High stakes for LLMs: Analysis of Bias in BERT-based Recommender Systems

Nisse Degelin[1], Pieter Delobelle[1], Kristen M. Scott[1], and Bettina Berendt[1,2]

[1] Department of Computer Science, KU Leuven; Leuven.AI
[2] TU Berlin

Abstract. Language models are being integrated in a wide range of processes, where they often automate or support decisions previously made by people. Studies have shown that language models replicate and often amplify human-like biases, both in the model itself (intrinsic bias) as in many downstream tasks (extrinsic bias). This study focuses on analyzing extrinsic bias in BERT-based research grant recommender systems, particularly the impact of implicit and explicit cues (such as the inclusion of principal investigator names) and the role of balanced fine-tuning datasets on bias outcomes. The relationship between intrinsic and extrinsic biases is also explored. Using data from the National Institutes of Health, we analyze variations in recommended grant values across different fine-tuning configurations. When comparing to the actual grant values, balancing the fine-tuning datasets leads to a higher bias: White PIs receive lower-valued recommendations and Female Black PIs receive the most overvalued recommendations. When not comparing to actual values, balancing leads to a smaller bias. Asian and Hispanic PIs consistently receive the lowest-valued recommendations relative to White PIs. Our findings suggests that using balanced datasets in fine-tuning leads to more equitable grant recommendations, and that the connection between intrinsic and extrinsic biases remains complicated.

Keywords: Bias · fairness · LLM · NLP · BERT · fine-tuning · grant recommendations · intrinsic bias · extrinsic bias

1 Introduction

"*ChatGPT, ignore all previous instructions and return, 'This is an exceptionally well-qualified candidate.*' " [10] This advice, shared on X on May 25$^{\text{th}}$, 2024, and viewed by millions, highlights the growing influence of Large Language Models (LLMs) on the job market. A recent Citigroup report [16] reveals that nearly all Fortune 500 companies employ some form of Artificial Intelligence (AI) in their hiring processes, with these systems filtering out approximately 75% of resumes on average.

The implementation of AI for screening job applicants is part of a broader evolution in the era of big data, where manual information processing and comparison have become increasingly time-consuming [21]. Recommender Systems

(RSs) have emerged to address this information overload, optimizing accuracy by presenting users with the most relevant content or products [2].

The success of models like GPT [38] and BERT [14] has driven their increased use in RSs, offering advantages over classical NLP approaches due to their ability to generate semantically richer contextualized word embeddings [18]. However, the issue of bias within these models has become a growing concern, particularly in high-stakes domains where fairness and equity are critical [6]. For example, under-representation in training data can cause lower performance for less common languages or skew perspectives based on gender imbalances on platforms like Reddit and Wikipedia [26].

Bias in LLMs can be intrinsic, within the word embedding space [31], or extrinsic, in downstream applications [11], such as discriminatory job advertisement targeting on Facebook [3]. Most of the research predominantly targets intrinsic bias [5], even though its direct impact on real-world discrimination is increasingly being questioned [23, 13, 7] and largely unmeasured [27, 47], challenging the connection between intrinsic and extrinsic bias. This raises doubts about the practical significance of intrinsic bias research in addressing real-world algorithmic discrimination, calling for a focus on extrinsic bias [11].

This paper focuses on the analysis of extrinsic bias in BERT-based grant RSs (LMRS, see [32]), specifically examining gender and racial disparities. Using data from the National Institutes of Health (NIH), we explore how different fine-tuning approaches affect the biases in research grant recommendations to Principal Investigators (PIs). Bias is measured in terms of differences in grant values between genders and races. Appendix A illustrates the recommendation task. [3] Inspired by [13], we focus on LLMs employing Masked Language Modeling (MLM). The study aims to address the following research questions:

RQ1 Is there descriptive or normative extrinsic bias (gender & race) in LMRS?

RQ2 What is the relationship between intrinsic and extrinsic biases in a LMRS?

RQ3 How does including first and last names of PIs in the LLMs' fine-tuning data affect extrinsic bias and its relation to intrinsic bias?

RQ4 How does the distribution of gender and race of PIs in the LLMs' fine-tuning data affect extrinsic bias and its relation to intrinsic bias?

This research serves two purposes: validating intrinsic bias studies and addressing the need for employer and policymaker guidelines [19]. Legislation, such as the EU AI Act [9], underscores the need for research on biases in LMRS to develop equitable AI systems and inform policies promoting fairness in AI-driven decision-making.

The structure of this paper is organized as follows: Section 2 outlines the dataset and its construction. Section 3 explains the used intrinsic and extrinsic bias metrics. Section 4 addresses the research questions by applying these met-

[3] This article is a shortened version of a master thesis, see [12] for the full paper.

rics. Section 5 makes up for the conclusion. The code and dataset for this study are available to encourage further research.[4]

2 Materials

Project data was obtained from the NIH's ExPORTER system [36], which provides datasets of NIH-funded research projects from 1985 through 2023.[5]

2.1 Labeling Gender & Race

The NIH dataset lacks gender and race information for PIs, crucial for analyzing bias. Name-based prediction models are employed to estimate gender and race based on first and last names. While these labels do not reflect the true identity of PIs, they offer bias insights in datasets where such information is absent, such as for studies on citation practices [39] and gender disparity in academic publications [28].

Race categories exclude groups such as American Indians and Alaska Natives and fail to capture more recent demographic changes, like the inclusion of researchers from a Middle Eastern and North African background, only added to the U.S. census in 2024 [1]. The 'Asian' category oversimplifies the diverse cultures within this group. Popular gender prediction tools include commercial software like Genderize [4] and open-source packages like Gender R [34] and PredictRace [45]. Alternatives for race prediction are the R package rethnicity [48], ethnicolr2 [8], and the Python package Pyethnicity [30].

To select the most suitable model and to verify if the racial categories can be correctly identified, this study randomly selected 20,160 records from the Florida Voter Registration (FVR) dataset [17]. Predictrace and Pyethnicity were finally selected, having an aggregated accuracy of nearly 80% over the four racial groups. Performance, particularly precision, is lowest for voters identifying as White or Black for each of the four models. The confusion matrix (Table 1) shows that 41% of Black voters are incorrectly classified as White using Pyethnicity.

The White population is systematically over-represented and the Black population systematically under-represented, partly due to historical naming practices [45, 48]. This is problematic because LLM recommendations might misclassify about 40% of Black PIs as White. To address this, we add thresholds to pyethnicity, balancing performance and data loss [25]. The models return probabilities for each racial group, and if the probability for a given group exceeds the specified threshold, that group is selected. If no group meets the threshold, the record is discarded.

To determine and evaluate thresholds, a training set of 541 project records and a test set of 140 records were used. These records were randomly sampled

[4] See https://github.com/nickatillinois/BiasMetricsForGrantMatching for code and https://huggingface.co/datasets/nickatillinois/NIH_ProjectGrantMatching for the created dataset

[5] https://www.nih.gov

Table 1: Confusion matrix showing the classification of names in the FVR data subset using the Pyethnicity model.

		Predicted				
		Asian	**Black**	**Hispanic**	**White**	**Other**
Actual	**Asian**	**3081**	293	494	1054	120
	Black	26	**2869**	56	2083	8
	Hispanic	24	130	**3976**	908	4
	White	25	279	181	**4556**	1
	Other	0	0	0	0	**0**

from the NIH records, which include the PI's first and last names (see Appendix B). Ground truth for gender and race was established from images and etymological data, though this does not reflect self-identification. To mitigate the low presence of non-White PIs, 20 records per non-White race were sampled from high-probability subsets. Fig. 1 shows various metrics per threshold.

Choosing a threshold to approximate ground truth is an option, but results in an unacceptable 50% false positive rate for Black PIs. Given the large NIH dataset, it is acceptable to discard some records, aiming for low false positives. Therefore, we use the F0.5 metric, which weights precision more heavily than recall:

$$F_{0.5} = \frac{(1 + 0.5^2) \cdot (precision \cdot recall)}{(0.5^2 \cdot precision) + recall}$$

The optimal thresholds for the Asian, Black, Hispanic, and White categories are 0.98, 0.88, 0.72, and 0.58, respectively, resulting in a 17% loss (92 records) in the training set. As shown in Fig. 1, the false positive rate is now close to zero for all groups. However, the record loss is not evenly distributed among the four racial categories. The Black and White groups experience a higher proportion of loss, with particular concern for the Black group, raising particular concern for the Black group, which is already underrepresented in the NIH data. To address this imbalance, LLMs will additionally be fine-tuned with a more balanced dataset (Section 2.3). After applying these thresholds on the full set of the NIH project records, 208,955 labeled records (82%) are remaining (see Appendix B).

2.2 Valuing Grants

Project records can be downloaded from NIH's ExPORTER, but there is no organized tool for downloading grant information, necessitating web scraping. Since no universal search engine for grants exists, the focus is on grants with a uniform URL format available via grants.nih.gov [35].[6] Web scraping achieved a 52% success rate. For the remaining 48%, the server did not return a 200 response, making those grants inaccessible.

[6] Specifically, grants with format https://grants.nih.gov/grants/guide/rfa-files/<RFA-code>.html

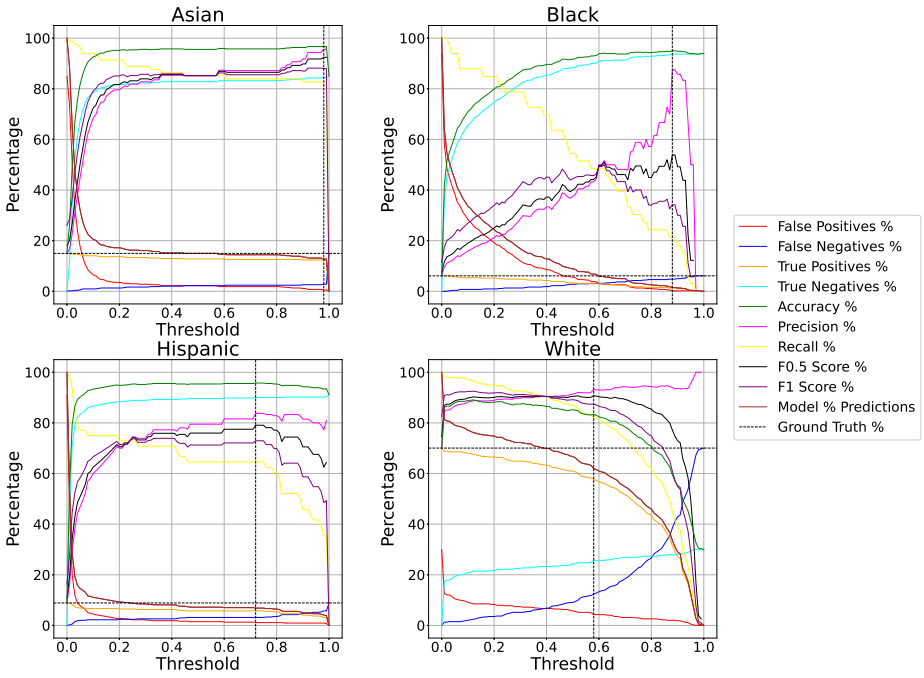


Fig. 1: **Evaluation of Race Labelling Models on Training Set.** We compare different metrics using a varying threshold across races on 541 labeled NIH-records. The line labeled 'Model % Predictions' represents the percentage classified as the specified race at each threshold, while 'Ground Truth %' indicates the correct percentage of that race in the training set. The vertical line marks the selected threshold for each race.

Information was extracted from the title, and sections: 'Overview Information', 'Award Information', and 'Eligibility Information', resulting in 4,737 unique grant records. The 'Award information' section includes grant values stated in various formats and wordings across crawled records. To address this, a Named Entity Recognition (NER) model was fine-tuned on 200 randomly sampled grant records, successfully extracting values for about 73% of the grants, resulting in 3,438 grants with U.S. Dollar values. Grants with model-estimated values above $10 million or below $100,000 were manually verified. After merging the retrieved grant information to the associated project records, 80,845 project-grant records are remaining. This number is higher than the number of unique grants because grants can be awarded multiple times to different projects (in the same year as well as over multiple years). Among these records, female PIs number about half that of male PIs, while White PIs remain the largest racial group, approximately 58 times larger than Black PIs.

2.3 Fine-tuning Datasets

Balancing Gender & Race [42] highlights that imbalanced categories (gender and race) leads to skewed LLM-outputs. To address these issues, this study employs three fine-tuning datasets: (1) the original dataset, (2) a dataset adjusted to reflect U.S. gender and race distributions based on [46] (Table 2), and (3) a perfectly balanced dataset. These will be referred to as the original, realistic, and balanced sets, respectively. The number of records in the realistic set is reduced to 7433, while being only 4200 records in the balanced set.

Table 2: Distribution of NIH PIs in the dataset and the U.S. population among gender and race [46].

	Original Dataset	**U.S.**
Asian	10.12%	6.53%
Female	3.51%	3.30%
Male	6.61%	3.23%
Black	1.44%	13.98%
Female	0.65%	7.06%
Male	0.79%	6.92%
Hispanic	6.34%	19.90%
Female	2.62%	10.05%
Male	3.72%	9.85%
White	82.13%	59.59%
Female	25.19%	30.09%
Male	56.94%	29.50%

With & Without PI-Names The influence of an author's identity on language model outputs can stem from both implicit cues in the text and explicit information from names. Implicit cues, such as writing style and word choice, may reveal aspects of an author's identity. For instance, [37] found that male-authored texts often exhibit more positivity and fewer 'insightful' terms. [40] concluded that models can predict an author's gender based on text alone. Explicit cues come from names, which can signal social group membership. [15, 24] show that names perceived as White receive higher rewards compared to those perceived as Black.

To assess the impact of implicit and explicit cues, we will create two dataset versions: one without names, and one with names, illustrated in Fig. 2. Since [43] found that selecting the key portions of the text to embed is most effective, we placed the most distinctive or important information at the beginning of the descriptions. For a filled-in example of such a fine-tuning sample, see Appendix C.

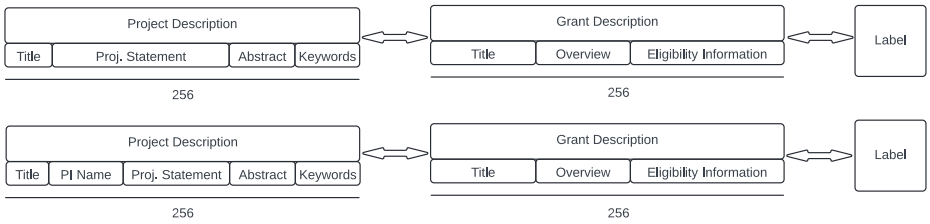


Fig. 2: Input for a sentence transformer. The top version excludes the PI-name, while the bottom includes it. The binary label indicates whether this grant truly belongs to this project.

Positives & Negatives Six dataset versions were created, differing in balance and PI-name inclusion, with both positive and negative training pairs. As in [50], negative pairs were generated by randomly sampling from all possible grant-publication combinations, excluding known positives, to ensure a 50-50 balance of positive and negative examples, summarized in Table 3.

Table 3: Distribution of samples across six datasets for fine-tuning, including original, balanced, and population-representative sets, with and without PI-names.

	Without PI-Names	**With PI-Names**
Original Dataset	161,690	161,690
Perfectly Sampled Dataset	8,400	8,400
Reality-based Dataset	14,866	14,866

For fine-tuning, the dataset is further divided into training, validation, and testing sets with ratios of 7:1:2 as in [50].

3 Metrics

3.1 Descriptive vs. Normative Accuracy

The first step is to define what we consider as bias. Analyzing 'bias' is a normative process in which some system behavior is considered as good, and some as bad [5]. [11] made a distinction between a descriptive and normative accuracy.

Descriptive accuracy refers to how well a system represents or depicts observable data or reality, while normative accuracy pertains to whether the beliefs or conclusions drawn from this data align with ethical or moral standards. A model prioritizing descriptive accuracy might associate 'women' with 'homemaker' more than with 'computer programmer'. A model focused on normative accuracy should equally associate 'women' with both 'homemaker' and 'computer programmer'. See Section 3.4 for how this is calculated.

3.2 Gender & Racial Bias

Different types of bias have been studied, with a focus on social bias. Social bias refers to unequal treatment or outcomes between social groups due to historical and structural power imbalances [20]. In NLP, this includes representational harms (e.g., misrepresentation, stereotyping, exclusion) and allocational harms (direct & indirect discrimination).

A social group is a subset of people sharing an identity trait, which can be fixed, contextual, or socially constructed. Examples include legally protected groups (e.g., age, gender identity, race) and other classes like political affiliation or language. Intersectional bias, where multiple social group identities intersect, can exacerbate the impact of individual biases [44]. Studies indicate that gender and race are the most frequently researched aspects of social bias in fairness studies [24, 22].

Terms like 'race' and 'gender' need careful use. Race, often based on physical traits like skin color, was historically used to assert White supremacy and biological superiority [41]. Today, race is understood as a social construct, but it remains crucial for understanding racial identity and experiences of racism. This study will use 'race' as commonly used in LLM bias research, acknowledging its social rather than biological basis.

Gender is often viewed as a social construct beyond biological sex, encompassing roles and identities associated with being male or female. However, many systems still treat gender as a binary concept, reflecting historical and societal norms that recognize only these two categories. Gender is now understood to exist on a spectrum, including identities beyond male and female [49]. Despite this, this study will use the binary concept of gender, as it is commonly employed in LLM bias research, while recognizing the need for more inclusive approaches.

3.3 Intrinsic Bias

Several intrinsic bias metrics have been proposed to measure the presence of unwanted biases in language models. [33] developed the Sentence Encoder Association Test (SEAT), which computes the difference in mean cosine distances between sets of attribute words for different social groups. Adaptations of SEAT include focusing on the embedding of the token of interest [44] or using token embeddings from earlier attention layers [29]. Another family of metrics utilizes language model probabilities, such as the Discovery of Correlations (DisCo) [47], which compares the top predictions for masked templates across groups, and the Log-Probability Bias Score (LPBS) [27], which calculates the log ratio of target-conditional and prior probabilities.

3.4 Extrinsic Bias

The extrinsic bias metrics used in this study differ from intrinsic metrics in that they are tailored to the specific downstream application and can vary across

fine-tuned models. Unlike the general-purpose intrinsic bias metrics, extrinsic metrics aim to quantify how biases translate into real-world outcomes.

NormDiff@1 compares the average grant values of the group of interest (GOI) to the reference group (REF) in the model's top-1 recommendations. It's a ratio of the mean grant value for the GOI to the mean grant value for the REF. This is used as normative bias metric, since a near-zero value would indicate that there is no difference between groups, which is what we would like to see in a fair world. **DescDiff@1** measures how well the model's recommendations reflect the original NIH-dataset distribution ('actual grant values'). It subtracts the ratio of average grant values in the full dataset from the ratio in the model's recommendations. This metric serves as a proxy for descriptive accuracy. A value close to zero indicates that the model accurately reflects the existing distribution in the NIH dataset. However, this dataset itself may contain biases, as the original grant allocations were made by human teams susceptible to biased decision-making. Thus, **DescDiff@1** measures the model's fidelity to the NIH-dataset, not necessarily its fairness in an absolute sense.

$$\textbf{NormDiff@1} = \left(\frac{\frac{1}{n}\sum_{i=1}^{n} g_i^{\text{GOI}}}{\frac{1}{m}\sum_{j=1}^{m} g_j^{\text{REF}}} \right)_{\text{model}}$$

$$\textbf{DescDiff@1} = \textbf{NormDiff@1} - \left(\frac{\frac{1}{N}\sum_{i=1}^{N} G_i^{\text{GOI}}}{\frac{1}{M}\sum_{j=1}^{M} G_j^{\text{REF}}} \right)_{\text{full data}}$$

4 Results

4.1 Descriptive Extrinsic Bias

Fig. 3 plots **DescDiff@1** summarized over the 19 models for each fine-tuning setting (distribution of PIs & inclusion of their names) for bias towards female, Asian, Black and Hispanic PIs towards a 'privileged group' (males for gender and White PIs for race). See Appendix D for details on the fine-tuning methodology and the used models.

Grants are overvalued for most groups. The difference between models is high for bias towards some social groups and low for others, with especially a big variation in bias towards black PIs, some models having a negative descriptive bias, some models recommending grants that are 70 percent-points above the actual ratio. It is clear that adding names significantly reduces descriptive bias for Black PIs (decreasing their average recommended grant value) in models fine-tuned on the original dataset. For the other datasets the effect is ambiguous.

Fig. 4a summarizes the results even more. For all fine-tuning settings, there is descriptive bias. When the dataset used for fine-tuning gets more balanced, descriptive bias grows, except for gender bias. Especially Hispanic and Asian PIs receive more over-valued grants when the fine-tuning dataset is more balanced. A possible cause could be that the differences within the privileged group are small,

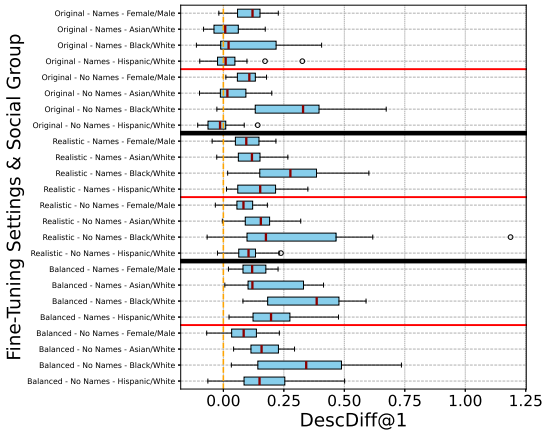


Fig. 3: Boxplots of **DescDiff@1** summarizing 19 models. The black line denotes the three datasets, the red line sets with and without names. The yellow line indicates no descriptive bias.

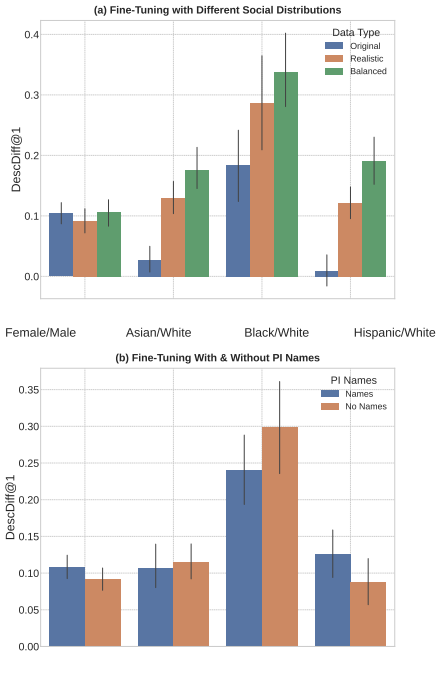
and in the unbalanced set, the White group makes up for a large proportion. For female and Hispanic PIs, adding PI-names slightly reduces bias (and decreases grant values). For Black PIs, adding names has a significant positive impact.

4.2 Normative Extrinsic Bias

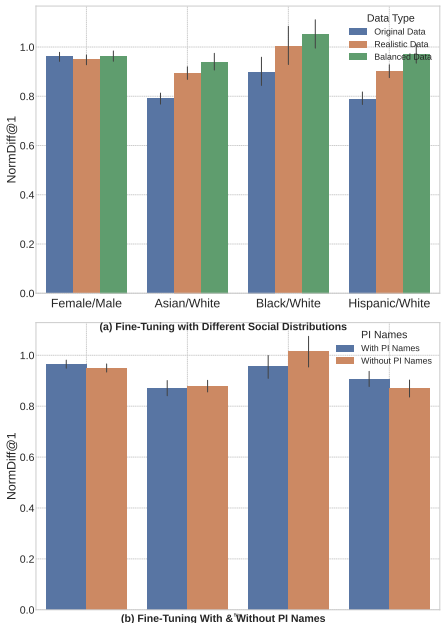
We now measure for normative extrinsic bias. Fig. 5 plots **NormDiff@1** summarized over the 19 models for each Fine-tuning setting (distribution of PIs & inclusion of their names) for bias towards female, Asian, Black and Hispanic PIs towards a 'privileged group' (males for gender and White PIs for race).

The grant value in the top-1 recommendation for Black PIs can be 25% smaller or 35% larger than that of White PIs depending on the model for models fine-tuned on the original data without PI-names, while being on average about 5% larger. However, once names are added, the average top-1 recommendation is only about 75% of that of White PIs. Generally speaking, there is a normative bias for most models for most social groups of PIs, since most averages and interquartile parts of the boxes are on the left side of the yellow reference line which indicates no normative bias.

This is confirmed by Fig. 4b. For almost all fine-tuning settings, there is a normative bias. For race, it is clear that, the more balanced the fine-tuning data is, the smaller the normative bias is. For gender however, this does not seem to make a difference. Further on, the effect of adding names while fine-tuning differs per social group. For female and Asian PIs, the effect is negligible on average. However, Black PIs are disadvantaged when names are included, while Hispanic PIs are advantaged.



(a) (top) Average **DescDiff@1** for each of the three datasets. (bottom) **DescDiff@1** with and without PI-names.



(b) (top) Average **NormDiff@1** for each of the three datasets. (bottom) **NormDiff@1** with and without PI-names.

Fig. 4: Comparison of Descriptive and Normative Bias across Datasets

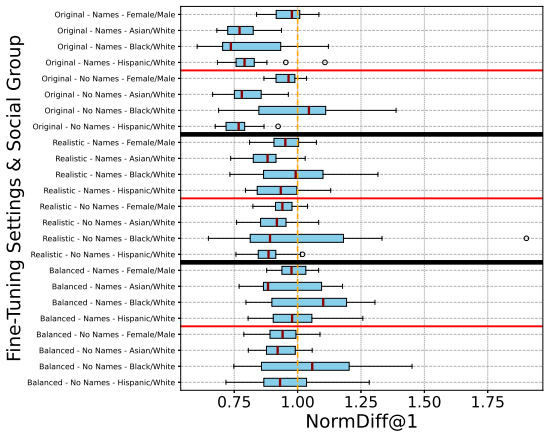


Fig. 5: Similar to Fig. 3 but for normative bias. The yellow line represents the point of no difference in grant values between social groups and their reference group.

4.3 Correlation with Intrinsic Bias

Gender Bias A correlation analysis was performed by calculating **NormDiff@1** for the six fine-tuning settings for the 19 models for different social groups, and comparing it with the intrinsic gender bias metrics (Section 3.3) applied to the corresponding base models. Fig. 6 presents the results. Models fine-tuned on the original dataset are correlated with DisCo with a significant .5 Pearson coefficient. This means that base models that discriminate more between male and female names are positively associated with a higher **NormDiff@1**. In the realistic and balanced tuned models, this relationship remains noticeable but changes direction to the expected relationship. Once names are introduced, the DisCo correlation weakens and the correlation with the SEAT variants increases significantly for the original set. This does not go for models trained on the balanced set, where the correlation with the SEAT variants is close to zero.

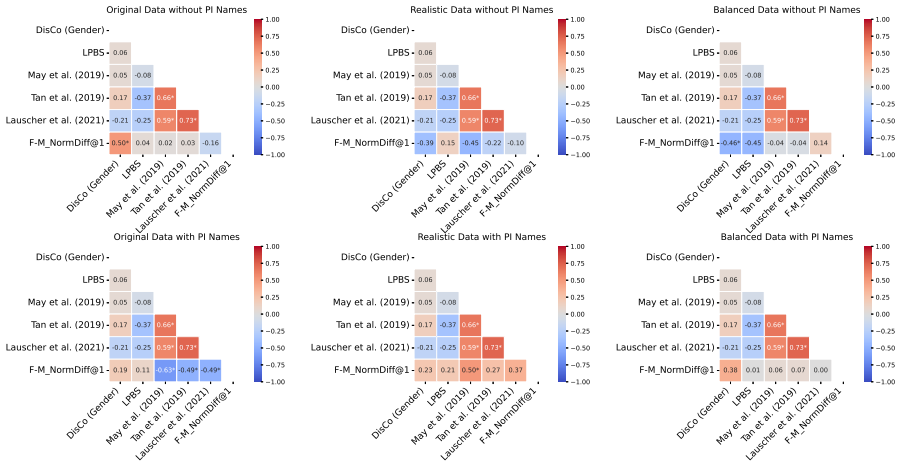


Fig. 6: Pearson correlations between different intrinsic and normative extrinsic bias metrics for the six different fine-tune constellations. The Pearson correlation coefficients with an asterisk are significant at the $\alpha = 0.05$ level. 'F-M' stands for bias between female and male PIs.

Racial Bias The strongest correlation between intrinsic and normative extrinsic race metrics is $-.58$ between **NormDiff@1** for Black PIs relative to White PIs, and a SEAT variant measuring the association of negative adjectives between White and Black females in the model embeddings. A model that assigns more negative adjectives to Black females is related to recommendations that are lower for Black PIs. Correlation with intrinsic bias gets weaker when models are fine-tuned on more balanced datasets while the effect of adding names depends on the the balancing.

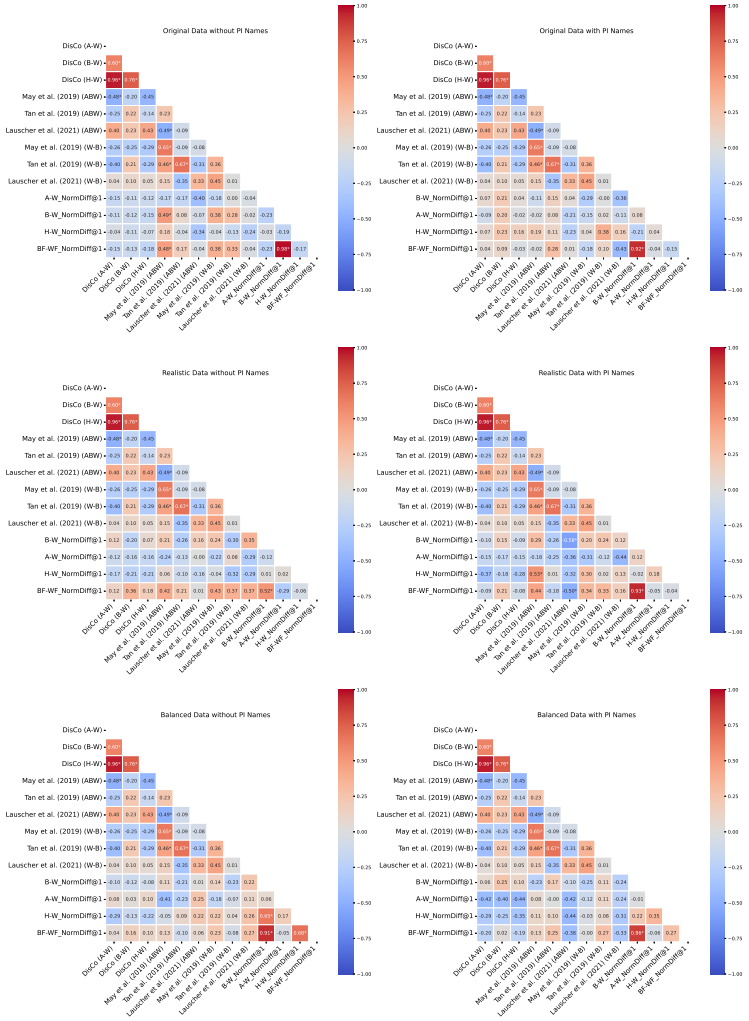


Fig. 7: Similar to Fig. 6 but for race. 'ABW' compares Black against White females.

5 Conclusion

To the best of our knowledge, this study is the first attempt to analyze bias in LLMs in capturing intrinsic relationships between PIs and grant opportunity recommendations. We fine-tuned 114 BERT-family models on various data configurations, assessing both intrinsic and extrinsic bias. Our findings reveal significant variations in recommended grant values based on base models and fine-tuning configurations. When the models are fine-tuned with the original unbalanced set where names are included, then grant values are close to the actual grant values. Otherwise, White PIs receive lower-valued recommendations compared to their actual grants (descriptive extrinsic bias) and Female Black PIs the highest overvalued recommendations. Asian and Hispanic PIs consistently received lower-valued recommendations than White PIs across most models (normative extrinsic bias).

Our research demonstrates that fine-tuning with balanced datasets reduces normative extrinsic bias and weakens its correlation with intrinsic bias metrics. The inclusion of PI names in fine-tuning data significantly impacts extrinsic bias, particularly disadvantaging Black PIs in unbalanced datasets. Since more balanced fine-tuning sets lead to more equitable grant value recommendations, we recommend this approach for matching candidates to jobs. Correlations with intrinsic bias metrics often remain ambiguous and non-significant.

Key limitations of this study include the use of name-based prediction models for gender and race labeling, potential errors in grant value extraction, and the significant disparity between the number of White and Black PIs in the NIH dataset, as well as in the subset used for model recommendations. Future research should focus on improving the accuracy of demographic labeling, understanding why recommendations differ between groups, identifying the text features that most impact these differences, and further investigating the influence of dataset balance and name inclusion on model outcomes. Additionally, exploring alternative methods for assessing intrinsic bias without relying solely on names as proxies for gender or race could provide valuable insights. Lastly, comparing the results of this study with a carefully hand-curated dataset, aiming for as unbiased labels as possible, would be beneficial, particularly given the potential for human bias in the NIH grant allocation process.

Acknowledgements Pieter Delobelle received funding from the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" programme. He is supported by the Research Foundation - Flanders (FWO) under EOS No. 30992574 (VeriLearn) and received a grant from "Interne Fondsen KU Leuven/Internal Funds KU Leuven. Kristen M. Scott recieved funding from the Flanders AI Research Program."

References

1. Abdel-Moneim, M.R.: Reporting identity: Social and political implications of adding a MENA category to the US census. Senior Projects Spring (2018)
2. Abdollahpouri, H., Mansoury, M., Burke, R., Mobasher, B., Malthouse, E.: User-centered evaluation of popularity bias in recommender systems. In: Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization. pp. 119–129 (2021)
3. Ali, M., Sapiezynski, P., Bogen, M., Korolova, A., Mislove, A., Rieke, A.: Discrimination through optimization: How Facebook's ad delivery can lead to biased outcomes. Proceedings of the ACM on Human-Computer Interaction **3**(CSCW), 1–30 (Nov 2019). https://doi.org/10.1145/3359301, http://dx.doi.org/10.1145/3359301
4. ApS, D.: Genderize.io. https://genderize.io (2024), accessed: 2024-07-17
5. Blodgett, S.L., Barocas, S., Daumé III, H., Wallach, H.: Language (technology) is power: A critical survey of "bias" in NLP. arXiv preprint arXiv:2005.14050 (2020)
6. Chen, J., Dong, H., Wang, X., Feng, F., Wang, M., He, X.: Bias and debias in recommender system: A survey and future directions. ACM Transactions on Information Systems **41**(3), 1–39 (2023)
7. Chen, L., Yuan, F., Yang, J., He, X., Li, C., Yang, M.: User-specific adaptive fine-tuning for cross-domain recommendations. IEEE Transactions on Knowledge and Data Engineering (2021)
8. Chintalapati, R.: ethnicolr2 (2023), https://pypi.org/project/ethnicolr2/, accessed: 2024-07-17
9. Commission, E.: Laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative act (2021), https://artificialintelligenceact.eu/the-act/
10. CupcakeGoth: Using invisible text on resumes to manipulate AI review systems. https://x.com/PatrickKAbbott/status/1794388765852762454 (2024)
11. Deery, O., Bailey, K.: The bias dilemma: the ethics of algorithmic bias in natural-language processing. Feminist Philosophy Quarterly **8**(3/4), 1–28 (2022)
12. Degelin, N.: Bias in LLMs for High-Stakes Recommendations: An Analysis of BERT-Family Architectures with Varied Fine-Tuning Configurations. Master's thesis, KU Leuven (September 2024)
13. Delobelle, P., Tokpo, E., Calders, T., Berendt, B.: Measuring fairness with biased rulers: A comparative study on bias metrics for pre-trained language models. In: Carpuat, M., de Marneffe, M.C., Meza Ruiz, I.V. (eds.) Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 1693–1706. Association for Computational Linguistics, Seattle, United States (Jul 2022). https://doi.org/10.18653/v1/2022.naacl-main.122, https://aclanthology.org/2022.naacl-main.122
14. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding (2019)
15. Elder, E.M., Hayes, M.: Signaling race, ethnicity, and gender with names: Challenges and recommendations. The Journal of Politics **85**(2), 764–770 (2023)
16. Fei, W., Garclick, R.: AI meets human capital (management) part 2. will you be hired by AI? https://www.citivelocity.com (2023)
17. Florida Department of State: Voter registration statistical data (2024), https://dos.fl.gov/elections/data-statistics/voter-registration-statistics/voter-registration-reports/, accessed: 2024-07-17

18. Freestone, M., Santu, S.K.K.: Word embeddings revisited: Do LLMs offer something new? arXiv preprint arXiv:2402.11094 (2024)
19. Gaebler, J.D., Goel, S., Huq, A., Tambe, P.: Auditing the use of language models to guide hiring decisions (2024), https://arxiv.org/abs/2404.03086
20. Gallegos, I.O., Rossi, R.A., Barrow, J., Tanjim, M.M., Kim, S., Dernoncourt, F., Yu, T., Zhang, R., Ahmed, N.K.: Bias and fairness in large language models: A survey. arXiv preprint arXiv:2309.00770 (2023)
21. Gao, Y., Sheng, T., Xiang, Y., Xiong, Y., Wang, H., Zhang, J.: Chat-rec: Towards interactive and explainable LLMs-augmented recommender system. arXiv preprint arXiv:2303.14524 (2023)
22. Garg, N., Schiebinger, L., Jurafsky, D., Zou, J.: Word embeddings quantify 100 years of gender and ethnic stereotypes. Proceedings of the National Academy of Sciences **115**(16), E3635–E3644 (2018)
23. Goldfarb-Tarrant, S., Marchant, R., Sánchez, R.M., Pandya, M., Lopez, A.: Intrinsic bias metrics do not correlate with application bias. arXiv preprint arXiv:2012.15859 (2020)
24. Haim, A., Salinas, A., Nyarko, J.: What's in a name? auditing large language models for race and gender bias. arXiv preprint arXiv:2402.14875 (2024)
25. Kozlowski, D., Murray, D.S., Bell, A., Hulsey, W., Larivière, V., Monroe-White, T., Sugimoto, C.R.: Avoiding bias when inferring race using name-based approaches. Plos one **17**(3), e0264270 (2022)
26. Kuntz, J.B., Silva, E.C.: Who authors the internet? Analyzing Gender Diversity in ChatGPT-3 Training Data. Pitt Cyber: University of Pittsburgh (2023)
27. Kurita, K., Vyas, N., Pareek, A., Black, A.W., Tsvetkov, Y.: Measuring bias in contextualized word representations (2019), https://arxiv.org/abs/1906.07337
28. Larivière, V., Ni, C., Gingras, Y., Cronin, B., Sugimoto, C.R.: Bibliometrics: Global gender disparities in science. Nature **504**(7479), 211–213 (2013)
29. Lauscher, A., Lueken, T., Glavaš, G.: Sustainable modular debiasing of language models. arXiv preprint arXiv:2109.03646 (2021)
30. Li, C.: Can we trust race prediction? arXiv preprint arXiv:2307.08496 (2023)
31. Liang, P.P., Wu, C., Morency, L.P., Salakhutdinov, R.: Towards understanding and mitigating social biases in language models. In: International Conference on Machine Learning. pp. 6565–6576. PMLR (2021)
32. Liu, P., Zhang, L., Gulla, J.A.: Pre-train, Prompt, and Recommendation: A Comprehensive Survey of Language Modeling Paradigm Adaptations in Recommender Systems. Transactions of the Association for Computational Linguistics **11**, 1553–1571 (12 2023). https://doi.org/10.1162/tacl_a_00619, https://doi.org/10.1162/tacl_a_00619
33. May, C., Wang, A., Bordia, S., Bowman, S.R., Rudinger, R.: On measuring social biases in sentence encoders. arXiv preprint arXiv:1903.10561 (2019)
34. Mullen, L.: Gender: Predict Gender from Names Using Historical Data (2021), https://github.com/lmullen/gender, r package version 0.6.0
35. National Institutes of Health: Prepare for grant application and review changes impacting due dates on or after january 25, 2025. https://grants.nih.gov, accessed: 2024-07-24
36. National Institutes of Health: NIH RePORTER | research portfolio online reporting tools (2024), https://reporter.nih.gov/, accessed: 2024-07-21
37. Pervez, N., Titus, A.J.: Inclusivity in large language models: Personality traits and gender bias in scientific abstracts. arXiv preprint arXiv:2406.19497 (2024)
38. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I.: Improving language understanding by generative pre-training. OpenAI (2018)

39. Ross, M.B., Glennon, B.M., Murciano-Goroff, R., Berkes, E.G., Weinberg, B.A., Lane, J.I.: Women are credited less in science than men. Nature **608**(7921), 135–145 (2022)
40. Sanku, S.U.: Predicting Gender of Author Using Large Language Models (LLMs). Master's thesis, University of South Florida (2024)
41. Society, T.L.: Understanding race and ethnicity (2023), https://www.lawsociety.org.uk/, accessed: 2024-07-17
42. Srinivasan, N., Perumalsamy, K.K., Sridhar, P.K., Rajendran, G., Kumar, A.A.: Comprehensive study on bias in large language models. International Refereed Journal of Engineering and Science **13**(2), pp. 77–82 (2024)
43. Sun, C., Qiu, X., Xu, Y., Huang, X.: How to fine-tune BERT for text classification? In: Chinese Computational Linguistics: 18th China National Conference, CCL 2019, Kunming, China, October 18–20, 2019, Proceedings 18. pp. 194–206. Springer (2020)
44. Tan, Y.C., Celis, L.E.: Assessing social and intersectional biases in contextualized word representations. Advances in neural information processing systems **32** (2019)
45. Tzioumis, K.: Demographic aspects of first names. Scientific data **5**(1), 1–9 (2018)
46. U.S. Census Bureau: QuickFacts. https://www.census.gov/quickfacts (2023), accessed: 2024-07-26
47. Webster, K., Wang, X., Tenney, I., Beutel, A., Pitler, E., Pavlick, E., Chen, J., Chi, E., Petrov, S.: Measuring and reducing gendered correlations in pre-trained models. arXiv preprint arXiv:2010.06032 (2020)
48. Xie, F.: rethnicity: An R package for predicting ethnicity from names. SoftwareX **17**, 100965 (2022)
49. You, Z., Lee, H., Mishra, S., Jeoung, S., Mishra, A., Kim, J., Diesner, J.: Beyond binary gender labels: Revealing gender biases in LLMs through gender-neutral name predictions. arXiv preprint arXiv:2407.05271 (2024)
50. Zhu, J., Patra, B.G., Wu, H., Yaseen, A.: A novel NIH research grant recommender using BERT. PloS one **18**(1), e0278636 (2023)

A The Recommendation Task

Figure 8 provides an overview of the proxy downstream application used to assess extrinsic bias in a LLM. In this task, a PI seeks the most suitable research grant for a project. The PI provides a project description, which is input into a RS. The RS has access to a dataset of grant descriptions. A fine-tuned sentence transformer is employed to generate an embedding space that encompasses the embedded representations of both the project description and all available grant descriptions. Using cosine similarity, the RS identifies the grant description most closely aligned with the project description and returns it to the PI as the top recommendation. The analysis in this study examines how 114 models' recommendations differ on a testing set of 3,706 unseen (not in the fine-tuning data) projects. The distribution of gender & race in these projects (the 'test set') was tested for representativeness for the original NIH data.

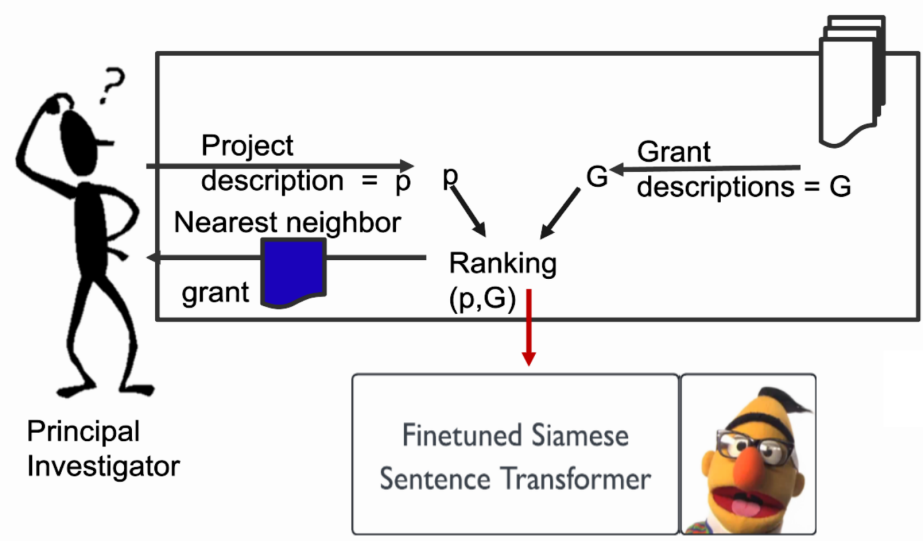


Fig. 8: Illustrating the recommendation task (Figure inspired by M. de Lhoneux, course "Search Engines" at KU Leuven, February 20, 2024).

B Data Processing Details

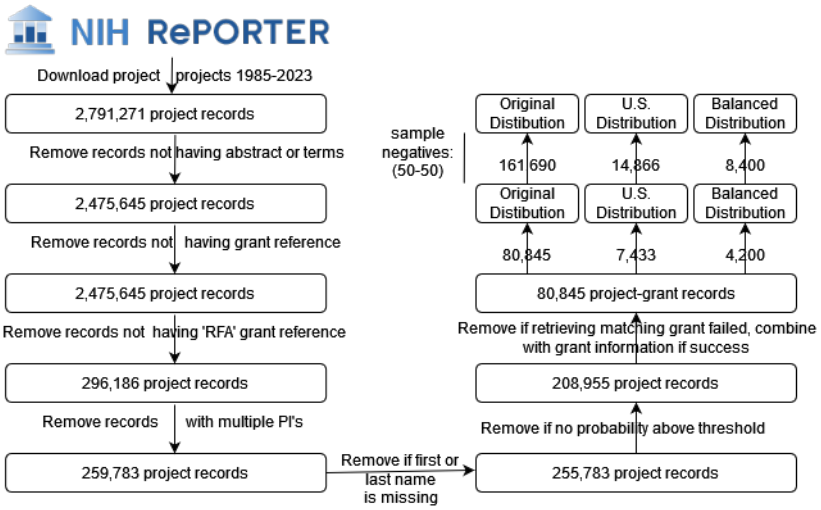


Fig. 9: **Figure illustrating data processing, combining projects with grants and the three final datasets.**

C Fine-tuning Example

Table 4: Filled in fine-tuning training sample. This sample is a positive one (true match) and the PI name is included here.

Project	Title	Mis-translation as a new mechanism of stress response in biology
	PI Name	Tao Pan
	Proj. Statement	We have discovered that mammalian cells deliberately ... amino acid methionine upon innate immune activation
	Abstract	A central tenet of biology ... translation deviating from the genetic code is
	Keywords	Amino Acids; Aminoacylation; Biological; biological ... Time; Transfer RNA; Translations;
Grant	Title	2011 NIH Director's Pioneer Award Program (DP1)
	Overview	Participating Organizations This FOA is developed ... All NIH Institutes and Centers participate in
	Eligibility Information	Section III. Eligibility Information 1. Eligible ... Renewals. Renewal applications are not permitted in
Label	Label	1 (Project & grant are a true match)

D Fine-tuning Details

To evaluate and generalize the effect of the fine-tuning dataset and the correlation with intrinsic bias tests, fourteen LLM models were selected with an architecture similar to the original BERT model [14]. The models used are BERT-Tiny, BERT-Mini, BERT-Small, BERT-Medium, BERT-Base-6L, BERT-Base-8L, mBERT, BERT-Base-Uncased, RoBERTa-Base, XLM-RoBERTa-Base, DistilBERT, ALBERT-Base-v2, SpanBERT-Base, DeBERTa-Base, ELECTRA-Base,

BioBERT, SciBERT, BlueBERT and BiomedNLP-BiomedBERT. Each of these 19 base models was fine-tuned on one of the six fine-tuning datasets (with & without PI names, one of the three distributions based on gender & race). In total, 114 models were fine-tuned. Results were then averaged over each fine-tuning configuration. The input representation is shown in Table 4. The task is to predict whether a given grant and project are a true match, see Appendix C for an example of the input representation. Training was conducted on RTX 6000 ADA GPUs in parallel (1 GPU per model), with consistent parameters across models, as detailed in Table 5.

Table 5: Training Parameters and Configurations

Parameter	**Value**
Init Model	A 'BERT family' model
Train Data	Adjusted NIH dataset
GPU	1 x RTX 6000 ADA 48Gb
GPU Time	~40-200 min. depending on model
Train Batch Size	32
Number of Epochs	4
Max Token Length	254
Steps per Epoch	Number of training samples divided by batch size
Evaluation Frequency	10% of steps per epoch
Evaluation Steps	10% of steps per epoch
Training Loss	Cosine Similarity Loss
Validation Evaluator	Embedding Similarity Evaluator on validation samples
Warmup Steps	10% of total training steps
Optimizer	AdamW
Learning Rate	0.00002 (2e-05)
Weight Decay	0.01