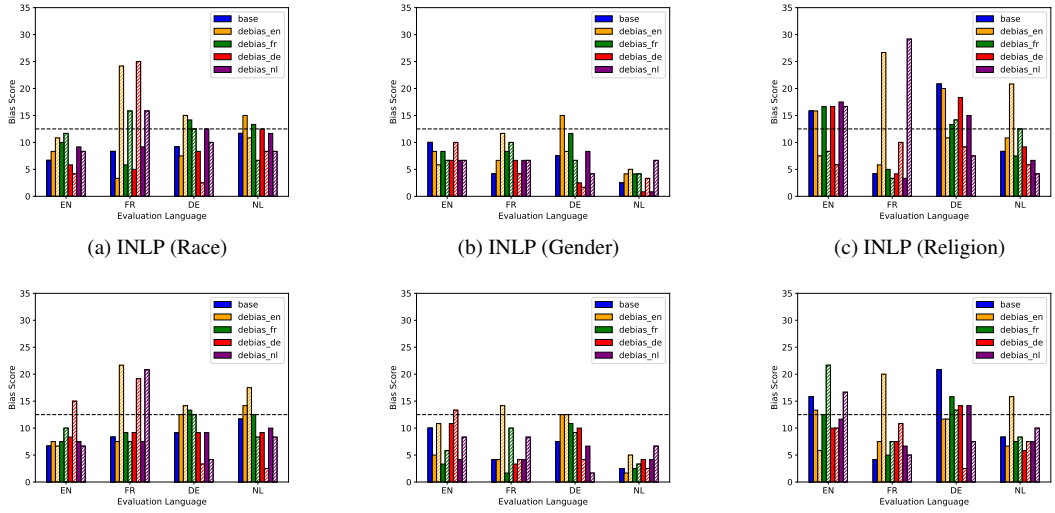


(e) SentDebias (Gender)



(f) SentDebias (Religion)

Figure 4: Bias score of Aya-expanse-8B on different evaluation language when applying INLP/SentDebias in the original space and latent space respectively, with different debiasing languages. The horizontal line at score of 12.5 refers to the significant threshold. Color represents different debiasing languages. Solid bar refers to the result of debiasing in original space and dashed bar refers to debiasing in the latent space. Blue bar refers to the base performance without debiasing.



(a) INLP (Race) (b) INLP (Gender) (c) INLP (Religion)

(d) SentDebias (Race) (e) SentDebias (Gender) (f) SentDebias (Religion)

Figure 5: Bias score of Llama3.1-8B-Instruct on different evaluation language when performing INLP/SentDebias in the original space and latent space respectively, with different debiasing languages. The horizontal line at score of 12.5 refers to the significant threshold. Color represents different debiasing languages. Solid bar refers to the result of debiasing in original space and dashed bar refers to debiasing in the latent space. Blue bar refers to the base performance without debiasing.