What is in a name? Mitigating Name Bias in Text Embedding Similarity via Anonymization

Sahil Manchanda and Pannaga Shivaswamy
Pocket FM, India
{sahilm1992, pannaga.datta}@gmail.com

Abstract

Text-embedding models are often used for finding similarity between texts using cosine similarity in a variety of tasks. Most models exhibit biases arising from the data on which they are trained on. In this paper, we examine a hitherto unexplored bias in text-embeddings in similarity tasks: bias arising from the presence of *names* such as persons, locations, organizations, etc., in the text. Our study shows how the presence of *name-bias* in text-embedding models can potentially lead to erroneous conclusions in the assessment of thematic similarity. *Text-embeddings can mistakenly indicate similarity between texts based on names in the text, even when their actual semantic contents do not have similarity or indicate dissimilarity simply because of the names in the text, even when the texts match semantically*. We first demonstrate the presence of name bias in different text-embedding models and then propose *text anonymization* during inference, which involves removing references to names while preserving the core theme of the text. The efficacy of the anonymization approach is demonstrated on three downstream NLP tasks involving embedding similarities, achieving significant performance gains. Our simple and training-optimization-free approach offers a practical and easily implementable solution to mitigate name bias. The code of our work can be found at `https://github.com/sahilm1992/name_bias`.

1 Introduction

Text-embedding models, which give a concise numerical representation to sentences/paragraphs have become fundamental tools for downstream NLP tasks in fields such as healthcare, education, law and scientific research (Chrysostomou and Aletras, 2022; Reimers, 2019; Tenney, 2019; Nie et al., 2024; Sun et al., 2019). A cosine similarity between embeddings is generally used (Zhang et al., 2019; Mathur et al., 2019) in a variety of tasks. With a similarity measure, the goal is to find which two texts are similar to or different from each other.

Text-embedding models are often trained on large amounts of text on the Internet. This data can inadvertently contain biases of various kinds, reflecting social prejudices and stereotypes. As a result, these models can generate biased embeddings, reinforcing harmful stereotypes or discriminating against certain cultural groups, genders, etc. (Gallegos et al., 2024; Li et al., 2023a; Rakivnenko et al., 2024). Furthermore, the presence of bias in models could lead to embeddings that disproportionately emphasize particular parts of the text, consequently failing to capture the true semantics and themes within the text (Rakivnenko et al., 2024).

While important, existing studies on biases, predominantly examine biases in text-embedding models mostly related to gender, geography, race, religion etc. (Rakivnenko et al., 2024; May et al., 2019; Bolukbasi et al., 2016; Kotek et al., 2023; Nghiem et al., 2024). In this paper, we demonstrate that text-embedding models exhibit significant bias towards *names* within the text. To illustrate this, we begin with a motivating example in Table 1. We present a simple narrative (*Story 1*). We then show a similar plot while substituting the name of the main character in (*Story 2*). In the third narrative (*Story 3*), we introduce a distinct and contradicting storyline from *Story 1* while retaining the original character names. We embed all three stories using text-embedding models. We observe that the similarity between Story 1 and Story 3, despite their differing plots, is consistently higher than the similarity between Story 1 and Story 2, which share highly semantically similar plots but differ in character names. This is very counter-intuitive since the text-embedding models seem to prioritize name similarity over the text's narrative structure. While this is admittedly an illustrative example, we proceed to generate numerous such narratives and conduct a thorough investigation of

Findings of the Association for Computational Linguistics: ACL 2025, pages 17759–17781
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

this bias in our experiments.

Our observation reveals a critical issue that can significantly impact applications that rely on semantic similarity, including semantic search, information retrieval, and plagiarism detection (Minaee et al., 2024; Pudasaini et al., 2024): consider the challenge of accurately assessing the similarity between two stories/plots with identical underlying meanings but distinct character names. Current methods may erroneously classify these stories as dissimilar, leading to inconsistent and unreliable results. Further, based upon our investigation, we would like to mention upfront that the issue is not confined to certain cultures, cross-culture, but is universal in the sense that the name bias issue occurs in a very broad sense.

Story Id	Text
Story 1	*Alejandro* gently examined the injured bird. He gave it food.
Story 2	*Jelani* tenderly inspected the wounded bird and gave it a meal to eat.
Story 3	*Alejandro* tracked the injured bird. He used it as his food.

Model	Cosine Similarity	
	Story1, Story 2 ↑	Story 1, Story 3 ↓
all-mpnet-base-v2	0.755	0.778
all-distilroberta-v1	0.780	0.798
all-MiniLM-L6-v2	0.660	0.853
gemini	0.864	0.848
multi-qa-distilbert-cos-v1	0.579	0.907
paraphrase-MiniLM-L6-v2	0.775	0.855
distiluse-base-multilingual-cased-v1	0.752	0.889
distiluse-base-multilingual-cased-v2	0.742	0.875
paraphrase-multilingual-MiniLM-L12-v2	0.836	0.840
msmarco-distilbert-cos-v5	0.584	0.817
multi-qa-mpnet-base-cos-v1	0.694	0.854
voyage-3-lite	0.780	0.868
text-embedding-3-small	0.755	0.826
text-embedding-3-large	0.741	0.808
gte-large	0.928	0.970
gte-base	0.900	0.973
e5-base-v2	0.916	0.949
e5-large-v2	0.918	0.935
Mistral	0.890	0.917

Table 1: **Impact of names on similarity**: We see that Story 1 is similar to Story 2 but has different person names(*Alejandro*, *Jelani*). Story 3 is different from Story 1 but has same name (*Alejandro*) as Story 1. We observe that, in most embedding models a different story with opposite meaning and same name(*Alejandro*) is getting a higher similarity score in comparison to the same story with different names.

Having briefly revealed the issue of name bias in text-embedding models, we outline our contributions in the work: First, we identify bias arising from names in textual content. Although several forms of biases have been studied in the past (see Sec.2), to the best of our knowledge, our work is the first that specifically looks at bias associated with names and how they can influence the embeddings coming out of embedding models. Toward this end, we propose a benchmarking study to comprehensively assess this bias. Second, we propose a simple *inference-time text-anonymization* technique designed to overcome the identified bias. Our method does not require any model fine-tuning or retraining of the text-embedding models. The approach offers a simple, intuitive, and effective way to mitigate the problem rather than relying on complex computations. Third, we conducted extensive experiments to study the identified problem in detail on a variety of text-embedding models and tasks. Our results demonstrate that our anonymization approach effectively reduces name bias within embeddings in semantic similarity and downstream tasks.

We emphasize here that our study investigates thematic and semantic similarities within textual data while acknowledging certain applications involving text tied to specific individuals or locations, our primary focus lies on the broader thematic context rather than characters in the text. For example, if an article is about Albert Einstein, then the name Albert Einstein would be very important for the article and replacing it with an arbitrary name or removing it from the article would not make much sense. However, in a fictitious story, whether the name is James or Michael may matter much less. Our work focuses on such tasks where the task is not about the entity itself but rather on the semantic meaning of it regardless of the names.

2 Related Work

Biases in Text-embedding models: Text-embedding models while powerful, can inadvertently reflect and amplify existing biases and prejudices; there is vast research understanding and mitigating bias in such models. For example, there is work focusing on models that investigate under-representation or misrepresentation of specific groups, such as LGBTQ+ individuals, leading to skewed or inaccurate outcomes (May et al., 2019; Bolukbasi et al., 2016; Cheng et al., 2021). Another type of study focuses on gender bias in word embeddings models (Rakivnenko et al., 2024). The study highlights a concerning issue i.e., many embedding models associate specific occupations with particular genders. Nikolaev and Padó (2023) studied biases at a sentence-level in sentence transformers influenced by different parts of speech such as common nouns, adverbs etc. While we discuss text-embedding models, it is important to highlight works that investigate bias within Large Language Models (LLMs) for text-generation which are a part of this ecosystem (Gallegos et al., 2024). Schwöbel et al. (2023) observed "geographical erasure" where certain regions are underrepresented in LLM outputs. Manvi et al. (2024) showed that LLMs often favor developed regions and exhibit negative biases towards locations with lower socioeconomic conditions, particularly on subjective topics such as attractiveness and intelligence. Further, some works have also investigated cross-cultural biases in LLMs for text generation (Naous et al., 2023; Ramezani and Xu, 2023; Cao et al., 2023; Arora et al., 2022). Compared to the above work, we investigate name-bias in text-embeddings, an area not previously explored in existing research to the best of our knowledge.

Debiasing methods: Various approaches have been proposed to tackle different kinds of biases in text-embedding models highlighted above. One common technique to remove such biases is to update the training dataset and make it unbiased and re-train the model (Brunet et al., 2019; Ngo et al., 2021). Another paradigm involves applying approaches such as disentanglement or alignment where models are fine-tuned to remove biases associated with concepts such as gender, religion etc. (Kaneko and Bollegala, 2021; Guo et al., 2022; Kenneweg et al., 2024). An alternative approach involves post-processing of the embeddings. Specifically, it involves adding a debiasing module after encoders to filter out certain biases in the representations Cheng et al. (2021). For more details on this topic, we refer the reader to survey by Li et al. (2023a) for more details.

We emphasize some key considerations based upon the discussions above. Firstly, all the aforementioned techniques require an optimization phase, involving either retraining the initial model, fine-tuning with a modified loss or post-processing of the generated embeddings. Secondly, these methods are often designed to address specific bias types, such as social, gender, or religious biases. Notably, the identification and mitigation of name bias has not been previously explored to our knowledge.

3 Understanding name bias

In this section, we investigate the presence of bias within text-embedding models related to names. Our primary objective is to investigate the influence of names containing identity-specific information on the resulting text embeddings, while ensuring the semantic structure of the text remains unchanged.

3.1 Benchmarking Methodology

To understand the impact of bias associated with names, we systematically replace instances of names in text with alternatives. For the sake of simplicity, in this section, we focus on person names and country names[1]. Given a text, we first identify instances of person and country names in the text.[2] To study bias w.r.t. person names, we replace each person name in the text with a randomly sampled name from a list of person names. In the text, all instances of the same person are replaced by the same sampled name. Similarly, country names are replaced with a random country name sampled from a predefined list of countries. This process only changes the person names and countries and does not change the original structure or meaning of the text.

Formally, given a universe of n person names $P = \{p_1, p_2, p_3 \cdots p_n\}$, and l country names $C = \{c_1, c_2, c_3 \cdots c_l\}$, we apply a straight-forward algorithm (Appendix A Algorithm 1) that randomly replaces names and countries for a given text T from a list of names and countries to obtain a perturbed text T'.

We generate K=20 such perturbations capturing a wide range of person and country names. The names used for replacement are shown in Table 14 in the Appendix and we have names from many different cultures/countries. An illustrative example of a perturbation is presented in Table 2.

Original Text(T)	Perturbed Text (T')
Mike has been living in **Belgium** for five years and made a fortune by winning a lottery. **Mike** spent most of his money on treatment of his brother **Donald** who was suffering from Lung Cancer.	**Dwayne** has been living in **France** for five years and made a fortune by winning a lottery. **Dwayne** spent most of his money on treatment of his brother **Shawn** who was suffering from Lung Cancer.

Table 2: Example of text perturbation.

The objective is to determine the degree of semantic divergence observed between perturbed text instances, resulting from the replacement of names and countries, by examining their embeddings. As discussed above, for a text T we create K perturbations $\{T'_i \mid 1 \leq i \leq K\}$. Each of these K perturbed text versions were processed through a

[1] We study the impact of perturbation of person names only in the Appendix.

[2] More details of the datasets are described in Sec. 3.3

text-embedding model, to obtain its corresponding embedding. Subsequently, to capture the distance between the perturbed text's embeddings with each other, we calculate the pairwise cosine similarity (Pedregosa et al., 2011) between all K embeddings. For example, if a text sample has K=20 perturbations, we get $\frac{K\times(K-1)}{2} = 190$ similarity scores. Given N such text samples in a dataset, to arrive at a single metric, we first compute pairwise cosine similarities(between the perturbed text embeddings) for a given text, excluding the self-similarity comparisons (i.e., the similarity of a perturbed text embedding to itself). For N samples, we obtain $N \times \frac{K\times(K-1)}{2}$ similarity scores. Let emb_{si} refer to the embedding of i^{th} perturbation of sample s where $s \in \{1, 2, ..., N\}$ and $i \in \{1, 2, ..., K\}$. Then, average similarity across N samples is defined as:

$$\frac{1}{N \times \frac{K(K-1)}{2}} \sum_{s=1}^{N} \left[\sum_{i=1}^{K} \sum_{\substack{j=1 \\ j\neq i}}^{K} Sim(emb_{si}, emb_{sj}) \right]$$

A higher average similarity indicates that the perturbed texts are closer to each other in the semantic space, suggesting less deviation. Conversely, a lower average similarity score suggests a higher degree of deviation from the expected semantic relationship. It suggests that the embedding model exhibits a bias towards names in the text, potentially affecting its ability to accurately capture the theme of the text.

3.2 Candidate Text-embedding Models

We analyzed a diverse set of leading text embedding models from the academia and the industry. This includes models explicitly trained on diverse languages and tasks such as semantic search, question-answering etc. We include models such as *multi-qa-distilbert-cos-v1* and *multi-qa-mpnet-base-cos-v5* for question answering, and *paraphrase-MiniLM-L6-v2* and *paraphrase-multilingual-MiniLM-L12-v2* for identifying semantic similarity (Reimers, 2019). Other notable models include *all-mpnet-base-v2, all-distilroberta-v1*, *all-MiniLM-L6-v2*, designed for general-purpose text representation (Reimers, 2019), embedding models *e5-large-v2, e5-base-v2* (Wang et al., 2022). Further models such as *gte-base, gte-large* (Li et al., 2023b) designed for semantic textual similarity and text ranking were included. Additionally, multilingual models like *distiluse-base-multilingual-cased-v1* and *distiluse-base-multilingual-cased-v2* are also included (Reimers and Gurevych, 2020). We also include *msmarco-distilbert-cos-v5* specialized model for search (Reimers, 2019). Mistral(via *mistral-embed* API) (Jiang et al., 2023) is also included which is trained on a diverse range of textual data, including books, articles, and websites. Additionally, we also choose cutting-edge models which are not open-source namely *text-embedding-3-small* and *text-embedding-3-large* from Open AI (OpenAI, 2024), *gemini-1.5-pro(text-embedding-004)* from Google (Team et al., 2023) and *voyage-3-lite* from Voyage AI (AI, 2024).

3.3 Benchmark Datasets

CMU Movie Dataset (Bamman et al., 2013): The CMU Movie dataset primarily consists of 6,559 textual plot summaries of movies spanning multiple sentences. These summaries are typically short, concise descriptions of the main events and storylines within a film. They often include key characters, conflicts, and resolutions. This dataset consisted of an average of around 117 words per story and an average of about 8.2 characters/countries per story.

CMU Book Dataset (Bamman and Smith, 2013): Similar to CMU Movie, the core of this dataset consists of concise multiple sentence summaries of 42,306 books. These summaries capture the main plot points, key characters, and themes. The summaries had an average of 129 words per summary and about 7.9 characters/countries per summary.

We select plots where the number of words are less than 250 which is within token limit of most models under consideration[3].

3.4 Analyzing Bias

We then perturb each instance in the two data sets with the procedure outlined in Section 3.1. In Table 3 and 4 we observe a significant deviation in the average cosine similarity which should be close to one if the cosine similarity captured the real semantic similarity rather than information in names present in the text[4]. Any deviation from one indicates that the embeddings are

[3] For each embedding model, we evaluate its performance only on samples which are within the limits of its maximum context window.

[4] We also experimented by using euclidean distance instead of cosine similarity in Tab. 18 in Appendix. The conclusion remained similar and therefore we proceeded with cosine similarity for remaining experiments.

Model Name	Cosine sim per perturbation pair
all-mpnet-base-v2	0.774 ± 0.001
gte-base	0.95 ± 0.0
gte-large	0.948 ± 0.0
e5-base-v2	0.923 ± 0.0
e5-large-v2	0.918 ± 0.0
all-distilroberta-v1	0.768 ± 0.001
all-MiniLM-L6-v2	0.706 ± 0.001
gemini	0.885 ± 0.0
multi-qa-distilbert-cos-v1	0.733 ± 0.001
paraphrase-MiniLM-L6-v2	0.742 ± 0.001
distiluse-base-multilingual-cased-v1	0.786 ± 0.001
distiluse-base-multilingual-cased-v2	0.795 ± 0.001
paraphrase-multilingual-MiniLM-L12-v2	0.75 ± 0.001
msmarco-distilbert-cos-v5	0.681 ± 0.001
multi-qa-mpnet-base-cos-v1	0.743 ± 0.001
text-embedding-3-small	0.742 ± 0.0
text-embedding-3-large	0.779 ± 0.0
voyage-3-lite	0.76 ± 0.0
Mistral	0.926 ± 0.0

Table 3: **Bias Measurement on CMU Movie dataset**. For each show, we create $K{=}20$ **perturbations** by replacing person names and country names. In this experiment, we used plot samples that contain both person and country names but does not mention any city/town/village/nationality keywords (Spanish, American etc.) in order to minimize the impact of other variables. We report the mean and the std. error rounded off to 3 decimal places.

Model Name	Cosine sim per perturbation pair
all-mpnet-base-v2	0.777 ± 0.001
gte-base	0.952 ± 0.0
gte-large	0.953 ± 0.0
e5-base-v2	0.929 ± 0.0
e5-large-v2	0.927 ± 0.0
all-distilroberta-v1	0.778 ± 0.001
all-MiniLM-L6-v2	0.693 ± 0.001
gemini	0.89 ± 0.0
multi-qa-distilbert-cos-v1	0.743 ± 0.001
paraphrase-MiniLM-L6-v2	0.735 ± 0.002
distiluse-base-multilingual-cased-v1	0.777 ± 0.001
distiluse-base-multilingual-cased-v2	0.785 ± 0.001
paraphrase-multilingual-MiniLM-L12-v2	0.746 ± 0.002
msmarco-distilbert-cos-v5	0.707 ± 0.001
multi-qa-mpnet-base-cos-v1	0.75 ± 0.001
text-embedding-3-small	0.761 ± 0.001
text-embedding-3-large	0.795 ± 0.001
voyage-3-lite	0.781 ± 0.001
Mistral	0.933 ± 0.000

Table 4: **Bias Measurement on CMU Books dataset.** We follow the same evaluation setup as in Table 3.

heavily biased by the choice of names rather than from the similarity of the text. Models like $msmarco{-}distilbert{-}cos{-}v5$ exhibit significant sensitivity to changes in person and country names, as evidenced by an average cosine similarity ≈ 0.7. This suggests that the model's embeddings may be heavily influenced by specific entities rather than capturing the underlying semantic meaning of the text. Observations from the evaluation of both datasets suggest that only a few models are able to even achieve a score of 0.9 or above. However, we observe that even for all those models, the scores are still far away from one indicating further room for improvement.

In the above experiment, we replaced the names of people and countries and generated a perturbed text. One may ask: how much of the bias is from country name versus person names? To study this, we considered an experiment in which we perturbed the text by only replacing person names while keeping the country names as they were in the original text. We also examined variations in which all the perturbed names are sampled from the same country and demonstrate that bias persists even if text samples differ only by person names even from the same country. These results can be found in Appendix C. Further, apart from the experiments presented in this section, we performed a perturbation experiment on one of the datasets used in the experiments section and see that a similar conclusion holds in Table 19.

4 Methodology: Overcoming Bias through Anonymization

Previously, we showed that how just changing person names/country names can impact the embeddings significantly. In this section, we introduce a simple *inference-time anonymization* technique to mitigate the bias caused by names. The core idea is to mitigate the influence of names on embeddings, and making the resulting *debiased* anonymized embeddings to be more generalizable and less prone to biases related to particular individuals or entities.

The anonymization of a text T during inference is achieved through the following process. We first identify in T, occurrences of desired entities such as person names, locations and organizations relevant to the use case. We *anonymize* the text by removing those occurrences from T. The anonymized text referred to as T_{anon} retains the overall structure and meaning of the original text T while removing any specific references to person names etc. This anonymization can be achieved via tools such as Large Language Models(LLMs) (Zhao et al., 2023) or Named Entity Recognition tools (Jehangir et al., 2023). In our work, we used *gemini* and *anthropic.claude-3-5-sonnet* text-generative models for anonymization using prompts. Depending upon the use-case, different names in text such as person names, cities, countries, organizations can be removed. We would like to clarify that the same process of anonymization can also be done through Named-Entity Recognition(NER) tools (Jehangir et al., 2023), however in our initial experiments we found LLMs to be more accurate. Sample prompts for anonymization are presented in Table 5. Post anonymization, the embeddings become independent of identity specific details such as person names/ country names etc. [5] Overall, the *debiased* embeddings generated on anonymized text promise reduced sensitivity to biases associated with particular individuals or entities. Note that the embeddings generated for sentences that differ solely in their named entities (e.g., character names) will now have a cosine similarity of 1. An alternate to removing *named* content for anonymization is to replace names with specific non-identifying placeholder words. This approach with its associated challenges is further examined in Appendix K.

Purpose	Prompt
Remove person names and location names	Given below text, please COMPLETELY DELETE all Person/Character names which are PROPER NOUNS and City/ Country/ Village/ Town/ Continent/ River/ Organization names which are PROPER NOUNS etc. Wherever they occur replace with empty string. Completely remove them and not anything else. Do not delete monument/landmark names like Eiffel tower etc. Do not remove He/She/him/her etc.. Output contains the modified text only.... The text is provided below ::::
Remove person names only	Given below text, please COMPLETELY DELETE all Person/Character names which are PROPER NOUNS. Wherever they occur replace with empty string. Completely remove them and not anything else. Do not remove He/She/him/her etc.. Output contains the modified text only.... The text is provided below ::::

Table 5: Prompts for Anonymization. In our experiments, we select the first prompt. Based upon the use case, the suitable prompt can be selected or modified accordingly.

5 Can anonymization help in down-stream tasks that use similarity from text-embedding models?

In this section, we investigate the performance of the anonymized text embeddings on three downstream tasks. The tasks are based on obtaining a similarity score between pieces of texts. These tasks are primarily based upon semantic similarity which find applications in areas such as information retrieval, clustering, plagiarism detection, question answering etc. (Reimers, 2019). The tasks that we evaluate on differ in various aspects such as the nature of the task, evaluation methodology, the judgment score available, etc. On these tasks, our experiments show that embeddings based on anonymized text can significantly help in downstream tasks.

5.1 Task 1: Semantic Similarity Between Query and Text-Pairs with Binary Labels.

Recall from Sec. 3 that altering only the names/locations in two otherwise identical stories/paragraphs significantly impacted their text embeddings. In this section, we investigate whether anonymization technique proposed in Sec. 3.4 can effectively mitigate this type of bias. Towards this, we explore the Semantic Textual Similarity (STS) task.

Semantic similarity seeks to determine the degree to which two pieces of text convey similar meaning (Muennighoff et al., 2022; Reimers, 2019). This goes beyond simple word matching, aiming to understand the underlying meaning within the text. In today's era of deep learning (Reimers et al., 2016; Muennighoff et al., 2022), achieving accurate semantic similarity relies heavily on high-quality embeddings, which represents sentences as dense vectors in a continuous space.

In this experiment we investigate whether the text-embeddings are able to capture the semantic nuances within the text or are they biased towards names? Ideally, a good embedding model should be able to differentiate reasonably well between two stories/paragraphs which have very different themes even if they contain same names. To investigate this, we create a dataset of 50 paragraph triplets. Each triplet includes a *query* paragraph, a *positive* paragraph that is *highly semantically similar* but with distinct person and location names, and a *negative* paragraph that is semantically *dissimilar* to the query text but has same person names/location names as in query text. For each triplet, (*query*, *positive*) pair is assigned a label 1 (positive) and *(query, negative)* pair is assigned a label 0 (negative). Two sample examples can be found in Table 7 in the the rows marked as *Original*. The entire set of generated triplets with labels are present in Appendix G. We evaluate the performance of different models on the STS task using *AUC ROC score* between cosine similarity scores of embeddings and the ground truth.

Peformance on Semantic Similarity. Tab. 6 presents the AUC-ROC scores for different models on the STS task. The results indicate that the AUC scores for the majority of models are significantly below 0.5. This finding suggests a critical

[5] The type of anonymization i.e removing person names and/or country names and/or city names etc. used determines the exact level of independence.

Model	*Original text:* The *(query, positive)* paragraphs share the same meaning but different person/location names. The *(query, negative)* paragraphs share different meaning but same person/location names.	*Identical Names:* The (*query*, *positive*, *negative*) paragraphs in the same triplet contain the same person/location names.	*Anonymized text:* Anonymization applied to (*query*, *positive*, *negative*) paragraphs.
all-mpnet-base-v2	0.233	0.856	0.934 ± 0.003
gte-large	0.206	0.938	0.955 ± 0.001
gte-base	0.208	0.913	0.927 ± 0.004
e5-base-v2	0.173	0.932	0.923 ± 0.002
e5-large-v2	0.276	0.933	0.936 ± 0.006
all-distilroberta-v1	0.247	0.920	0.932 ± 0.001
all-MiniLM-L6-v2	0.056	0.841	0.845 ± 0.003
multi-qa-distilbert-cos-v1	0.046	0.800	0.835 ± 0.005
paraphrase-MiniLM-L6-v2	0.129	0.954	0.895 ± 0.003
distiluse-base-multilingual-cased-v1	0.205	0.777	0.797 ± 0.002
distiluse-base-multilingual-cased-v2	0.244	0.770	0.818 ± 0.001
paraphrase-multilingual-MiniLM-L12-v2	0.161	0.932	0.918 ± 0.005
msmarco-distilbert-cos-v5	0.055	0.812	0.851 ± 0.022
multi-qa-mpnet-base-cos-v1	0.038	0.896	0.924 ± 0.007
gemini	0.584	1.000	0.996 ± 0.001
text-embedding-3-small	0.047	0.937	0.954 ± 0.010
text-embedding-3-large	0.114	0.977	0.980 ± 0.008
voyage-3-lite	0.134	0.899	0.898 ± 0.000
Mistral	0.049	0.883	0.934 ± 0.011

Table 6: **Evaluation on Task 1: Semantic Similarity Task.** AUC scores obtained on Semantic Similarity Task. Our proposed strategy of anonymization achieves high quality results across all models. Mean and standard error are reported based on results from two separate LLM runs for anonymization.

	Query	Pos/Neg	Sim score	Label
Original	Alejandro quickly ran to the store to buy a cold drink. He was eager to have a glass of cold drink.	**POS:** Quickly, Hiroki dashed to the local market to procure some cold drinks. He was yearning for a chilled glass of cold drink.	0.58	1
		NEG: Alejandro has stopped buying cold drinks from market. He only drinks cold drinks made at home.	0.69	0
Anonymized	quickly ran to the store to buy a cold drink. He was eager to have a glass of cold drink.	**POS:** Quickly, dashed to the local market to procure some cold drinks. He was yearning for a chilled glass of cold drink.	0.80	1
		NEG: has stopped buying cold drinks from market. He only drinks cold drinks made at home.	0.47	0
Original	Ganga and Yamuna are two mighty rivers. They are lifelines for millions of people in the region.	**POS:** Yangtze is a mighty river. It is a long river and is the lifeline for millions of people in the region.	0.54	1
		NEG: Ganga and Yamuna are two sisters. They had their schooling in the region and schooling provided a lifeline for them.	0.70	0
Anonymized	and are two mighty rivers. They are lifelines for millions of people in the region.	**POS:** is a mighty river. It is a long river and is the lifeline for millions of people in the region.	0.70	1
		NEG: and are two sisters. They had their schooling in the region and schooling provided a lifeline for them.	0.56	0

Table 7: Examples showing impact of anonymization on semantic similarity using embeddings created by *msmarco-distilbert-cos-v5*.

issue, as even a random classifier would be expected to achieve an AUC score of approximately 0.5. The fact that most of the AUC is much below 0.5 suggests that the cosine similarity based ranking got the ordering wrong! Gemini's AUC is better than random, however, it also gets improved significantly after anonymization. Such low AUC scores strongly imply that the embeddings used in these models are primarily capturing identity-related information, leading to a significant bias in the model's embeddings and predictions. Next, we observe that the AUC-ROC results post anonymization. We see that anonymization can improve the model's capacity to grasp the core semantic meaning in the text as reflected in the significantly higher AUC-ROC numbers (closer to 1). Additionally, it is important to note that all models attain high AUC scores when all stories share identical names. This indicates that the models can effectively distinguish between sentences conveying the same or different meanings when identity information remains constant. The aforementioned observations highlights that anonymization is crucial to avoid situations where semantically equivalent paragraphs are assigned unique embeddings solely based on the presence of identity information (such as names). Conversely, it's essential that when texts have significant semantic variations, even if they contain

model	Spearman-correlation (Original Text)	Spearman-correlation (Anonymized)	Pearson-correlation (Original Text)	Pearson-correlation (Anonymized)
all-mpnet-base-v2	0.262	**0.344** ± **0.001**	0.321	0.364 ± 0.002
gte-large	**0.351**	0.338 ± 0.005	**0.408**	0.382 ± 0.000
gte-base	0.326	**0.328** ± **0.003**	**0.381**	0.367 ± 0.003
e5-base-v2	0.356	**0.374** ± **0.003**	0.393	**0.398** ± **0.005**
e5-large-v2	0.330	**0.376** ± **0.003**	0.386	**0.409** ± **0.008**
all-distilroberta-v1	0.245	**0.327** ± **0.010**	0.302	**0.370** ± **0.004**
all-MiniLM-L6-v2	0.251	**0.330** ± **0.004**	0.282	**0.354** ± **0.008**
gemini	0.381	**0.390** ± **0.001**	**0.456**	0.436 ± 0.004
multi-qa-distilbert-cos-v1	0.240	**0.292** ± **0.003**	0.269	**0.316** ± **0.009**
paraphrase-MiniLM-L6-v2	0.283	**0.352** ± **0.006**	0.317	**0.370** ± **0.000**
distiluse-base-multilingual-cased-v1	0.282	**0.356** ± **0.001**	0.325	**0.386** ± **0.003**
distiluse-base-multilingual-cased-v2	0.308	**0.357** ± **0.000**	0.345	**0.389** ± **0.004**
paraphrase-multilingual-MiniLM-L12-v2	0.261	**0.332** ± **0.001**	0.281	**0.364** ± **0.005**
msmarco-distilbert-cos-v5	0.232	**0.304** ± **0.002**	0.262	**0.333** ± **0.007**
multi-qa-mpnet-base-cos-v1	0.274	**0.323** ± **0.002**	0.317	**0.353** ± **0.000**
text-embedding-3-small	0.374	**0.382** ± **0.003**	0.416	**0.422** ± **0.006**
text-embedding-3-large	0.366	**0.382** ± **0.009**	0.428	**0.429** ± **0.017**
voyage-3-lite	**0.359**	0.322 ± 0.006	**0.400**	0.352 ± 0.003
Mistral	0.313	**0.335** ± **0.000**	**0.370**	0.368 ± 0.004

Table 8: **Evaluation on Task 2: Semantic similarity with graded relevance.** The table presents correlation between cosine similarity between human & machine summaries and relevance(ground truth) provided by human evaluators .

identical identity information, their embeddings are able to able to capture it.

Examples of similarity post-anonymization. In Tab. 7, we show some instances of how similarity values between embeddings change between *(query, positive)* pair and *(query, negative)* pair post anonymization. Before anonymization, the models assigned higher similarity scores to negative pairs and lower similarity scores to positive pairs in a counterintuitive way. Anonymization resulted in the models predominantly attending to the semantic structure of the text, which is accurately reflected in similarity scores. We would like to highlight that these samples are a subset of examples used for AUC computation on the STS task in Tab. 6.

5.2 Task 2: Semantic Similarity With Graded Human Relevance.

In the previous task, a binary approach was employed to assess text pair similarity, categorizing text-pairs as either similar or dissimilar. In the task proposed in this section, we employ a more refined approach for evaluation by utilizing a graded relevance scale from 1 to 5 between a pair of text. The graded scale enables a more nuanced and granular assessment of semantic similarity between pairs, providing a richer understanding of their relationship. To evaluate this, we use the machine summary evaluation task from Muennighoff et al. (2022), which involves automatically assessing the relevance of machine-generated summaries, commonly assessed by calculating the semantic similarity between the embeddings of the summary and the original document/human summaries.

For this task, we follow the same evaluation setup as Muennighoff et al. (2022) which we describe next. We use the SummEval dataset (Fabbri et al., 2021; Muennighoff et al., 2022) with 100 text samples, each containing 16 machine and 10 human summaries. Human relevance scores $(1-5)$ are assigned to each machine summary. We first obtain summary embeddings using text-embedding models for each machine summary and human summary in all 100 samples. Without loss of generality, for a given text sample out of the 100 samples, for each machine summary $\{m_i \mid 1 \leq i \leq 16\}$, we get its predicted score based on its maximum cosine similarity to any human summary $\{h_j \mid 1 \leq j \leq 10\}$ within the same text sample i.e $machine\_pred\_score(m_i) = max_{1 \leq j \leq 10}\, cos\_sim(m_i, h_j)$. This yields 16 machine summary quality predicted scores for each sample i.e., one predicted score for each machine summary. Further, as mentioned earlier, we have a human relevance score assigned to each machine summary. Overall, across all text samples, we get 1600 *machine summary predicted scores* and its corresponding *human relevance scores*. We then correlate these two scores using Pearson and Spearman coefficients (Muennighoff et al., 2022). Higher correlations indicate better alignment between model-assigned scores and human judgments, suggesting more reliable evaluation.

Impact of Anonymization Table 8 shows that post-anonymization, the performance of various text-embedding models significantly improves in predicting graded human-rated summary quality. Spearman and Pearson correlation coefficients increase substantially, indicating that the model's assessment of summary quality after anonymization better aligns with human evaluations. This improvement is substantial, with some models like *all-distilroberta-v1* showing a performance increase of around 30%.

5.3 Task 3: Paraphrase Detection Task

model	AUC (Original)	AUC (Anonymized)
all-mpnet-base-v2	0.793	0.804
all-distilroberta-v1	0.736	0.781
all-MiniLM-L6-v2	0.757	0.775
gemini	0.828	0.829
multi-qa-distilbert-cos-v1	0.729	0.754
paraphrase-MiniLM-L6-v2	0.777	0.798
distiluse-base-multilingual-cased-v1	0.770	0.803
distiluse-base-multilingual-cased-v2	0.768	0.799
paraphrase-multilingual-MiniLM-L12-v2	0.786	0.798
msmarco-distilbert-cos-v5	0.719	0.757
multi-qa-mpnet-base-cos-v1	0.719	0.757
text-embedding-3-small	0.759	0.795
text-embedding-3-large	0.771	0.804
voyage-3-lite	0.738	0.761
gte-large	0.749	0.792
gte-base	0.742	0.797
e5-base-v2	0.778	0.796
e5-large-v2	0.784	0.808
Mistral	0.755	0.802

Table 9: **Task 3: Results on Paraphrase Detection Task**.

The Paraphrase Detection task in NLP is about automatically determining whether two given text snippets convey the same meaning (Xu et al., 2015). The task is binary classification, i.e, to predict whether a sentence pair is paraphrased or not. For evaluation purposes, we use the Microsoft Research Paraphrase Corpus [6]. To evaluate the impact of anonymization of names, we only considered samples that had at least one occurrence of a person name. This constituted 608 samples. We computed the embeddings of sentences before and after applying anonymization and then calculated their cosine similarity, which served as the predicted score. The results are shown in Table 9. The results show that the AUC improves significantly after anonymization of text.

In summary, the results of the three downstream tasks demonstrate a substantial enhancement in the semantic similarity post-anonymization.

6 Conclusion

In this work, we highlight the bias in text embeddings stemming from the presence of names in the text. We showed concrete examples, over multiple text-embedding models, that similarities between embeddings can be dominated by names in the text rather than the semantic meanings of the text. We then proposed a method to mitigate bias by performing anonymization at inference time. This involved the removal of names from the text and using the anonymized text to create the embeddings. Our findings demonstrate that anonymized text embeddings significantly outperform deanonymized text embeddings on tasks involving semantic similarity. While we proposed one way to mitigate the issue through anonymization, a deeper question that remains is: how to train text-embedding models such that the embeddings capture the semantics more than the names in the text?

7 Acknowledgment

We thank Prateek Dixit for supporting this work. We thank Othmane Rifki for his useful feedback on an earlier version of this paper and his help with a couple of models. We also thank the anonymous reviewers for providing useful feedback about this work that lead to substantial improvements in this paper.

8 Limitations

Below we discuss the limitations of the proposed work.

1. In this work we focused on evaluating/mitigating name bias in text-embedding models using texts from English language. The work presented here does not cover other languages. Further, the work also does not cover name bias issues arising in multi language texts.

2. While our proposed anonymization solution enhances thematic similarity, it is not ideal for situations requiring the preservation of identity that we are removing through anonymization. A partial and straightforward solution might involve anonymizing only non-critical identifying information depending upon the use-case. Many real world use cases may require dynamically balancing identity and thematic preservation to suit the specific needs of each use case.

[6] https://www.kaggle.com/datasets/doctri/microsoft-research-paraphrase-corpus/

3. In our work, we adopted similarity between text-embeddings as a proxy for their semantic similarity. While commonly used, it is still an estimate of semantic similarity and may overlook deeper semantic relationships that require reasoning. A limitation of this work is that we capture thematic similarity only to the extent that it is captured by the cosine similarity.

9 Broader Impact

This research uncovers *name-bias* in text-embedding models. It reveals how the presence of names can skew similarity judgments, leading to incorrect conclusions about thematic similarities. This impacts a wide range of NLP applications, potentially compromising accuracy in tasks from information retrieval to sentiment analysis. The major impact of this paper is uncovering such bias and how it can be mitigated at inference time. This work contributes to inspiring further investigation into building more robust text-embedding models.

References

Voyage AI. 2024. Embeddings.

Arnav Arora, Lucie-Aimée Kaffee, and Isabelle Augenstein. 2022. Probing pre-trained language models for cross-cultural differences in values. *arXiv preprint arXiv:2203.13722*.

David Bamman, Brendan O'Connor, and Noah A Smith. 2013. Learning latent personas of film characters. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 352–361.

David Bamman and Noah A Smith. 2013. New alignment methods for discriminative book summarization. *arXiv preprint arXiv:1305.1319*.

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.

Marc-Etienne Brunet, Colleen Alkalay-Houlihan, Ashton Anderson, and Richard Zemel. 2019. Understanding the origins of bias in word embeddings. In *International conference on machine learning*, pages 803–811. PMLR.

Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. 2023. Assessing cross-cultural alignment between chatgpt and human societies: An empirical study. *arXiv preprint arXiv:2303.17466*.

Pengyu Cheng, Weituo Hao, Siyang Yuan, Shijing Si, and Lawrence Carin. 2021. Fairfil: Contrastive neural debiasing method for pretrained text encoders. *arXiv preprint arXiv:2103.06413*.

George Chrysostomou and Nikolaos Aletras. 2022. Flexible instance-specific rationalization of nlp models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10545–10553.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

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*, pages 1–79.

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.

Basra Jehangir, Saravanan Radhakrishnan, and Rahul Agarwal. 2023. A survey on named entity recognition—datasets, tools, and methodologies. *Natural Language Processing Journal*, 3:100017.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Masahiro Kaneko and Danushka Bollegala. 2021. Debiasing pre-trained contextualised embeddings. *arXiv preprint arXiv:2101.09523*.

Philip Kenneweg, Sarah Schröder, Alexander Schulz, and Barbara Hammer. 2024. Debiasing sentence embedders through contrastive word pairs. *arXiv preprint arXiv:2403.18555*.

Hadas Kotek, Rikker Dockum, and David Sun. 2023. Gender bias and stereotypes in large language models. In *Proceedings of the ACM collective intelligence conference*, pages 12–24.

Yingji Li, Mengnan Du, Rui Song, Xin Wang, and Ying Wang. 2023a. A survey on fairness in large language models. *arXiv preprint arXiv:2308.10149*.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023b. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*.

Rohin Manvi, Samar Khanna, Marshall Burke, David Lobell, and Stefano Ermon. 2024. Large language models are geographically biased. *arXiv preprint arXiv:2402.02680*.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2019. Putting evaluation in context: Contextual embeddings improve machine translation evaluation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2799–2808.

Chandler May, Alex Wang, Shikha Bordia, Samuel R Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. *arXiv preprint arXiv:1903.10561*.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. Large language models: A survey. *arXiv preprint arXiv:2402.06196*.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2022. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*.

Tarek Naous, Michael J Ryan, Alan Ritter, and Wei Xu. 2023. Having beer after prayer? measuring cultural bias in large language models. *arXiv preprint arXiv:2305.14456*.

Huy Nghiem, John Prindle, Jieyu Zhao, and Hal Daumé III. 2024. " you gotta be a doctor, lin": An investigation of name-based bias of large language models in employment recommendations. *arXiv preprint arXiv:2406.12232*.

Helen Ngo, João G M Araújo Cooper Raterink, Ivan Zhang, Carol Chen, Adrien Morisot, and Nicholas Frosst. 2021. Mitigating harm in language models with conditional-likelihood filtration.(aug. *arXiv preprint arXiv:2108.07790*.

Zhijie Nie, Zhangchi Feng, Mingxin Li, Cunwang Zhang, Yanzhao Zhang, Dingkun Long, and Richong Zhang. 2024. When text embedding meets large language model: A comprehensive survey. *arXiv preprint arXiv:2412.09165*.

Dmitry Nikolaev and Sebastian Padó. 2023. Representation biases in sentence transformers. *arXiv preprint arXiv:2301.13039*.

OpenAI. 2024. Embeddings.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Shushanta Pudasaini, Luis Miralles-Pechuán, David Lillis, and Marisa Llorens Salvador. 2024. Survey on plagiarism detection in large language models: The impact of chatgpt and gemini on academic integrity. *arXiv preprint arXiv:2407.13105*.

Vasyl Rakivnenko, Nestor Maslej, Jessica Cervi, and Volodymyr Zhukov. 2024. Bias in text embedding models. *arXiv preprint arXiv:2406.12138*.

Aida Ramezani and Yang Xu. 2023. Knowledge of cultural moral norms in large language models. *arXiv preprint arXiv:2306.01857*.

N Reimers. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Nils Reimers, Philip Beyer, and Iryna Gurevych. 2016. Task-oriented intrinsic evaluation of semantic textual similarity. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 87–96.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Pola Schwöbel, Jacek Golebiowski, Michele Donini, Cédric Archambeau, and Danish Pruthi. 2023. Geographical erasure in language generation. *arXiv preprint arXiv:2310.14777*.

Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. 2019. Bert4rec: Sequential recommendation with bidirectional encoder representations from transformer. In *Proceedings of the 28th ACM international conference on information and knowledge management*, pages 1441–1450.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

I Tenney. 2019. Bert rediscovers the classical nlp pipeline. *arXiv preprint arXiv:1905.05950*.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

Wei Xu, Chris Callison-Burch, and Bill Dolan. 2015. SemEval-2015 task 1: Paraphrase and semantic similarity in Twitter (PIT). In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 1–11, Denver, Colorado. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

A Name Perturbation Algorithm

The perturbation algorithm detailed in Section 3 is shown in Algorithm 1. Essentially it replaces the names of entities in the text with another random name from a list containing the names of similar entities.

Algorithm 1 Perturb Text for Benchmarking

```
Require: $P$ : List of Person names, $C$ : List of Country names.
Input: Text $T$
Output: Text $T'$ with replaced entities
Initalize: $T' \leftarrow T$
Identify Entities:
    Identify all occurrences of person names $IP$ in $T'$.
    Identify all occurrences of country names $IC$ in $T'$.
Perturbation:
for each identified person $ip \in IP$ in text $T'$ do
    Randomly select a name $p_k \in P$ without replacement.
    Replace all occurrences of $ip$ with $p_k$ in text $T'$.
end for
for each identified country $ic \in IC$ in text $T'$ do
    Randomly select a country name $c_k \in C$ without replacement.
    Replace all occurrences of $ic$ with $c_k$ in text $T'$.
end for
Return $T'$ {Perturbed Text}
```

Model Name	Cosine sim per perturbation pair
all-mpnet-base-v2	0.842 ± 0.0002
all-distilroberta-v1	0.852 ± 0.0002
all-MiniLM-L6-v2	0.784 ± 0.0002
gemini	0.93 ± 0.0
multi-qa-distilbert-cos-v1	0.82 ± 0.0002
paraphrase-MiniLM-L6-v2	0.806 ± 0.0004
distiluse-base-multilingual-cased-v1	0.837 ± 0.0003
distiluse-base-multilingual-cased-v2	0.838 ± 0.0003
paraphrase-multilingual-MiniLM-L12-v2	0.82 ± 0.0003
msmarco-distilbert-cos-v5	0.799 ± 0.0002
multi-qa-mpnet-base-cos-v1	0.815 ± 0.0002
voyage-3-lite	0.847 ± 0.0001

Table 10: **Bias Measurement: Names from same country.** Perturbation of person names and replacing them with names from ***Spain***. We used CMU Book dataset for this experiment and set number of perturbations K=20. In this experiment we use samples without mention of country/city/town/other location names, nationality etc.

B Names used for perturbation in Benchmarking

Table 14 presents the universe of names used for perturbation in the benchmarking experiment in Sec. 3. These names represent a diverse range of geographies. The person names were selected to span a diverse range of geographies such as Asia, Americas, Africa, Europe, and Australia etc. It is not an exhaustive list as there are too many names in the world. However, it is easy to see that Table 14 does represent a very diverse set of names representing a variety of countries and cultures.

Model Name	Cosine sim per perturbation pair
all-mpnet-base-v2	0.840 ± 0.0002
all-distilroberta-v1	0.838 ± 0.0002
all-MiniLM-L6-v2	0.757 ± 0.0003
gemini	0.931 ± 0.0000
multi-qa-distilbert-cos-v1	0.806 ± 0.0002
paraphrase-MiniLM-L6-v2	0.790 ± 0.0004
distiluse-base-multilingual-cased-v1	0.830 ± 0.0003
distiluse-base-multilingual-cased-v2	0.833 ± 0.0003
paraphrase-multilingual-MiniLM-L12-v2	0.815 ± 0.0004
msmarco-distilbert-cos-v5	0.786 ± 0.0002
multi-qa-mpnet-base-cos-v1	0.810 ± 0.0002
voyage-3-lite	0.843 ± 0.0001

Table 11: **Bias Measurement: Names from same country.** Perturbation of person names and replacing them with names from ***France***. We used CMU Book dataset for this experiment and set number of perturbations K=20. In this experiment we use samples without mention of country/city/town/other location names, nationality etc.

Model Name	Cosine sim per perturbation pair
all-mpnet-base-v2	0.816 ± 0.0002
all-distilroberta-v1	0.828 ± 0.0002
all-MiniLM-L6-v2	0.750 ± 0.0003
gemini	0.931 ± 0.0000
multi-qa-distilbert-cos-v1	0.79 ± 0.0002
paraphrase-MiniLM-L6-v2	0.778 ± 0.0004
distiluse-base-multilingual-cased-v1	0.880 ± 0.0002
distiluse-base-multilingual-cased-v2	0.887 ± 0.0002
paraphrase-multilingual-MiniLM-L12-v2	0.796 ± 0.0004
msmarco-distilbert-cos-v5	0.780 ± 0.0002
multi-qa-mpnet-base-cos-v1	0.805 ± 0.0002
voyage-3-lite	0.850 ± 0.0001

Table 12: **Bias Measurement: Names from same country.** Perturbation of person names and replacing them with names from ***India***. We used CMU Book dataset for this experiment and set number of perturbations K=20. In this experiment we use samples without mention of country/city/town/other location names, nationality etc.

C Bias measurement with only person name perturbations

In the benchmarking study in Sec. 3, we investigated the divergence in text embeddings when person names and locations were perturbed. In this section, we examine the impact of replacing only

Model Name	Cosine sim per perturbation pair
all-mpnet-base-v2	0.815 ± 0.0001
all-distilroberta-v1	0.821 ± 0.0001
all-MiniLM-L6-v2	0.749 ± 0.0002
gemini	0.910 ± 0.0000
multi-qa-distilbert-cos-v1	0.787 ± 0.0001
paraphrase-MiniLM-L6-v2	0.773 ± 0.0003
distiluse-base-multilingual-cased-v1	0.843 ± 0.0002
distiluse-base-multilingual-cased-v2	0.848 ± 0.0002
paraphrase-multilingual-MiniLM-L12-v2	0.790 ± 0.0003
msmarco-distilbert-cos-v5	0.752 ± 0.0001
multi-qa-mpnet-base-cos-v1	0.795 ± 0.0001
voyage-3-lite	0.821 ± 0.0001

Table 13: **Bias Measurement on CMU Books dataset with perturbation of person names only.** For each show, we create K=20 perturbations by replacing person names. We compute the average cosine similarity for each perturbation pair and the standard error. The country/city/town names remain unchanged. .

person names on the text embeddings.

C.1 Perturbations of only person names

In this study, we only perturb person names and keep the location names unchanged to understand the impact of only perturbing person names. As shown in Table 13, performing only person name perturbations on book plots also reveals a significant deviation from the ideal score of 1 across all evaluated models.

C.2 Person name perturbations on text samples without mention of country/city/town names

In this section, we investigate impact of person name perturbations when using samples which don't have mention of any country/city/town etc. names. The objective is to minimize the impact of these variables and study divergence solely w.r.t person names. As shown in Table 15, benchmarking on the CMU Book dataset on samples without having any mention of country/city/town etc. also reveals a significant deviation from the ideal score of one in cosine similarity across all evaluated models when only person names are perturbed.

C.3 Bias measurement with person name perturbations from the same geographical area

In previous studies, we perturbed names by replacing them from a diverse set of person names. In this study we investigate whether the issue of divergence in embeddings persists when all the perturbed names are from the same geography. This study aims to minimize the impact of cultural differences in analysis in text-embeddings. Table 16 shows the country wise names used for benchmarking. In tables 10, 11, and 12, we observe that the divergence issue persists even when the replaced names belong to the same geography. This demonstrates that the issue is not present in names from certain cultures, cross-culture, but is universal in the sense that the name bias issue occurs in a very broad sense.

D Similarity Heatmaps

In this section, we show examples of cosine similarity heatmaps based upon embeddings generated by different text-embedding models. We use the following example:

CHARACTER_NAME, a seasoned physician, meticulously analyzed a patient's intricate heart condition. He later realised she was his school friend.

To obtain different perturbations, we replace **"CHARACTER_NAME"** with different person names and generate embedding for each of the perturbation. The similarity heatmaps are present in figs. 1 and 2. The heatmaps clearly reveal that only changing the person names can significantly impact the text embeddings. This suggests that the text embedding model is highly sensitive to the specific names used within the text, even when the overall context and meaning remains completely unchanged. These kind of variations can lead to misleading results in various downstream tasks. For example, if the goal is to cluster documents into topics, changing the person names could lead to different clusters being formed, even if the underlying topics are the same. Similarly, if the text embedding model is used to classify documents as positive or negative, changing the person names could lead to different classifications being assigned, even if the overall sentiment and theme of the text remains the same.

E Length-wise results for the CMU movie benchmarking

We compute results based on sentence lengths (number of words) for CMU movie benchmarking dataset that was studied in Section 3. The results are presented in Table 17. Each bin contains approximately the same number of samples. We observe that name-bias exists at different lengths.

Person names	Aaron, Adrian, Aiden, Akira, Alex, Alexander, Alfred, Anders, Andreas, Andrew, Anthony, Archer, Arthur, Ayden, Benjamin, Bernard, Blake, Boris, Bradley, Brandon, Brayden, Brian, Caleb, Cameron, Carlos, Carl, Charles, Charlie, Christopher, Connor, Cooper, Daichi, Daniel, David, Dean, Dennis, Dylan, Edward, Elijah, Elliot, Emil, Eric, Ethan, Evan, Ezra, Fabian, Felix, Finn, Francis, Gavin, George, Giovanni, Gregory, Haakon, Han, Harry, Hayden, Henry, Hiroki, Hugo, Hunter, Ian, Isaac, Ivan, Jack, Jacob, Jake, James, Jason, Jasper, Jayden, Jeremy, Jesse, Jin, Joaquim, Johan, John, Jonathan, Jordan, Joseph, Joshua, Juan, Kai, Kaiden, Kazuma, Keanu, Ken, Kenneth, Kevin, Liam, Logan, Lucas, Luis, Luke, Luka, Magnus, Mark, Martin, Mateo, Matthew, Max, Maximilian, Michael, Mikael, Nathan, Nathaniel, Nicolas, Noah, Oliver, Oscar, Owen, Pablo, Patrick, Paul, Pedro, Peter, Phillip, Phoenix, Rafael, Rajiv, Ralf, Ramón, Raphael, Ravi, Raymond, Reuben, Richard, Robert, Robin, Rohan, Roland, Ronan, Ryan, Samuel, Santiago, Sebastian, Sean, Silas, Simon, Stefan, Stephen, Thomas, Timothy, Tyler, Victor, Vincent, Walter, William, Xavier, Yan, Yang, Yao, Youssef, Zachary, Zane, Zayd, Zephyr, Zidan, Zinedine, Zubin, Alistair, Anders, Arjun, Arthur, Axel, Bartosz, Ben, Björn, Bruno, Caleb, Caoimhín, Cillian, Cormac, Daisuke, Damien, Darius, Deniz, Dorian, Eamon, Emile, Enzo, Fionn, Florian, Gabriel, Gideon, Gustaf, Hassan, Héctor, Igor, Ishaan, Ivan, Jasper, Kai, Leo, Levi, Liam, Luca, Lucian, Luis, Magnus, Marcel, Matteo, Max, Milan, Noah, Oliver, Oscar, Otto, Pavel, Quentin, Rafael, Ravi, Rémy, Ren, Robin, Samuel, Santiago, Sebastian, Silas, Soren, Theo, Thomas, Tristan, Viktor, William, Xavier, Yannik, Zane, Aditya, Ajeet, Ajit, Akash, Amar, Amit, Arjun, Aryan, Ashish, Avinash, Bharat, Bhuvan, Chirag, Darshan, Dev, Dheeraj, Dhruv, Gaurav, Harsh, Harsha, Hemant, Ishan,Shubham, Karan, Karthik, Kumar, Manav, Manoj, Mihir, Nikhil, Niranjan, Nivaan, Pradeep, Pranav, Raj, Rajeev, Rahul, Ramesh, Ranjit, Ravi, Rohan, Rohit, Roop, Sachin, Sandeep, Sanjay, Sanket, Sarthak, Satish, Shaan, Shahrukh, Shankar, Sharad, Shivam, Siddhant, Siddharth, Soham, Somesh, Suresh, Tejas, Uday, Varun, Vijay, Vikram, Vinay, Vishal, Yash, Yogesh, Yuvraj, Adil, Amine, Anas, Fayçal, Hakim, Hicham, Mazen, Mehdi, Nassim, Rafik, Sami, Sofiane, Tarik, Yacine, Yassine, Abiodun, Ade, Adekunle, Adewale, Ayodeji, Chidi, Chijioke, David, Ebuka, Emeka, Godwin, Ikechukwu, Ikenna, Kolade, Kunle, Nonso, Obinna, Olamide, Olusegun, Onyeka, Paul, Peter, Samuel, Taiwo, Uche, Victor, Yemi, Yinka, Aiden, Callum, Connor, Declan, Dylan, Eoghan, Finn, Jack, James, Jamie, Jason, Jayden, Kian, Liam, Logan, Lucas, Luke, Mason, Max, Michael, Noah, Oliver, Oscar, Rory, Ryan, Samuel, Sean, Thomas, William, Charlie, Freddie, George, Harry, Jacob, Leo, Oliver, Oscar, Teddy, Arthur, Freddie, George, Harry, Jacob, Leo, Oliver, Oscar, Teddy, Aiden, Alexander, Charlie, Ethan, Jacob, James, Leo, Mason, Michael, Noah, Oliver, William, Benjamin, Charlie, Jacob, Leo, Oliver, Oscar, Thomas, William, Aiden, Charlie, Ethan, Jacob, Leo, Oliver, Oscar, Thomas, William, Shrey,Venkatesh,Nguyen,Vishwanathan , Priya, Patricia, Jennifer, Linda, Barbara, Susan, Camille, Sophie, Julie, Claire, Yuki, Sakura, Hana, Aiko, Emi, Li, Xiao, Mei, Fang, Jing, Maria, Ana, Isabel, Carmen, Dolores, Amina, Layla, Nadia, Olga, Irina, Svetlana, Ekaterina, Giulia, Francesca, Anna, Elena, Heidi, Greta, Lena, Marta, Sofia, Valentina, Martina, Paula, Clara, Laura, Mia, Emily, Sophia, Charlotte, Anita, Kavita, Lalita, Meena, Lucy, Megan, Hannah, Jessica, Amelia, Chloe, Manon, Lea, Elodie, Amandine, Haruka, Miyu, Rina, Yuna, Nao, Chen, Hua, Ling, Qing, Yan, Lucia, Pilar, Rosa, Nour, Sara, Hiba, Mona, Rania, Anastasia, Natalia, Daria, Polina, Vera, Mariana, Gabriela, Beatriz, Rafaela, Camila, Juliana, Evelyn, Amanda, Milla, Ines, Susana, Leonor, Bianca, Livia, Helena, Marina, Fernanda, Eduarda, Victoria, Andressa, Denise, Raquel, Isis, Elisa, Julia, Luana, Milena, Yasmin, Alessandra, Claudia, Veronica, Larissa, Bia, Silvia, Vanessa, Leticia, Nicole, Daniele, Eva, Alice, Milena, Leonie, Mila, Lisa, Sarah, Emma, Helena, Anja, Tina, Ingrid, Lucija, Noor, Samira, Dana, Kalila, Arwa, Eman, Latifa, Nahla, Sang, Jin, Hye, Soo, Mi, Eun, Yeon, Ji, Sun, Abeba, Hadia, Fatou, Maimouna, Nia, Asha, Kamaria, Mira, Joan, Fiona, Leanne, Orla, Ava, Siobhan, Niamh, Sienna, Poppy, Lara, Freya, Florence, Rosie, Summer, Ivy,Sunidhi, Amara, Chidinma, Ngozi, Sunaina, Matilda, Harper, Willow, Aarushi, Ananya, Bhavna, Chandni, Deepa, Esha, Hina, Sneha, Jaya, Kiran, Lata, Maya, Nisha, Shrishti, Isabella, Saanvi, Drishti
Country Names	Afghanistan, Albania, Algeria, Andorra, Angola, Antigua and Barbuda, Argentina, Armenia, Australia, Austria, Azerbaijan, Bahamas, Bahrain, Bangladesh, Barbados, Belarus, Belgium, Belize, Benin, Bhutan, Bolivia, Bosnia and Herzegovina, Botswana, Brazil, Brunei, Bulgaria, Burkina Faso, Burundi, Cabo Verde, Cambodia, Cameroon, Canada, Central African Republic, Chad, Chile, China, Colombia, Comoros, Congo, Costa Rica, Croatia, Cuba, Cyprus, Czech Republic, Denmark, Djibouti, Dominica, Dominican Republic, Ecuador, Egypt, El Salvador, Equatorial Guinea, Eritrea, Estonia, Eswatini, Ethiopia, Fiji, Finland, France, Gabon, Gambia, Georgia, Germany, Ghana, Greece, Grenada, Guatemala, Guinea, Guinea-Bissau, Guyana, Haiti, Honduras, Hungary, Iceland, India, Indonesia, Iran, Iraq, Ireland, Israel, Italy, Jamaica, Japan, Jordan, Kazakhstan, Kenya, Kiribati, Kuwait, Kyrgyzstan, Laos, Latvia, Lebanon, Lesotho, Liberia, Libya, Liechtenstein, Lithuania, Luxembourg, Madagascar, Malawi, Malaysia, Maldives, Mali, Malta, Marshall Islands, Mauritania, Mauritius, Mexico, Micronesia, Moldova, Monaco, Mongolia, Montenegro, Morocco, Mozambique, Myanmar, Namibia, Nauru, Nepal, Netherlands, New Zealand, Nicaragua, Niger, Nigeria, North Korea, North Macedonia, Norway, Oman, Pakistan, Palau, Panama, Papua New Guinea, Paraguay, Peru, Philippines, Poland, Portugal, Qatar, Romania, Russia, Rwanda, Saint Kitts and Nevis, Saint Lucia, Saint Vincent and the Grenadines, Samoa, San Marino, Saudi Arabia, Senegal, Serbia

Table 14: Universe of names used for replacement in benchmarking.

The impact in general is slightly higher on smaller texts.

F Perturbation results on STS dataset

We conducted the benchmarking experiment of Section 3 on the STS SummEval dataset that we originally used in Sec. 5.2 (considering the original texts instead of summaries), and performing text perturbation using the same methodology that was used in the benchmarking. We observe in Table 19 that the name bias holds on this dataset as well.

Model Name	Cosine sim per perturbation pair
all-mpnet-base-v2	0.796 ± 0.0002
all-distilroberta-v1	0.803 ± 0.0002
all-MiniLM-L6-v2	0.731 ± 0.0003
gemini	0.906 ± 0.0001
multi-qa-distilbert-cos-v1	0.766 ± 0.0002
paraphrase-MiniLM-L6-v2	0.758 ± 0.0004
distiluse-base-multilingual-cased-v1	0.825 ± 0.0003
distiluse-base-multilingual-cased-v2	0.828 ± 0.0003
paraphrase-multilingual-MiniLM-L12-v2	0.770 ± 0.0004
msmarco-distilbert-cos-v5	0.747 ± 0.0002
multi-qa-mpnet-base-cos-v1	0.778 ± 0.0002
voyage-3-lite	0.810 ± 0.0001

Table 15: **Bias Measuremenent on CMU Books dataset on samples without mention of country/city/town names.** Perturbation of person names only. For each show, we create K=20 perturbations by replacing person names. We compute the average cosine similarity for each perturbation pair and the standard error.

Country	Person Names
France	Max, Tom, Léo, Noé, Paul, Jules, Hugo, Arthur, Louis, Clément, Jean-Baptiste, Jean-Pierre, Jean-Paul, Charles-Henri, François-Xavier, Constantin, Gaspard, Côme, Yanis, Kilian, Maël, Thibault, Raphaël, Jérémie, Vincent, Antoine, Pierre, Louis, Jacques, Baptiste, Émile, Gustave, Henri, Laurent, Marcel, Nicolas, Olivier, Pascal, Quentin, Rémi, Sébastien, Théodore, Ulysse, Valentin, Wilfried, Xavier, Yves, Zacharie, Adrien, Bernard, Eva, Zoé, Jade, Lou, Alice, Chloé, Léa, Lina, Louise, Éléonore, Solène, Héloïse, Camille, Marie, Jeanne, Sophie, Claire, Isabelle, Ambre, Lilou, Maëlys, Victoire, Clémence, Valentine, Juliette, Aurélie, Angélique, Amandine, Brigitte, Catherine, Delphine, Édith, Fanny, Gabrielle, Hélène, Inès, Joséphine, Karine, Laure, Manon, Nathalie, Océane, Pascale, Quitterie, Rosalie, Stéphanie, Thérèse, Ursule
India	Aarav, Aditya, Aryan, Ayush, Dev, Ishaan, Ramesh, Krishna, Mihir, Rohan, Sahir, Samarth, Shaurya, Vihaan, Vrijesh, Aakash, Advait, Vinayak, Atharv, Venkatesh, Dhruv, Eshan, Hrithik, Kabir, Karan, Krish, Mahesh, Nakul, Pranav, Rudra, Siddharth, Soham, Tanmay, Uday, Vaibhav, Vedant, Vikram, Yash, Yuvraj,Sachin, Ahaan, Gaurav, Arjun, Daksh, Devansh, Ishan, Vishwanathan, Mayank, Parichay, Krishnanshu, Sahir, Rishi, Samyak, Brajesh, Vivaan, Ayan, Rudra,Rakesh, Zain, Aarohi, Bhavya, Charvi, Devika, Eshani, Falguni, Garima, Harini, Ishita, Jahnvi, Kavya, Lavanya, Madhavi, Niharika, Ojasvi, Prisha, Qara, Radhika, Saanvi, Tara, Urvashi, Vanya, Wamika, Xara, Yamini, Zara, Anvi, Bhumika, Chaitali, Dharini, Ekta, Fiza, Gauri, Himani, Ira, Jiya, Kriti, Lata, Meera, Nisha, Oviya, Pallavi, Rhea, Sakshi, Tanisha, Uma, Vaidehi, Yashika, Zaina, Aditi
Spain	Mateo, Santiago, Lucas, Marcos, Daniel, David, Samuel, Benjamín, Ezequiel, Noé, Salvador, Ismael, Aarón, Elías, Jonás, Jeremías, Iker, Unax, Aitor, Ander, Martín, Rodrigo, Fernando, Alfonso, Enrique, Felipe, Carlos, Javier, Jorge, Luis, Antonio, José, Juan, Manuel, Pedro, Francisco, Ignacio, Rafael, Víctor, Álvaro, Diego, Gabriel, Miguel, Pablo, Ricardo, Sergio, Tomás, César, Gonzalo, Leonardo, Emiliano, Matías, Nicolás, Sebastián, Thiago, Sofía, Camila, Valentina, Martina, Emilia, Emma, Olivia, Luna, Zoe, Mia, Isabella, Victoria, Sara, Lucía, María, Laura, Paula, Andrea, Ana, Elena, Carmen, Alba, Carla, Daniela, Julia, Natalia, Ximena, Aitana, Noa, Mía, Isabel, Beatriz, Blanca, Clara, Inés, Irene, Marta, Patricia, Rocío, Silvia, Teresa, Verónica, Alicia, Amelia, Ángela, Aurora, Bárbara, Carolina, Dolores, Eva, Gloria, Lidia, Lorena, Mónica, Nuria, Olga, Raquel, Sandra,Xiomara, Yamile

Table 16: Universe of names for country wise name replacement in benchmarking experiments in Sec. C

G Semantic Similarity Task Dataset

Below we present the STS dataset used in Sec. 5.1. Each sample is a triplet of the form: $< Query, Positive\ sample, Negative\ sample >$. We present 10 samples below, and the full dataset of 50 samples is available at https://github.com/sahilm1992/name_bias.

1. **Query:** Nikolai and Deborah met on a rainy Tuesday in New York. The city's hustle and bustle couldn't dim the spark between them. Deborah, with her radiant smile and infectious laughter, had captured Nikolai's heart from the moment he saw her. Nikolai, a charming and witty gentleman, returned her affection with equal fervor.

   - **Positive:** Kashvi and Oluwafemi met on a rainy Tuesday in Northampton. The city's bustling streets couldn't dim the spark between them. Kashvi, with her radiant smile and infectious laughter, had captured Oluwafemi's heart from the mo-

model_name	cosine sim	cosine sim	cosine sim
	[22, 85]	[22, 85]	[138.66, 249]
all-distilroberta-v1	0.759 ± 0.001	0.764 ± 0.001	0.781 ± 0.001
gemini	0.887 ± 0.00	0.887 ± 0.00	0.883 ± 0.00
multi-qa-distilbert-cos-v1	0.729 ± 0.001	0.736 ± 0.001	0.733 ± 0.001
multi-qa-mpnet-base-cos-v1	0.742 ± 0.001	0.752 ± 0.001	0.735 ± 0.001
text-embedding-3-small	0.731 ± 0.001	0.743 ± 0.001	0.751 ± 0.001
text-embedding-3-large	0.755 ± 0.001	0.783 ± 0.001	0.799 ± 0.001
voyage-3-lite	0.760 ± 0.001	0.768 ± 0.001	0.753 ± 0.001
gte-base	0.936 ± 0.00	0.954 ± 0.00	0.959 ± 0.00
gte-large	0.938 ± 0.00	0.951 ± 0.00	0.956 ± 0.00
e5-base-v2	0.921 ± 0.00	0.925 ± 0.00	0.922 ± 0.00
e5-large-v2	0.918 ± 0.00	0.921 ± 0.00	0.916 ± 0.00

Table 17: **Length-wise results for the CMU movie benchmarking.** It can be seen that the name bias exists even at higher lengths of the sentences and not just in smaller length sentences.

Model Name	Euclidean Distance per perturbation pair
all-mpnet-base-v2	0.642 ± 0.0016
all-distilroberta-v1	0.641 ± 0.0015
all-MiniLM-L6-v2	0.766 ± 0.0017
gemini	0.460 ± 0.0007
multi-qa-distilbert-cos-v1	0.694 ± 0.0014
paraphrase-MiniLM-L6-v2	3.398 ± 0.0153
distiluse-base-multilingual-cased-v1	0.638 ± 0.0020
distiluse-base-multilingual-cased-v2	0.630 ± 0.0021
paraphrase-multilingual-MiniLM-L12-v2	2.726 ± 0.0108
msmarco-distilbert-cos-v5	0.742 ± 0.0016
multi-qa-mpnet-base-cos-v1	0.679 ± 0.0016
text-embedding-3-small	0.670 ± 0.0013
text-embedding-3-large	0.616 ± 0.0013
voyage-3-lite	0.647 ± 0.0010

Table 18: **Bias Measurement on CMU Book dataset with Euclidean distance as distance function between embeddings**. A distance close to 0 is better.

ment he saw her. Oluwafemi, a charming and witty gentleman, returned her affection with equal fervor.

- **Negative:** Nikolai and Deborah staying in New Jersey, once inseparable, were now worlds apart. Deborah, the trusted confidante, had betrayed Nikolai's trust, revealing his secrets to their rivals. The city's hustle and bustle mirrored the chaos within Nikolai's heart, as he grappled with the bitter reality of love turned treachery.

2. **Query:** Alejandro quickly ran to the store to buy a cold drink. He was eager to have a glass of cold drink.

    - **Positive:** Quickly, Hiroki dashed to the local market to procure some cold drinks. He was yearning for a chilled glass of cold drink.
    - **Negative:** Alejandro has stopped buying cold drinks from market. He only drinks cold drinks made at home.

3. **Query:** Mayatoshi and Alex had a deep, passionate love for each other. Their bond was unbreakable, a love that transcended all obstacles. They shared dreams, hopes, and aspirations, and their love was the foundation of their happiness.

    - **Positive:** Priyanka and Yuan were deeply in love. Their affection for each other was profound and unwavering. They shared a strong connection, a love that was the source of their joy and contentment.
    - **Negative:** Despite their intense hatred for each other, Mayatoshi and Alex were bound by a strange, twisted connection. Their animosity fueled a toxic relationship, a constant battle of wills. Their lives were intertwined, a dark, destructive dance of love and hate.

4. **Query:** Amazon and Apple are two American corporations. Amazon's main business is online shopping and Apple is a phone maker giant

    - **Positive:** Alibaba and Xiaomi are two Chinese corporations. Alibaba's main business is online shopping and Xiaomi is a producer of phones
    - **Negative:** Amazon is a river in South America. Apples are not grown in the Amazon basin.

model_name	Cosine sim per perturbation pair
all-distilroberta-v1	0.8196 ± 0.0026
gemini	0.9000 ± 0.0005
multi-qa-distilbert-cos-v1	0.8022 ± 0.0026
multi-qa-mpnet-base-cos-v1	0.7973 ± 0.0027
text-embedding-3-small	0.8612 ± 0.0007
text-embedding-3-large	0.8709 ± 0.0007
voyage-3-lite	0.8684 ± 0.0007
gte-base	0.9347 ± 0.0008
gte-large	0.9443 ± 0.0007
e5-base-v2	0.9491 ± 0.0006
e5-large-v2	0.9503 ± 0.0006

Table 19: **Perturbation based benchmarking on the STS SummEval dataset.** We follow the same evaluation setup as in Table 3. We see that the results and conclusions are similar to what we observed in Table 3.

	Bin 1	Bin 1	Bin 1	Bin 2	Bin 2	Bin 2	Bin 3	Bin 3	Bin 3
model	orig	same	anon	orig	same	anon	orig	same	anon
all-mpnet-base-v2	0.446	0.982	0.986	0.179	0.765	0.921	0.09	0.778	0.892
gte-large	0.186	0.961	0.979	0.222	0.855	0.941	0.173	0.972	0.951
gte-base	0.148	0.982	0.986	0.234	0.878	0.914	0.19	0.913	0.885
e5-base-v2	0.328	0.982	0.972	0.066	0.921	0.910	0.131	0.941	0.844
e5-large-v2	0.505	0.986	0.972	0.085	0.918	0.929	0.193	0.948	0.917
all-distilroberta-v1	0.515	1.000	0.979	0.074	0.847	0.921	0.231	0.830	0.917
all-MiniLM-L6-v2	0.100	1.000	0.955	0.078	0.746	0.832	0.000	0.747	0.743
multi-qa-distilbert-cos-v1	0.117	0.930	0.930	0.007	0.777	0.777	0.006	0.730	0.833
paraphrase-MiniLM-L6-v2	0.169	0.982	0.899	0.125	0.867	0.867	0.076	0.982	0.917
distiluse-base-multilingual-cased-v1	0.204	0.899	0.889	0.207	0.691	0.707	0.211	0.754	0.788
distiluse-base-multilingual-cased-v2	0.173	0.903	0.923	0.195	0.691	0.714	0.332	0.726	0.785
paraphrase-multilingual-MiniLM-L12-v2	0.221	0.993	0.961	0.113	0.921	0.910	0.117	0.906	0.875
msmarco-distilbert-cos-v5	0.069	0.927	0.923	0.035	0.628	0.628	0.013	0.837	0.861
multi-qa-mpnet-base-cos-v1	0.076	0.972	0.937	0.011	0.890	0.949	0.000	0.889	0.820
gemini	0.678	1.000	1.000	0.414	1.000	1.000	0.674	1.000	0.986
text-embedding-3-small	0.100	0.982	0.965	0.003	0.925	0.921	0.024	0.934	0.920
text-embedding-3-large	0.169	0.989	1.000	0.019	0.988	0.964	0.179	0.944	0.937
voyage-3-lite	0.218	1.000	1.000	0.097	0.921	0.921	0.013	0.827	0.882

Table 20: **Results on STS Task 1 divided into bins of sentences of different lengths.** In the Table columns with Bin 1 represent average lengths between $[11.99, 27.88]$, columns with Bin 2 represent average lengths between $[27.88, 35.88]$ and columns with Bin 3 represent average lengths between $[35.88, 60.33]$.

5. **Query:** Ganga and Yamuna are two mighty rivers. They are lifelines for millions of people in the region.
    - **Positive:** Yangtze is a mighty river. It is a long river and is the lifeline for millions of people in the region.
    - **Negative:** Ganga and Yamuna are two sisters. They had their schooling in the region and schooling provided a lifeline for them.

6. **Query:** Alice and Bob often helped each other financially. Recently, Alice lent Bob a significant sum of money. Bob promised to return it soon.
    - **Positive:** Yuri and Haruto frequently helped each other out, including with money. Lately, Yuri had loaned Haruto a substantial amount of money, which Haruto assured her he'd repay promptly.
    - **Negative:** Alice and Bob had a disagreement about money. Alice believed Bob owed her money, but Bob denied it.

7. **Query:** John, a renowned lawyer, is defending his client, Mike, who is accused of a serious crime. John is determined to prove Mike's innocence and secure his acquittal.
    - **Positive:** Armaan, a man falsely accused of a heinous crime, is relying on his skilled lawyer, Udit, to exonerate him. Udit is committed to presenting a strong defense and clearing Armaan's name.
    - **Negative:** John, a cunning lawyer, is

manipulating the legal system to frame Mike for a crime he did not commit. John's goal is to secure a conviction and advance his own career, regardless of the truth.

8. **Query:** Dr. Alexander, a seasoned physician, meticulously analyzed patient Sarah's intricate heart condition. He prescribed a tailored regimen of medications and rigorous lifestyle modifications to significantly improve her cardiac health.
    - **Positive:** The esteemed doctor, Dr. Yerusha, conducted a thorough assessment of patient Reyan's complex symptoms of heart. She formulated a precise treatment plan, incorporating medications and day to day lifestyle changes, to alleviate Reyan's debilitating heart condition.
    - **Negative:** Dr. Alexander, a renowned doctor and surgeon, executed a high-risk heart surgical procedure on patient Sarah. After the complex operation Sarah did not recover.

9. **Query:** Mr. Smith, a dedicated teacher, guided his students, including the bright young minds of Miller and Pristina, towards academic excellence.
    - **Positive:** Mr. Yang, a committed educator, mentored his students, including the talented Shruti and Ren, to achieve academic success.
    - **Negative:** Mr. Smith , a rigid and punitive teacher, often unfairly targeted mischievous students like Miller and Pristina.

10. **Query:** Martinez gently examined the injured bird. He gave it food.
    - **Positive:** Yohan tenderly inspected the wounded bird and gave it a meal to eat.
    - **Negative:** The skilled hunter Martinez tracked the injured bird. He captured it for food.

H Length-wise performance on Semantic Similarity Task 1

To further understand the impact of the length of the sentences in STS task 1, in Table 20, we divide the results into bins based on the word count in the sentences. Each bin has almost the same number of samples. Word count of a sample is computed as the average number of words in the triplet <query, positive, negative>. The results show that while the results are much more pronounced on smaller length sentences, the results still exist in higher length bins.

I Impact of replacement of non-entity words

In this section, we study the impact of replacing other words in text except named entities, i.e, person names or country names. Towards this, we choose the CMU Book dataset as in Table 4 and generate 5 perturbations for each plot by replacing words that are not entities, such as verbs/common nouns, etc. We perform the same number of perturbations as there are person names and country names in the plot. After perturbation, we compute the cosine similarity between each perturbed plot and the original plot. We report the average cosine similarity along with the standard error in Table 22. We can observe that all the cosine similarities observed now are substantially higher than the observed cosine similarities when names were replaced in Section 3.

J Example of Semantic Similarity post-anonymization

In Table 21, we show impact of anonymization on STS tasks on embeddings crated by Open AI's $text-embedding-3-small$ model. We observe that in all cases performance after anonymization is superior. Specifically, post anonymization, we obtain relatively higher score for positive samples and lower for negative samples.

K Impact of Anonymization Strategy: Removal versus Replacement

This section investigates the effectiveness of removal of names vs. replacement of names in text for anonymization. In the replacement strategy, we replace names with non-identifying placeholder names instead of removing them from text. Example: person names with 'CHAR_ID', location names with 'LOC_ID' etc. Here ID can be replaced with $\{A, B, C \cdots\}$ or $\{1, 2, 3 \cdots\}$ etc. The detailed prompt is present in Table 24. In Table 25 we demonstrate that removal of names marginally outperforms replacement in the STS task. In the

	Query	Pos/Neg	Sim score	Label
Original	Alejandro quickly ran to the store to buy a cold drink. He was eager to have a glass of cold drink.	**POS:** Quickly, Hiroki dashed to the local market to procure some cold drinks. He was yearning for a chilled glass of cold drink.	0.66	1
		NEG: Alejandro has stopped buying cold drinks from market. He only drinks cold drinks made at home.	0.72	0
Anonymized	quickly ran to the store to buy a cold drink. He was eager to have a glass of cold drink.	**POS:** Quickly, dashed to the local market to procure some cold drinks. He was yearning for a chilled glass of cold drink.	0.83	1
		NEG: has stopped buying cold drinks from market. He only drinks cold drinks made at home.	0.57	0
Original	Ganga and Yamuna are two mighty rivers. They are lifelines for millions of people in the region.	**POS:** Yangtze is a mighty river. It is a long river and is the lifeline for millions of people in the region.	0.63	1
		NEG: Ganga and Yamuna are two sisters. They had their schooling in the region and schooling provided a lifeline for them.	0.73	0
Anonymized	and are two mighty rivers. They are lifelines for millions of people in the region.	**POS:** is a mighty river. It is a long river and is the lifeline for millions of people in the region.	0.76	1
		NEG: and are two sisters. They had their schooling in the region and schooling provided a lifeline for them.	0.46	0

Table 21: Example demonstrating impact of anonymization on semantic similarity using embeddings created by Open AI's *text-embedding-3-small* model. The text in color blue and red refer to the positive and negative paragraphs respectively.

model name	Avg Cosine Similarity
all-mpnet-base-v2	0.975 ± 0.002
all-distilroberta-v1	0.975 ± 0.002
all-MiniLM-L6-v2	0.960 ± 0.002
gemini	0.987 ± 0.001
multi-qa-distilbert-cos-v1	0.966 ± 0.002
paraphrase-MiniLM-L6-v2	0.955 ± 0.004
distiluse-base-multilingual-cased-v1	0.972 ± 0.003
distiluse-base-multilingual-cased-v2	0.974 ± 0.002
paraphrase-multilingual-MiniLM-L12-v2	0.971 ± 0.003
msmarco-distilbert-cos-v5	0.963 ± 0.002
multi-qa-mpnet-base-cos-v1	0.974 ± 0.002
text-embedding-3-small	0.967 ± 0.001
text-embedding-3-large	0.969 ± 0.001
voyage-3-lite	0.971 ± 0.001
gte-large	0.993 ± 0.000
gte-base	0.991 ± 0.000
e5-base-v2	0.989 ± 0.000
e5-large-v2	0.989 ± 0.000

Table 22: **Impact of replacement of non-entity words**.

Model Name	Size
all-mpnet-base-v2	420 MB
all-distilroberta-v1	290 MB
all-MiniLM-L6-v2	80 MB
multi-qa-distilbert-cos-v1	250 MB
paraphrase-MiniLM-L6-v2	90.9 MB
distiluse-base-multilingual-cased-v1	480 MB
distiluse-base-multilingual-cased-v2	480 MB
paraphrase-multilingual-MiniLM-L12-v2	420 MB
msmarco-distilbert-cos-v5	265 MB
multi-qa-mpnet-base-cos-v1	438 MB
gte-large	670 MB
gte-base	219 MB
e5-base-v2	438 MB
e5-large-v2	1340 MB

Table 23: Model information for open source models.

context of replacement strategy, one should note that the quality of embeddings derived is sensitive to the specific replacement placeholder terms used. For instance, substituting character names with with different placeholders such as "CHAR_A" / "CHARACTER_B" / "CHARACTER_1" or location names with "LOC_1" / "LOC_B" can impact the resulting embeddings differently. In order to mitigate this sensitivity and ensure consistent results and also based upon our findings we recommend using the name removal strategy for anonymization to mitigate name bias.

Replace person names, organizations and locations	Given below text, please convert all Person names(which are Proper Nouns) to a UNIQUE ID such as CHAR_A, CHAR_B, CHAR_C etc.. Keep it unique and for each UNIQUE Person name(which is a Proper Noun) use a UNIQUE ID. DO NOT KEEP THE ORIGINAL Person Names(which are Proper Nouns) in the generated paragraph text. Next, Replace all occurences City/Country/Village/Town/River/Continent etc. names which are PROPER NOUNS to a UNIQUE ID such as LOC_A, LOC_B, LOC_C, LOC_D etc.. Next, Replace all occurences of company/organization names which are PROPER NOUNS to a UNIQUE ID such as ORG_A, ORG_B, ORG_C, ORG_D etc.. Do not replace monument/landmark names like Eiffel tower etc. Output contains the modified text only.... The text is provided below ::::

Table 24: Prompt for Anonymization using replacement strategy described in Sec. K

Model	AUC ROC Original	AUC ROC Anonymization(Default)	AUC ROC Anonymization(Replacement)
all-mpnet-base-v2	0.19	0.980 ± 0.0071	**1.000** ± **0.0**
all-distilroberta-v1	0.36	**0.975** ± **0.0106**	0.945 ± 0.0106
all-MiniLM-L6-v2	0.09	**0.990** ± **0.0071**	0.970 ± 0.0071
gemini	0.71	**1.000** ± **0.0**	**1.000** ± **0.0**
multi-qa-distilbert-cos-v1	0.07	**0.970** ± **0.0071**	0.950 ± 0.0
paraphrase-MiniLM-L6-v2	0.14	0.980 ± 0.0	**0.990** ± **0.0071**
distiluse-base-multilingual-cased-v1	0.27	**0.940** ± **0.0**	0.935 ± 0.0106
distiluse-base-multilingual-cased-v2	0.26	**0.960** ± **0.0**	0.940 ± 0.0212
paraphrase-multilingual-MiniLM-L12-v2	0.21	0.990 ± 0.0	**1.000** ± **0.0**
msmarco-distilbert-cos-v5	0.10	**0.955** ± **0.0035**	0.875 ± 0.0035
multi-qa-mpnet-base-cos-v1	0.08	**1.000** ± **0.0**	0.985 ± 0.0035
text-embedding-3-small	0.12	1.000 ± 0.0	0.970 ± 0.0071
text-embedding-3-large	0.21	**1.000** ± **0.0**	**1.000** ± **0.0**
voyage-3-lite	0.18	**1.000** ± **0.0**	0.980 ± 0.0141

Table 25: Comparison of Removal based vs Replacement based Anonymization on Semantic Similarity task of Sec. 5.1. These results were computed for 10 samples presented in Sec. G.

L Implementation Details

L.1 Model Information and Computational budget

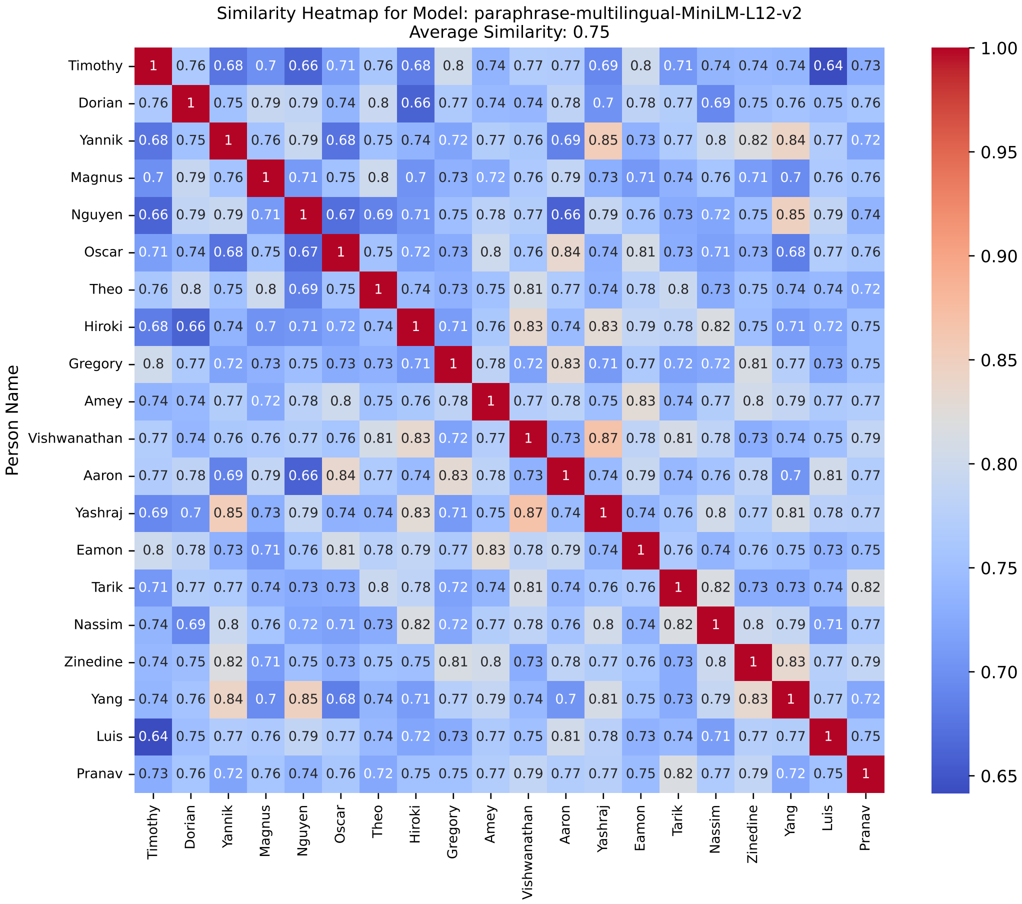
In Table 23, we present the model size of different open source models used. For our experiments, we consumed approximately 40 GPU hours with one 32 GB GPU.

L.2 Packages used

We used scikit-learn (Pedregosa et al., 2011) package for computing metrics such as cosine similarity and AUC-ROC.

L.3 Terms and License

For our implementation, we use sentence transformers library (Reimers, 2019), Gemini API, and OpenAI API which are under Apache License, Version 2.0. The Voyage API is licensed under MIT license. All the artifacts used in this paper are available for non-commercial scientific use.



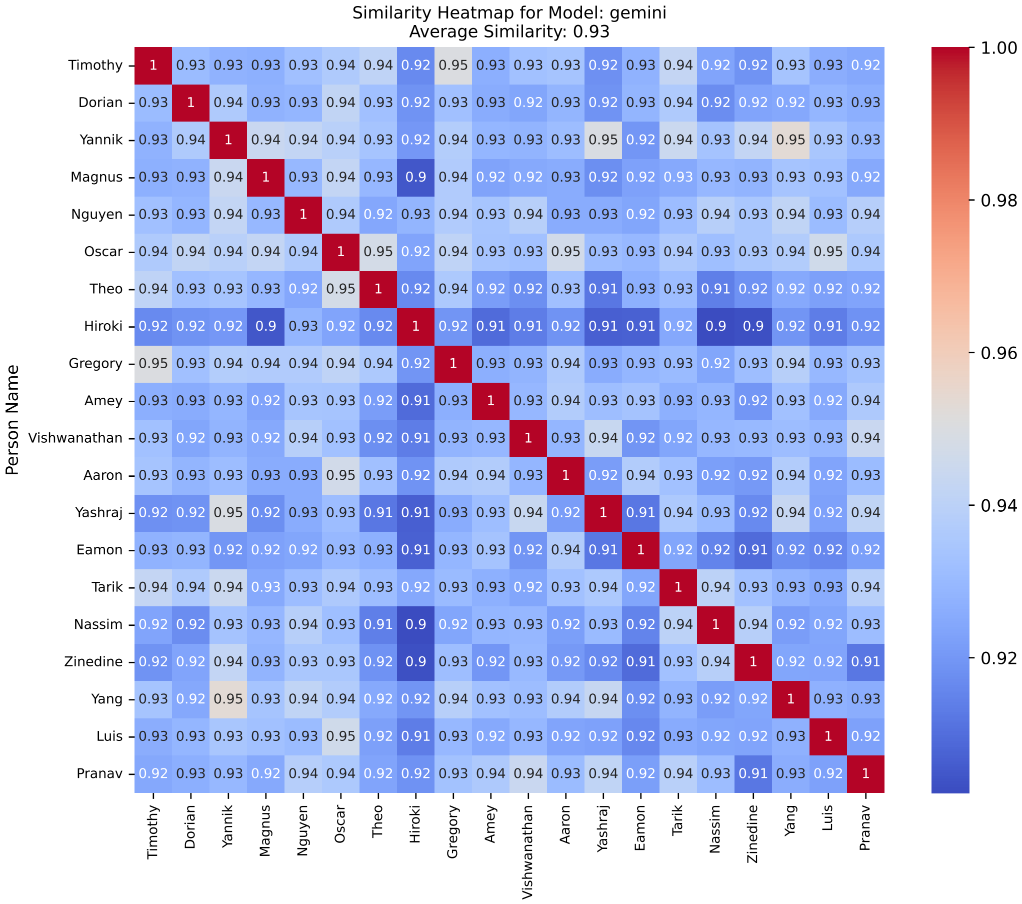


Figure 1: Cosine Similarity Heatmaps for example in Sec. D

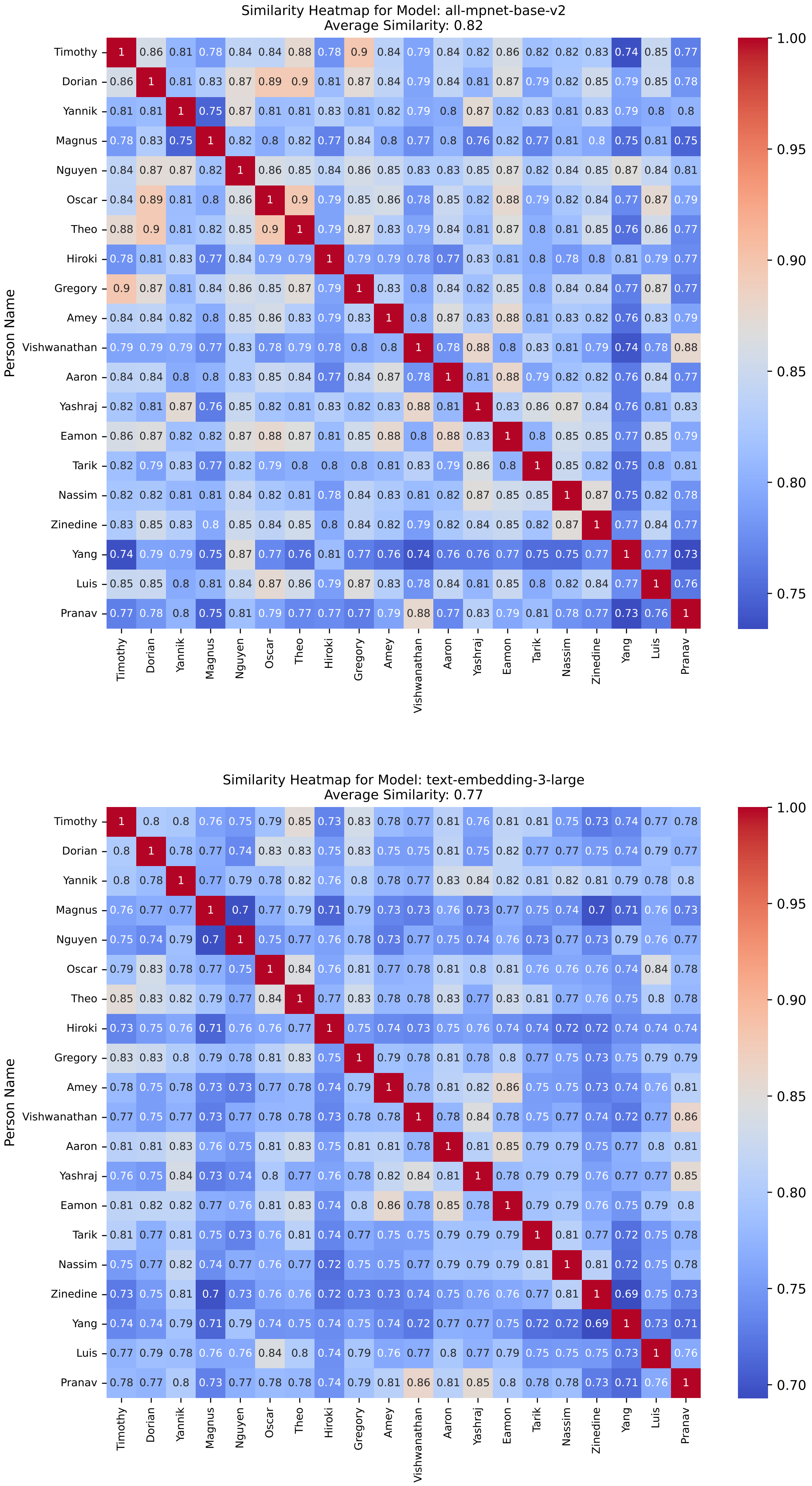


Figure 2: Cosine Similarity Heatmap for example in Sec. D