Selecting Shots for Demographic Fairness in Few-Shot Learning with Large Language Models

Carlos Aguirre, Kuleen Sasse, Isabel Cachola and **Mark Dredze**
Center for Language and Speech Processing
Johns Hopkins University
caguirre@cs.jhu.edu

Abstract

Recently, work in NLP has shifted to few-shot (in-context) learning, with large language models (LLMs) performing well across a range of tasks. However, while fairness evaluations have become a standard for supervised methods, little is known about the fairness of LLMs as prediction systems. Further, common standard methods for fairness involve access to models weights or are applied during finetuning, which are not applicable in few-shot learning. Do LLMs exhibit prediction biases when used for standard NLP tasks?

In this work, we analyze the effect of shots, which directly affect the performance of models, on the fairness of LLMs as NLP classification systems. We consider how different shot selection strategies, both existing and new demographically sensitive methods, affect model fairness across three standard fairness datasets. We find that overall the performance of LLMs is not indicative of their fairness, and furthermore, there is not a single method that fits all scenarios. In light of these facts, we discuss how future work can include LLM fairness into evaluations.

1 Introduction

Historically, evaluation of machine learning systems concerned only overall performance: how well did a trained system do on a held-out test set. More recently, practitioners have realized that dataset-level scores can mask uneven performance across different sets of data points (Barocas et al., 2019). This can be especially problematic when performance varies significantly between demographic groups, such as systems that do relatively worse on underrepresented and historically oppressed demographic groups (e.g., Zhang et al., 2020). These systems are often called unfair or biased. Fairness has implications for the quality of the user experience and system robustness, and can measure user experience in a manner not reflected by overall metrics. Additionally, fairness may have legal ramifications when AI regulations intersect with laws against discrimination (e.g., Kim, 2022). To address these disparities, researchers have developed methods for fairness that may be applied to training objectives, alignment after training, and evaluation metrics (Barocas et al., 2019).

A new approach to prediction relies on large language models (LLMs), in which an instance is accompanied by a prompt and an LLM relies on in-context learning to make a prediction (Brown et al., 2020). This type of learning, which requires no fine-tuning or other gradient updates, uses just a few examples at inference time as a “prompt” to guide inference on a final instance. Because in-context learning relies only on a few text examples during inference, the content of these examples can be very important for the quality of the emitted output (Dong et al., 2022). While prior work has shown that LLMs perform surprisingly well on various prediction tasks, models are measured once again on overall performance alone, not fairness, despite an understanding of the variable nature of LLM behavior (Chang and Bergen, 2023). To date, little to no work has measured the fairness of LLMs as prediction systems, despite numerous studies showing inherent biases in the generations of LLMs (Stanczak and Augenstein, 2021; Si et al., 2022). Furthermore, traditional methods for addressing unfair models, whether pre-, in-, or post-training, are not applicable to LLMs as the data they’re trained on is often proprietary, pre-training them is expensive, and many leading models are closed source.

Relying on the importance of the content of examples in few-shot learning, we analyze the fairness of LLMs as prediction systems considering how different demonstration selection methods affect the resulting social fairness of the model in classification tasks. Experiments with 7 popular models (Table 1) across 3 datasets find that

Proceedings of the Third Workshop on NLP for Positive Impact, pages 50–67
November 15, 2024 ©2024 Association for Computational Linguistics

LLMs are unfair predictors. We consider two types of demonstration selection methods to mitigate this unfairness: semantic and demographic-based, some novel and others from prior work. We conduct an in-depth analysis of the performance and fairness of each demonstration selection method for each model. While these selection methods can improve fairness in unpredictable scenarios, these inconsistent improvements across datasets and models suggest that future work is needed to better understand how to achieve prediction fairness of LLMs beyond shot selection, as well as methods that create more reliable and demographically-stable LLMs.

2 Related Work

In-Context Learning. Large Language Models are effective in a large number of classification and generative tasks (Devlin et al., 2019a; Radford et al., 2019; Liu et al., 2019a; Lewis et al., 2019). While finetuning a pretrained model is a popular paradigm (Devlin et al., 2019a), finetuning large models can be cost-prohibitive because of the compute required to do so. Furthermore, finetuning requires additional task-specific labeled data, which can also be prohibitively expensive to collect. Brown et al. (2020) evaluated *in-context learning*, or few-shot learning, for LLMs, a learning paradigm in which the model is given a few examples, or demonstrations, of a task and is then asked to complete the final example. In-context learning has shown impressive results in a variety of tasks, including question answering, translation, and natural language inference (Brown et al., 2020).

Work on in-context learning has focused on writing better prompts (Wei et al., 2022; Min et al., 2021a; Holtzman et al., 2021; Zhao et al., 2021), choosing better demonstrations (Liu et al., 2021; Rubin et al., 2021), and training with an in-context learning objective (Min et al., 2021b; Chen et al., 2021). There have also been explorations of the sensitivities of in-context learning, such as the format of the prompts (Gao et al., 2021a; Jiang et al., 2019) or the order of the demonstrations (Lu et al., 2021). However, prior work has not studied the effect of demonstration choice on social fairness, only on overall performance (Dong et al., 2022). Other work, like Ma et al. (2023) has evaluated the *label fairness*, i.e. performance differences across different labels or classes in a multi-class prediction setting, of LLMs in in-context learning by creating a system that chooses prompts to create a "fair" demonstration. Similar to our work, they focused on shot or demonstration choice and found that shot selection matters for performance. Thus, given the minimal amount of data used for in-context learning, we suspect that the choice of demonstrations has an effect on the social fairness of the model's output.

Social Fairness with Large Language Models. Work that identifies and measures the biases of language models have classified these harms in two general categories: *allocation* and *representation* harm (Stanczak and Augenstein, 2021). Representational harms happen when harmful concepts or relations are associated with demographic groups by a model; in language models these are often measured via token embeddings and model parameters with fill-in the blank, or complete the sentence templates (e.g., Nadeem et al., 2021; Nangia et al., 2020). Most bias studies in NLP have focused on representational harms: many studies have demonstrated how generations from LLMs exhibit bias towards specific groups, or generate text that can be considered offensive, harmful or toxic (Dodge et al., 2021; De-Arteaga et al., 2019; Bender et al., 2021; Nadeem et al., 2021; Si et al., 2022), generations from LLMs are more likely to generative negative sentiment for refugees, disabled people, AAVE sentences, nonbinary, muslim and women (Magee et al., 2021; Groenwold et al., 2020; Sheng et al., 2019). In this area, research has also investigated how shot selection and ordering affects the bias of models, finding that random ordering and representative shots helps reduce bias (Si et al., 2022). To understand the underlying bias source in the behavior of these models, researchers have evaluated the generations of LLMs under different conditions, like size and training procedure (Baldini et al., 2022; Tal et al., 2022; de Vassimon Manela et al., 2021; Nangia et al., 2020).

On the other hand, allocational harms are reflected on performance differences on data associated with different demographic groups (Stanczak and Augenstein, 2021), also known as fairness. Little work has focused on allocation harms from in-context learning in LLMs for classification settings. Salewski et al. (2023) found that impersonating roles improves performance for in-context learning on LLMs: impersonating an expert in a task can improve performance of the model for that task; however, these impersonations can also reveal biases in models by finding disparate performances from impersonating different roles, e.g. better performance when impersonating men than women. Perhaps the most related work is Zhang et al. (2022a), who investigates fairness re-programming techniques for models that cannot be re-trained or finetuned, e.g. in-context learning LLMs. They append token perturbations to the prompt, *fairness triggers*, that are learned from a helper model and show that they can decrease performance differences across demographic groups. We, instead, focus on investigating the role of choice of demonstrations or shots in the performance differences of LLMs on in-context learning settings.

3 Data

We consider three text classification datasets that include demographic information to evaluate the fairness of language models with regard to demographics: Bias in Bios (De-Arteaga et al., 2019), Twitter Sentiment (Blodgett et al., 2016), and HateXplain (Mathew et al., 2021).

Bias in Bios (demographics: gender) is a collection of English documents from CommonCrawl that contain biographies. The task is to predict the occupation from the biography, (*MIT license*.) De-Arteaga et al. (2019) found gender bias present in models for this task. Following Kaneko et al. (2022), we measure gender bias by comparing the relative performance of models across biographies written about men and women. We select professions (labels) that had more than 1000 examples of biographies for each gender in the test set.[1] This yields the following 8 labels: Attorney, Dentist, Journalist, Photographer, Physician, Professor, Psychologist, and Teacher. We randomly selected 500 for each gender from each profession to create a test set of 8,000 biographies. We then created a training set of 183,638 biographies by selecting all the biographies from the original train split with the professions listed above.

Twitter Sentiment (demographics: race) is a collection of English tweets where the task is to predict binary sentiment in a tweet. Tweets have also been annotated with a binary attribute corresponding to online text dialects: African-American English (AAE) or Standard American English (SAE), which has been previously correlated with parts-of-speech tagging performance difference in prior work (Blodgett et al., 2016). We use these text dialects as proxies for race and measure racial bias by comparing the relative performance of sentiment classification across the dialects, similar to Shen et al. (2022). To construct the dataset we follow Han et al. (2022) (*APACHE licence, v2.0*.) We then select 40k and 2k random tweets from each combination of dialect and sentiment for train and test, creating a train set with 160k examples and test set of 8k.

HateXplain (demographics: race) is a collection of posts from Gab and Twitter annotated with toxicity and hate speech labels, as well as demographic labels for the target group of the hate speech. While prior work has shown that there are performance differences for detecting hate speech for different target groups based on gender, religion, and race, we experiment only on race as it was the demographic characteristic with the reported highest disparities (Baldini et al., 2022, *MIT license*). We remove Indigenous and Indian examples from our race demographics as they do not appear in all data splits. To construct the dataset, we followed a similar procedure to Ye et al. (2021): we first reduced the space from multiclass to binary classification by combining the "offensive" and "hatespeech" labels to a singular "toxic" label while keeping the "normal" class the same. Because of HateXplain has multiple annotators per example for the labels and demographics, we take the majority label and the majority demographic. If there is not a majority in either, we discard the example.

4 Methods

We measure the effect of different demonstration selection methods on prediction fairness of LLMs. We hypothesize that, similar to how the choice of demonstrations has been shown to have an effect on performance, different methods of demonstration selection will affect social fairness of the model. This section describes the models evaluated, prompts, demonstration selection methods, and definitions of performance and fairness. Overall, we conduct experiments in 36 setups (3 tasks, 12 models), using 6 demonstration selection strategies.

4.1 Models

We consider the fairness of several different LLMs, including open and closed source models. We consider both pretrained only (LLaMA (Touvron

[1]i.e. professions with at least 1000 men and 1000 women

[2]https://openai.com/blog/chatgpt

Access Type	Model Name	Training Type	Parameters
Open Source	LLaMA	Pretrained	13B & 65B
	LLaMA2	Pretrained & chat	13B & 70B
	Alpaca	Instruction-tuned	7B & 13B
	UL2	Pretrained	20B
	Flan-UL2	Instruction-tuned	20B
Closed Source	`davinci-003`	Instruction-tuned	175B
	`gpt-3.5-turbo`	Instruction-tuned[2]	-

Table 1: The LLMs evaluated in this work.

et al., 2023a), UL2 (Tay et al., 2023), Llama2 (Touvron et al., 2023b)) and finetuned variants (Alpaca (Taori et al., 2023), Flan-UL2 (Chung et al., 2022), Llama2-chat). We also consider two model sizes to observe the effects of size on fairness: LLaMA 7B and 65B, Alpaca 7B and 13B, and Llama2 13B and 70B. Finally, we consider two closed source models (`davinci-003`, `gpt-3.5-turbo`). Table 1 shows the list of models tested in our experiments.

4.2 In-context Learning

The focus of our experiments is on the effect that demonstrations have on fairness, however other aspects such as model hyperparameters and prompt structure may affect the performance of the model. We controlled for temperature by conducting experiments varying temperature and choose the best (1.0) based on the results in Appendix C. Further, we controlled for prompt variability by utilizing existing prompts for each dataset where available. Otherwise, we adapted prompts from similar tasks. 2 shows the prompt templates. We choose the best prompt structures based on performance from past work, and leave exploration of the fairness effect of prompt structure to future work.

Bias in Bios: We adapted the prompt from Lin et al. (2022) to include information about the labels. **HateXplain**: We adopted the prompt from Kocielnik et al. (2023). **Twitter Sentiment**: Similar to Bias in Bios, we modified the prompt from Min et al. (2022) to include information about the labels. We prepended k samples (shots) from the training set as demonstrations; each demonstration follows the same prompt format. We evaluate models with zero-shot and 10-shot settings; we discontinued 5-shot evaluations after finding no meaningful differences in the results.

We note that it may be unrealistic to assume a large training set from which to draw demonstrations while also claiming a few-shot setting (Perez et al., 2021). If we indeed have hundreds or thousands of examples, train a model! Nevertheless, we evaluate in this setting to better understand the effects of demonstration selection on fairness. If one was going to annotate a small number of examples to include in a prompt, which type of examples should be included to maximize fairness? To answer this question, we rely on existing annotations (training sets) rather than creating our own.

4.3 Demonstration Selection Strategies

We evaluate existing demonstration selection methods for fairness: semantic similarity (Liu et al., 2022; Gao et al., 2021b) and diversity (Zhang et al., 2022b). We also experiment with demographic-aware selection methods: sampling only *within* the same demographic group and using a *representative* sample.

Zero-shot. We contextualize the performance and fairness of shot selection methods by including zero-shot baselines, i.e. no added demonstrations.

Random. We evaluate randomly selecting 10 demonstrations. While this may not be optimal for performance (Liu et al., 2022), the fairness of this method is unknown.

Similarity. Demonstrations are selected based on the query instance. We select the $k = 10$ most similar training examples as compared to the query instance. Similarity is measured based on the cosine distance of the SBERT (Reimers and Gurevych, 2019) embeddings, following (Gao et al., 2021b).[3]

Diversity. A single set of demonstrations is selected to include across all test instances to reflect a diversity of examples. Like Similarity selection, we obtain SBERT sentence embeddings and then use KMeans Clustering from the faiss library (Johnson et al., 2019) to produce $k = 10$ clusters. We selected the demonstrations with the vector closest to the centroid of each cluster (Zhang et al., 2022b), in order to obtain samples that are semantically diverse.

Within. We randomly select demonstrations that have the same demographic attribute as the test instance. For example, in Bias in Bios, if the example is a biography of a woman, we randomly select biography demonstrations only from women.

Representative. A single set of demonstrations is selected to include across all test instances to reflect a demographically representative set of instances. For example, in Bias in Bios, we randomly

[3]We use the `all-mpnet-base-v2` model which is the highest-performing sentence-embedding model at the time of writing.

Dataset	Prompt Structure
Bias in Bios	<Bio> \n Occupations: <List of Occupations> \nThe occupation of this person is <label>
Twitter Sent.	Post:<Tweet>\nQuestion: Is this post happy or sad? \nAnswer: <label>
HateXplain	Post:<Tweet> \nQuestion: Does this post contain offensive language?\n Answer: <label>

Table 2: Prompt templates used in our experiments. For each example, $k = \{0, 10\}$ demonstrations are constructed using the templates and prepended to the example which follows the same template but without the <label>.

sample 5 biography demonstrations from women and 5 from men, obtaining a representative sample.

In addition to the demonstration selection methods, we experiment with appending the demographic category, e.g. race, sex, etc. (`demographic-attribute` prompting), to the prompt in each demonstration and the test example. This is inspired by prior work that showed increased performance with demographically aware models (Hovy, 2015).

4.4 Evaluation

We obtain predictions by allowing each model to generate up to five tokens. Positive and negative labels are obtained by substring matching of the generated tokens. Specifically, for Bias in Bios models, we allowed the term "lawyer" as correct for "attorney". For performance, we report the macro-averaged F1 score of the model.

For the fairness evaluation, we use a modified 1-GAP metric originally introduced by De-Arteaga et al. (2019). GAP is the difference in recall scores (TPR) between two demographic groups, also called *equalized opportunity* (Hardt et al., 2016). We modified the definition to support multiple demographic groups by selecting the biggest recall difference across demographic groups, inspired by Ghosh et al. (2021). We define the set of all demographics as S, Y as the gold label, and $\hat{Y}$ as the prediction.

$$TPR_{s_i,y} = P\left(\hat{Y} = y \mid S = s_i, Y = y\right)$$

$$1 - GAP = \min_{s_i,s_j \in S} 1 - (TPR_{s_i,y} - TPR_{s_j,y})$$

1-GAP gives us a relative metric, where models closest to 1 are the fairest. However, to obtain a binary label for whether a model is fair, we obtain distributions of recall scores for each demographic by bootstrapping with 100 iterations. We then perform a Krukal-Wallis (KW) one-way analysis of variance to test whether the recall score samples for each demographic belong to the same distribution (fair model.)

4.5 Supervised and Other Baselines

To contextualize the performance of the LLMs for these tasks, we compare the in-context models with a *random classifier* baseline and BERT-based fine-tuned classification models with and without a fairness loss following Foulds et al. (2020). The BERT-based classifiers are encoder+classification layer models that were end-to-end finetuned with the training data and hyperparameter tuned with the available dev sets. The *fairness* variants of BERT-based classifiers are finetuned with a true positive rate (TPR or recall-parity) using the demographics available per dataset (Foulds et al., 2020). We use BERT-style encoders (Devlin et al., 2019b) with vocabulary that match the dataset domain: RoBERTa for the Bias in Bios dataset (Liu et al., 2019b) initialized with the `roberta-base` checkpoint,[4] and BERTweet for HateXplain and Twitter Sentiment (Nguyen et al., 2020), initialized with the `vinai/bertweet-base` checkpoint.[5] For more model training details, the hyperparameter search space, and details about fairness definitions and fairness finetuning, see Appendix B.

5 Results & Analysis

Table 3 shows the results of the models on all three datasets using the different demonstration selection methods. While the best performing LLMs are competitive compared to the supervised baselines, some settings perform below the random classifier baseline, as seen in table 3 (UL2, LLaMA-13B&65B, Alpaca-7B&13B, and Llama2-13B&70B).

For demographic fairness, we observe that the most fair models are often below random performance. Since the ultimate goal of fairness is to maximize the utility of the models across all demographic groups (rather than none), we do not take into account fairness results from models that perform below a random classifier, these are shaded on table 3. Comparing in-context models with

[4] https://huggingface.co/roberta-base
[5] https://huggingface.co/vinai/bertweet-base

BERT-based finetuned models, in-context models tend to be fairer but with a substantial loss in performance, with the most fair in-context model (zeroshot Llama2-70B-chat) performing ≈ 25 F1 points lower than the fair BERT-based counterpart. This is an extreme example of the fairness and accuracy trade-off, that is present in some of the LLMs we tested; fair models are fair because they perform poorly for all groups.

5.1 Model Choice

When considering the overall performance of models across all our settings, it becomes clear that the choice of model matters both in terms of performance and fairness. Flan-UL2, `davinci-003`, `gpt-3.5-turbo` and Llama2-13B-chat are the best-performing models across the three datasets. Some models, e.g. Alpaca and UL2, have better than random performance in only one dataset. In contrast, there is not a clear winner for fairness, with model fairness varying across all datasets. However, the more drastic fairness differences are at the dataset level, where the fairness of all models in Twitter Sentiment ($>$ .9 for all models) is much greater than, e.g. HateXplain. These dataset-specific differences could be due to overfitting to widely used benchmarks, as the Twitter Sentiment task is more often included benchmarks used to evaluate LLMs compared to HateXplain. When comparing finetuned vs pretrained variants of LLMs (FLAN-UL2 vs. UL2, LLaMA2 vs. LLama2-chat), finetuning seems to help in performance but have a varied effect on fairness.

Overall, we find that model selection for fairness cannot be generalized across datasets.

5.2 Performance and Fairness

1-GAP (fairness) has an inherent connection with F1 (performance) since both include recall. However, we can still have fair models at different ranges of accuracy. Many have postulated that there is a trade-off between fairness and performance; fairness comes at the expense of performance resulting in a negative correlation. Much recently, Islam et al. (2021) showed this trade-off is not always present empirically; some methods obtain high performance and fairness.

Our experiments (perhaps distressingly) exhibit both positive and negative correlations for certain models across datasets. Figure 1 shows the 1-GAP vs F1 plots for three models, which have a positive (Flan-UL2), no (Alpaca-7B) and negative correlation (UL2) between performance and fairness. This erratic relationship underscores the need for explicit evaluation of fairness rather than relying on performance alone.

5.3 Zero-shot Settings are Sometimes Better

How important is adding demonstrations (few-shot) to prompts compared to leaving them out (zero-shot) for fairness? The effect is especially pronounced for UL2, LLaMA, and Alpaca, e.g. Alpaca-7B goes from unusable performance in zero-shot (2.3 F1) to decent in few-shot (82.1 F1) in Bias in Bios. On the other hand, higher performing models (`davinci-003`, `gpt-3.5-turbo` and Flan-UL2) sometimes do better in the zero-shot setting; adding demonstrations hurts performance. Nevertheless, on average across models, zero-shot settings were always outperformed by all demonstration selection methods (see Table 4).

The relationship between demonstrations and fairness is more varied. In general, when both fairness and performance in zeroshot settings are high, adding demonstrations does not help and can even harm fairness. However, in average across models, zeroshot settings are generally more fair than other demonstration selection methods closely followed by *similarity*. While adding demonstrations helps performance, the effect on fairness is unpredictable. This again underscores the importance of evaluating prediction fairness of LLMs.

5.4 Which Demonstrations To Add

Adding demonstrations (Random vs. Zero-shot) usually improves model performance (∼70% of the time), but often made model fairness worse (∼60% of the time was worse). Care in demonstration selection is needed to ensure fairness.

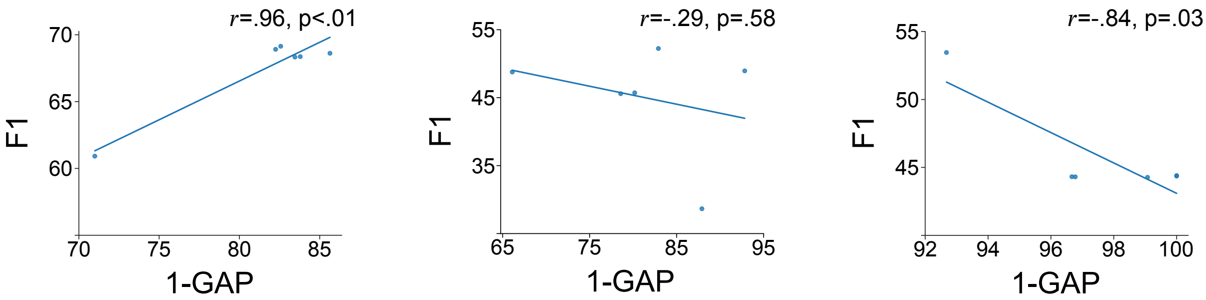
For *similarity* and *diversity* selection methods: *similarity* selection helps performance on average across datasets compared to random selection and zero-shot (table 4.) This same is generally true for fairness, but still less fair than zeroshot. In contrast, *Diversity* selection has less consistent behavior, where it helps LLaMA-65B and Flan-UL2, but hurts every other model. The fairness scores also fluctuate and vary by data and model.

The demographic-based demonstration selection strategies are less successful overall. Perhaps surprisingly, selecting demonstrations from *within* the same demographic was the least favored setting

[5] The recall scores from bootstrap samples (100) across demographics belong to the same distribution.

	zeroshot		random		similarity		diversity		within		representative	
HateXplain race												
	F1	1-GAP	F1	1-GAP	F1	1-GAP	F1	1-GAP	F1	1-GAP	F1	1-GAP
`davinci-003`	64.1	**84.7**	**70.0**	74.0	68.0	78.0	66.8	69.6	65.8	82.6	69.0	79.5
`gpt-3.5-turbo`	61.3	**85.6**	**69.1**	80.5	67.8	73.8	67.0	80.8	67.3	82.1	67.8	78.6
UL2	**53.5**	**92.7**	44.3	99.1	44.3	96.7	44.4	100.0*	44.4	100.0*	44.3	96.8
FLAN-UL2	60.9	71.0	68.4	83.8	68.6	**85.6**	68.3	83.5	68.9	82.3	**69.1**	82.6
LLaMA-13B	22.3	77.5	31.3	69.1	**48.5**	**52.6**	23.5	75.7	36.0	48.7	32.0	78.2
LLaMA-65B	40.5	84.6	44.7	76.4	**52.2**	**79.6**	49.6	60.7	47.2	71.3	48.8	68.7
Alpaca-7B	28.7	87.9	48.8	66.1	**52.2**	82.9	45.6	78.6	45.7	80.2	48.9	**92.8**
Alpaca-13B	27.7	85.7	34.9	84.8	38.3	78.5	37.1	74.7	35.5	76.9	36.6	77.1
LLaMA2-13B	33.0	86.5	46.1	**94.6**	47.1	85.2	**47.1**	93.5	46.0	88.7	43.9	92.6
LLaMA2-13B-chat	**63.4**	**93.5**	59.9	71.1	63.0	65.2	59.3	49.2	58.9	93.3	61.6	81.5
LLaMA2-70B	**46.1**	**90.9**	25.5	78.7	33.3	77.2	15.1	79.6	28.2	81.8	33.5	80.4
LLaMA2-70B-chat	48.5	**99.1**	**51.9**	68.2	42.4	74.6	31.7	82.2	46.4	72.0	51.1	77.2
avg	45.8	86.6	49.6	78.9	52.1	77.5	46.3	77.3	49.2	80.0	50.6	82.2
random class.	45.2											
BERTweet	72.7	40.0										
BERTweet Fair	73.2	86.9										
Bias in Bios												
	F1	1-GAP	F1	1-GAP	F1	1-GAP	F1	1-GAP	F1	1-GAP	F1	1-GAP
`davinci-003`	**82.8**	79.2	80.0	77.8	81.9	**85.6**	76.4	78.6	79.6	82.4	79.6	81.6
`gpt-3.5-turbo`	84.6	87.4	84.6	88.8	**86.7**	**92.4**	81.8	89.4	84.4	90.4	84.4	88.2
UL2	**19.2**	**99.6**	2.5	100.0*	11.5	100.0*	0.9	100.0*	2.4	100.0*	2.4	100.0*
FLAN-UL2	**86.7**	**92.8**	84.2	84.6	85.3	87.4	85.4	83.0	84.5	85.0	84.5	84.4
LLaMA-13B	11.5	99.8	74.2	82.0	**78.7**	**95.6**	78.3	83.0	73.0	78.4	73.6	81.8
LLaMA-65B	8.0	99.4	73.7	86.0	74.1	83.6	**82.1**	84.6	73.2	85.2	74.7	**88.4**
Alpaca-7B	2.3	99.8	76.7	78.2	**82.1**	79.8	80.6	**83.4**	76.3	78.4	76.1	79.6
Alpaca-13B	29.0	96.0	18.2	**99.2**	**34.0**	95.0	1.7	100.0*	18.4	98.4	17.7	98.4
LLaMA2-13B	2.1	100.0*	76.0	83.4	75.5	87.4	**83.6**	83.6	75.8	88.2	77.0	**91.8**
LLaMA2-13B-chat	65.0	**98.4**	84.7	93.2	**86.9**	88.2	83.7	94.2	85.1	95.6	84.9	95.4
LLaMA2-70B	5.2	99.6	**63.4**	91.0	50.0	94.4	54.7	**98.2**	62.9	94.4	43.7	95.8
LLaMA2-70B-chat	69.3	85.4	73.9	**94.6**	1.0	100.0*	**83.9**	82.4	73.5	93.8	73.6	89.2
avg	38.8	94.8	66.0	88.2	62.3	90.8	66.1	88.4	65.8	89.2	64.4	89.6
random class.	45.2											
RoBERTa	79.6	91.2										
RoBERTa Fair	77.5	92.0										
Twitter Sentiment												
	F1	1-GAP	F1	1-GAP	F1	1-GAP	F1	1-GAP	F1	1-GAP	F1	1-GAP
`davinci-003`	60.4	97.5	69.3	93.9	**71.1**	**99.5**	69.9	86.1	69.6	96.9	69.6	93.6
`gpt-3.5-turbo`	44.8	97.6	54.5	99.2	**61.2**	99.7*	57.0	**99.9***	54.7	98.2	54.9	97.7
UL2	58.1	98.6	48.2	92.6	**65.0**	**99.9***	33.5	100.0	47.8	83.6	47.9	94.1
FLAN-UL2	69.5	99.6*	69.7	99.1	**70.0**	**99.9***	69.6	98.8	69.8	98.8	69.8	98.6
LLaMA-13B	36.9	97.8	55.8	**97.0**	**64.5**	98.9	51.6	97.8	56.0	93.5	54.8	95.6
LLaMA-65B	0.4	99.8	54.7	**96.4**	**61.2**	93.6	49.9	93.4	54.6	92.5	54.3	94.5
Alpaca-7B	35.9	92.0	2.2	100.0*	10.2	98.9	0.0	100.0*	2.5	99.5	2.1	99.9
Alpaca-13B	21.9	97.2	35.7	98.8	36.5	99.4	24.6	97.4	35.6	95.4	36.7	98.0
LLaMA2-13B	8.3	96.0	20.2	95.2	52.1	96.5	**53.6**	**98.8**	21.8	87.2	21.0	96.0
LLaMA2-13B-chat	62.7	92.1	60.9	97.3	**63.2**	95.3	62.2	97.2	62.3	95.7	61.5	**97.8**
LLaMA2-70B	16.6	99.8	0.4	99.8	11.5	99.6	3.6	99.5	0.6	99.8	0.4	99.8
LLaMA2-70B-chat	**59.3**	**91.9**	43.2	96.0	44.6	91.1	51.5	91.6	43.5	93.9	42.7	95.7
avg	39.5	96.6	42.9	97.1	50.9	97.7	43.9	96.7	43.2	94.6	43.0	96.8
random class.	50.0											
BERTweet	76.6	83.9										
BERTweet Fair	76.5	88.7										

Table 3: Macro-averaged F1 score and 1-GAP of all models and demonstration selection methods for all of the three datasets. **Bold** is best per model×dataset and underlined is best per dataset (above a random baseline). Asterisk (*) denotes no significant difference in recall scores performing a Kruskal-Wallis test with 100 bootstrap iterations. We shade results that have an F1 score below a random baseline.



(a) Flan-UL2 (b) Alpaca-7B (c) UL2

Figure 1: F1 vs 1-GAP when varying demonstration selection methods for Flan-UL2, Alpaca-7B and UL2 in HateXplain dataset showing positive, no correlation and negative correlations respectively.

	HateXplain		Bias in Bios		Twitter Sent.	
	F1	1-GAP	F1	1-GAP	F1	1-GAP
zeroshot	45.8	**86.6**	38.8	**94.8**	39.6	96.6
random	49.6	78.9	66.0	88.2	42.9	97.1
similarity	**52.1**	77.5	62.3	90.8	**50.9**	**97.7**
diversity	46.3	77.3	**66.1**	88.4	43.9	96.7
within	49.2	80.0	65.8	89.2	43.2	94.6
representative	50.6	82.2	64.4	89.6	43.0	96.8

Table 4: Mean F1 & 1-GAP per selection strategy.

in both performance and fairness across models and datasets. We expected choosing data of the same type would help fairness; it did not. A *representative* selection of demonstrations had more success than *within* in both performance and fairness. These results are congruent with prior work that found that a representative selection of demonstrations aids in reducing bias in models (Si et al., 2022).

Combining these findings, our results suggest that LLMs more efficiently utilize examples with semantic similarity (*similarity*) as opposed examples with similarities in text due to demographic groups (*within*.)

5.5 Including Demographic Attributes

Perhaps having access to explicit demographic information can help LLMs reduce classification bias. Figure 2 shows the results of including demographic attributes with the demonstrations to open source models in the Bias in Bios dataset (all datasets are shown in Table 5). While adding demographic attributes helps in terms of performance, benefits appear to be model specific. For LLaMA and Alpaca, some settings have improved performance, but overall a mixed effect on fairness, e.g. for Alpaca-13B with demonstrations selected with *diversity* the performance increased from 2 F1 to 80 by simply adding the demographic attributes but, at the same time, reduced from perfect fairness (100) to 81 (Figure 2.) Adding demographic attributes affected the performance and fairness of Flan-UL2 models to a lesser effect. For these models, there was a general trade-off between increasing performance but decreasing fairness, and vice-versa.

Overall, adding demographic attributes seems to help LLaMA and Alpaca models the most in performance, perhaps because more information is provided, but the effect on fairness is mixed.

5.6 Other Selection Methods

Since *similarity* and *diversity* selection were more successful than demographic-based selection, we experimented with combining these and the *within* method. We test *within+similarity*, demonstrations that are most similar within the same demographic group, and *within+diversity*, demonstrations that are most diverse within the same demographic.

Figure 3 show results for Bias in Bios and Table 6 for all datasets. Unfortunately, combining *within* and *similarity* methods often drastically **decreases** model performance, but sometimes increases fairness (Flan-UL2.) This is interesting as these are the most similar methods, with $\sim 80\%$ of demonstrations selected by *similarity* being within the same demographic. Despite these similarities, we see that semantic *similarity* is generally more important than demographic similarity for both performance and fairness, and combining these two actually hinders the performance of the models.

On the other hand, combining *within* and *diversity* selection methods often helps in both performance and fairness! Contextualizing these results with the previous subsections, a rule-of-thumb is to select semantically diverse demonstrations within the same demographic group, or semantically similar demonstrations across all demographics.

While semantic similarity was not always the best performing, it provides the best performance and fairness trade-off.

6 Conclusion

Significant work has gone into evaluating different demonstration selection strategies in the performance of LLMs as classification systems. This paper represents one of the first studies to consider the fairness of these systems. Our study considers 7 widely used family of models (Table 1), three datasets, and multiple demonstration selection methods. We find that the classification fairness of LLMs doest not generalize across datasets, similar to prior work with other families supervised models (Zhang et al., 2020). Our results support the need for task-specific fairness evaluations and serve as a cautionary tale for over-reliance on performance metrics alone. On average, LLMs still underperform compared to supervised baselines often with a more drastic fairness vs performance trade-off. In terms of shot selection strategies, while adding demonstrations (with *similarity* having the most success) generally yields higher performing models (compared to *zero-shot*), it does not consistently yield fairer models.

Where do these results leave us? First, fairness **must** be evaluated alongside task performance when developing prompts, selection strategies, and models. We cannot assume any relationship between fairness and performance. Second, we need to better understand **why** LLMs are unfair in their predictions. While significant work has examined fairness in supervised training objectives (Delobelle et al., 2021), and other work demonstrates bias in LLM generations (Chang and Bergen, 2023), we need work that intersects these two. Third, how can we determine **when** a LLM is being unfair? Work examining confidence in LLM predictions (e.g., Portillo Wightman et al., 2023) can help automatically determine the accuracy of the system. Can we develop similar metrics for fairness? This would be especially helpful in cases where we do not have demographically labeled data. Finally, there is now a large focus on fine-tuning LLMs (e.g. RLHF (Ouyang et al., 2022), FLAN (Chung et al., 2022)). The goal of these methods has been better instruction following and improved accuracy on prediction tasks, but our results suggest they do not always make models fairer. How can we include fairness objectives in this training process?

7 Ethics Statement

We study the fairness of language models for three tasks: occupation classification, sentiment analysis, and hate speech detection. Occupation classification has direct applications in the automation of hiring procedures, which have been historically biased along many more demographic attributes than what we consider, e.g. age, disabilities, race, ethnicity, sexual orientation, and veteran status. The same is true of the other datasets in this paper. Additionally, often these inequities intersect across these social groups, further increasing the impact of applications that use these models outside of an academic environment. Because we were limited by the currently available datasets and the coverage they have on demographic attributes, we acknowledge that fairness as is discussed in this paper will not translate to social fairness in the wild without first considering all of these biases.

8 Limitations

We work with LLMs that are expensive to run (large GPUs to run big open source models) or costly to access (cost of APIs). This limits our ability to fully explore all possible models. For example, OpenAI API costs precluded our use of close-source models in some experiments Sections 5.5 and 5.6. Furthermore, our closed-source model evaluations may not be reproducible as we do not have control over updates to the underlying models and the model outputs are known to be inconsistent (Ye et al., 2023).

While we consider 12 models, there are now many different LLMs available for evaluation, with several released concurrent with this study, e.g. GPT4o, Falcon (Almazrouei et al., 2023) and Vicuna (Chiang et al., 2023). We cannot evaluate all models, but our results suggest that the fairness of these models will also be highly varied and there is no reason to believe this invalidates our findings. Additionally, other aspects of in-context learning may also affect the fairness of LLMs that we did not study, e.g. demonstration ordering (Lu et al., 2022) and prompt formatting (Wang et al., 2022). Further, we only test these models in English datasets limiting the breath of the type of biases we can capture; future work can expand this evaluation to other languages.

Acknowledgments

The authors gratefully acknowledge Elizabeth Salesky and Sharon Levy for helpful feedback, and we thank the anonymous reviewers for their helpful comments.

References

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Merouane Debbah, Etienne Goffinet, Daniel Heslow, Julien Launay, Quentin Malartic, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. Falcon-40B: an open large language model with state-of-the-art performance.

Ioana Baldini, Dennis Wei, Karthikeyan Natesan Ramamurthy, Moninder Singh, and Mikhail Yurochkin. 2022. Your fairness may vary: Pretrained language model fairness in toxic text classification. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2245–2262, Dublin, Ireland. Association for Computational Linguistics.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2019. *Fairness and Machine Learning: Limitations and Opportunities*. fairmlbook.org. `http://www.fairmlbook.org`.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic dialectal variation in social media: A case study of African-American English. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1119–1130, Austin, Texas. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *ArXiv*, abs/2005.14165.

Tyler A Chang and Benjamin K Bergen. 2023. Language model behavior: A comprehensive survey. *arXiv preprint arXiv:2303.11504*.

Yanda Chen, Ruiqi Zhong, Sheng Zha, George Karypis, and He He. 2021. Meta-learning via language model in-context tuning. *ArXiv*, abs/2110.07814.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 120–128, New York, NY, USA. Association for Computing Machinery.

Daniel de Vassimon Manela, David Errington, Thomas Fisher, Boris van Breugel, and Pasquale Minervini. 2021. Stereotype and skew: Quantifying gender bias in pre-trained and fine-tuned language models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2232–2242, Online. Association for Computational Linguistics.

Pieter Delobelle, Ewoenam Kwaku Tokpo, Toon Calders, and Bettina Berendt. 2021. Measuring fairness with biased rulers: A survey on quantifying biases in pretrained language models. *arXiv preprint arXiv:2112.07447*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019a. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019b. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1286–1305, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey for in-context learning. *arXiv preprint arXiv:2301.00234*.

James R Foulds, Rashidul Islam, Kamrun Naher Keya, and Shimei Pan. 2020. An intersectional definition of fairness. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, pages 1918–1921. IEEE.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021a. Making pre-trained language models better few-shot learners. *ArXiv*, abs/2012.15723.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021b. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3816–3830, Online. Association for Computational Linguistics.

Avijit Ghosh, Lea Genuit, and Mary Reagan. 2021. Characterizing intersectional group fairness with worst-case comparisons. In *Proceedings of 2nd Workshop on Diversity in Artificial Intelligence (AIDBEI)*, volume 142 of *Proceedings of Machine Learning Research*, pages 22–34. PMLR.

Sophie Groenwold, Lily Ou, Aesha Parekh, Samhita Honnavalli, Sharon Levy, Diba Mirza, and William Yang Wang. 2020. Investigating African-American Vernacular English in transformer-based text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5877–5883, Online. Association for Computational Linguistics.

Xudong Han, Aili Shen, Yitong Li, Lea Frermann, Timothy Baldwin, and Trevor Cohn. 2022. FairLib: A unified framework for assessing and improving fairness. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 60–71, Abu Dhabi, UAE. Association for Computational Linguistics.

Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29.

Ari Holtzman, Peter West, Vered Schwartz, Yejin Choi, and Luke Zettlemoyer. 2021. Surface form competition: Why the highest probability answer isn't always right. *ArXiv*, abs/2104.08315.

Dirk Hovy. 2015. Demographic factors improve classification performance. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 752–762, Beijing, China. Association for Computational Linguistics.

Rashidul Islam, Shimei Pan, and James R Foulds. 2021. Can we obtain fairness for free? In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 586–596.

Zhengbao Jiang, Frank F. Xu, J. Araki, and Graham Neubig. 2019. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Masahiro Kaneko, Danushka Bollegala, and Naoaki Okazaki. 2022. Debiasing isn't enough! – on the effectiveness of debiasing MLMs and their social biases in downstream tasks. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1299–1310, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Pauline T Kim. 2022. Race-aware algorithms: Fairness, nondiscrimination and affirmative action. *Cal. L. Rev.*, 110:1539.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Rafal Kocielnik, Sara Kangaslahti, Shrimai Prabhumoye, Meena Hari, Michael Alvarez, and Anima Anandkumar. 2023. Can you label less by using out-of-domain data? active and transfer learning with few-shot instructions. In *Proceedings of The 1st Transfer Learning for Natural Language Processing Workshop*, volume 203 of *Proceedings of Machine Learning Research*, pages 22–32. PMLR.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Annual Meeting of the Association for Computational Linguistics*.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2021. What makes good in-context examples for gpt-3? In *Workshop on Knowledge Extraction and Integration for Deep Learning Architectures; Deep Learning Inside Out*.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019a. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2021. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Annual Meeting of the Association for Computational Linguistics*.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland. Association for Computational Linguistics.

Huan Ma, Changqing Zhang, Yatao Bian, Lemao Liu, Zhirui Zhang, Peilin Zhao, Shu Zhang, Huazhu Fu, Qinghua Hu, and Bingzhe Wu. 2023. Fairness-guided few-shot prompting for large language models. *arXiv preprint arXiv:2303.13217*.

Liam Magee, Lida Ghahremanlou, Karen Soldatic, and Shanthi Robertson. 2021. Intersectional bias in causal language models. *arXiv preprint arXiv:2107.07691*.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14867–14875.

Sewon Min, Michael Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2021a. Noisy channel language model prompting for few-shot text classification. In *Annual Meeting of the Association for Computational Linguistics*.

Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2021b. Metaicl: Learning to learn in context. *ArXiv*, abs/2110.15943.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online. Association for Computational Linguistics.

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

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14, Online. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Ethan Perez, Douwe Kiela, and Kyunghyun Cho. 2021. True few-shot learning with language models. *Advances in neural information processing systems*, 34:11054–11070.

Jason Phang, Phil Yeres, Jesse Swanson, Haokun Liu, Ian F. Tenney, Phu Mon Htut, Clara Vania, Alex Wang, and Samuel R. Bowman. 2020. `jiant` 2.0: A software toolkit for research on general-purpose text understanding models. `http://jiant.info/`.

Gwenyth Portillo Wightman, Alexandra DeLucia, and Mark Dredze. 2023. Strength in numbers: Estimating confidence of large language models by prompt agreement. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, Toronto, CA. Association for Computational Linguistics.

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

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on*

Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2021. Learning to retrieve prompts for in-context learning. *ArXiv*, abs/2112.08633.

Leonard Salewski, Stephan Alaniz, Isabel Rio-Torto, Eric Schulz, and Zeynep Akata. 2023. In-context impersonation reveals large language models' strengths and biases. *arXiv preprint arXiv:2305.14930*.

Aili Shen, Xudong Han, Trevor Cohn, Timothy Baldwin, and Lea Frermann. 2022. Optimising equal opportunity fairness in model training. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4073–4084, Seattle, United States. Association for Computational Linguistics.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3407–3412, Hong Kong, China. Association for Computational Linguistics.

Chenglei Si, Zhe Gan, Zhengyuan Yang, Shuohang Wang, Jianfeng Wang, Jordan Boyd-Graber, and Lijuan Wang. 2022. Prompting gpt-3 to be reliable. *arXiv preprint arXiv:2210.09150*.

Karolina Stanczak and Isabelle Augenstein. 2021. A survey on gender bias in natural language processing. *arXiv preprint arXiv:2112.14168*.

Yarden Tal, Inbal Magar, and Roy Schwartz. 2022. Fewer errors, but more stereotypes? the effect of model size on gender bias. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 112–120, Seattle, Washington. Association for Computational Linguistics.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Yi Tay, Mostafa Dehghani, Vinh Q. Tran, Xavier Garcia, Jason Wei, Xuezhi Wang, Hyung Won Chung, Dara Bahri, Tal Schuster, Steven Zheng, Denny Zhou, Neil Houlsby, and Donald Metzler. 2023. UL2: Unifying language learning paradigms. In *The Eleventh International Conference on Learning Representations*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. Self-instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *ArXiv*, abs/2201.11903.

Qinyuan Ye, Bill Yuchen Lin, and Xiang Ren. 2021. CrossFit: A few-shot learning challenge for cross-task generalization in NLP. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7163–7189, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wentao Ye, Mingfeng Ou, Tianyi Li, Xuetao Ma, Yifan Yanggong, Sai Wu, Jie Fu, Gang Chen, Junbo Zhao, et al. 2023. Assessing hidden risks of llms: An empirical study on robustness, consistency, and credibility. *arXiv preprint arXiv:2305.10235*.

Guanhua Zhang, Yihua Zhang, Yang Zhang, Wenqi Fan, Qing Li, Sijia Liu, and Shiyu Chang. 2022a. Fairness reprogramming. *arXiv preprint arXiv:2209.10222*.

Haoran Zhang, Amy X Lu, Mohamed Abdalla, Matthew McDermott, and Marzyeh Ghassemi. 2020. Hurtful words: quantifying biases in clinical contextual word embeddings. In *proceedings of the ACM Conference on Health, Inference, and Learning*, pages 110–120.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2022b. Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493*.

Tony Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. *ArXiv*, abs/2102.09690.

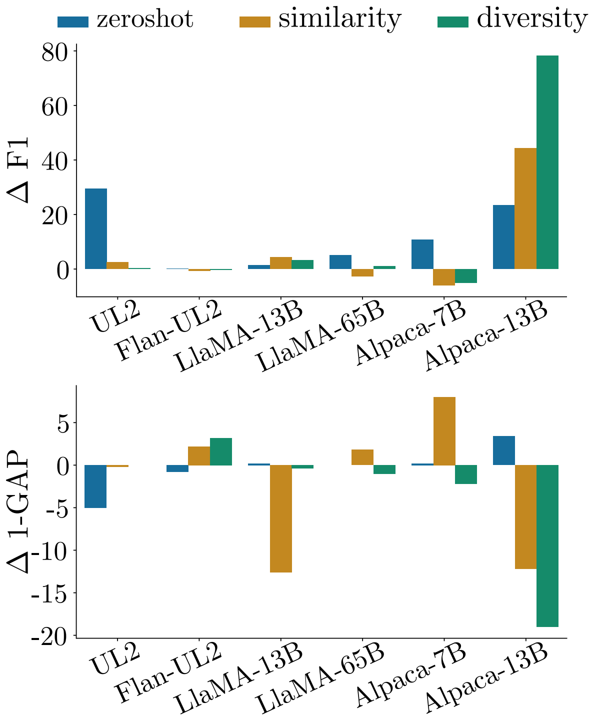


Figure 2: Δ F1 and Δ 1-GAP when including demographic attributes in prompt (Bias in Bios.)

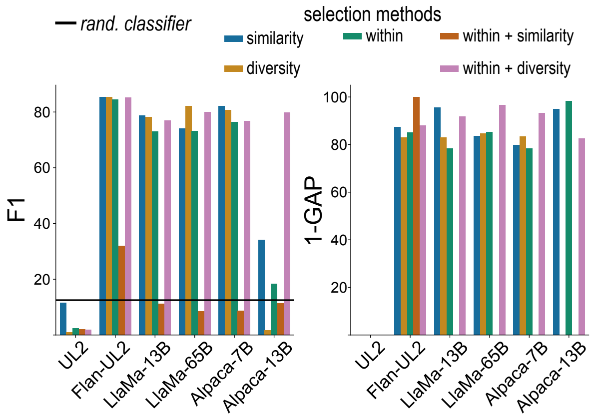


Figure 3: Performance (F1) and fairness (1-GAP) of combining *within* with semantic-based methods across models in the Bias in Bios dataset. For 1-GAP graph we show models with > rand. classifier performance.

A All Results

Here we present the performance of the models adding demographic attributes to the demonstrations and prompt in Table 5. And finally, we show the performance and fairness of the models when combining semantic and demographic based selection methods in Table 6, Figure 2 and Figure 3.

B BERT-based fine-tuning details

Baseline. We use BERT-style encoders (Devlin et al., 2019b) with a vocabulary that matches the domain of each dataset: RoBERTa for the Bias in Bios dataset (Liu et al., 2019b) initialized with the `roberta-base` checkpoint,[6] and BERTweet for HateXplain and Twitter Sentiment (Nguyen et al., 2020), initialized with the `vinai/bertweet-base` checkpoint.[7] We add a separate linear classification head for each task, with a Softmax output function to allow for multi-class classification (Bias in Bios) or a Sigmoid output function for binary classification (HateXplain and Twitter Sentiment.) The document representation for the classification head is a mean-pooled aggregation across all subword representations of the document taken at the top layer of the network.. Models were trained on Nvidia A100 GPUs, using `jiant` (Phang et al., 2020), a multi-task wrapper library.

Fairness Finetuning. In addition to a typical finetuning model, we also provide a finetuned model with an added fairness loss, to compare with a model that adds fairness to the objective. We utilize equalized opportunity, also known as GAP, as our fairness definition, which is the compliment of 1-GAP, the fairness definition in the main paper. We use ϵ-Differential Equalized Opportunity (ϵ-DEO), a variant of ϵ-DF (Foulds et al., 2020), that applies the equalized opportunity objective, to ensure that the recall rates are equal across demographic groups (Barocas et al., 2019) and that is learnable and differentiable.

Formally, let $s_1, ..., s_p$ be discrete-valued demographic attributes, $z = s_1 \times s_2 \times ... \times s_p$. A model $M(X)$ satisfies ϵ-DEO with respect to z if for all $x, \hat{y} \in \mathrm{Range}(M)$ and $y \in \mathrm{Range}(M)$,

$$e^{-\epsilon} \leq \frac{Pr(M_\theta(x) = 1 | s_i, y = 1)}{Pr(M_\theta(x) = 1 | s_j, y = 1)} \leq e^{\epsilon}, \quad (1)$$

for all $(s_i, s_j) \in z \times z$ where $Pr(s_i) > 0$, $Pr(s_j) > 0$; smaller ϵ is better, with $\epsilon = 0$ for perfect fairness. Perfect fairness results from a classifier with the same recall rates across groups of demographic attributes.

The standard approach to incorporating fairness metrics into learning objectives uses an additive term. For example, for a deep neural network classifier $M(X)$ with parameters θ, we obtain the following,

[6] https://huggingface.co/roberta-base
[7] https://huggingface.co/vinai/bertweet-base

		HateXplain race				Bias in Bios				Twitter Sentiment			
		F1 (Δ)		1-GAP (Δ)		F1 (Δ)		1-GAP (Δ)		F1 (Δ)		1-GAP (Δ)	
baseline	*random class.*	61.3				12.5				50.0			
model	selection method												
UL2	zero-shot	**53.5**		92.7		19.2		99.6		58.1		98.6	
	+demographic attributes	45.9	(-7.6)	**100**	(7.3)	**48.7**	(29.5)	94.6	(-5.0)	61.1	(3.0)	78.8	(-19.8)
	random	44.3		99.1		2.5		**100**		48.2		92.6	
	+demographic attributes	44.3	(0.0)	99.7	(0.6)	2.3	(-0.2)	**100**	(0.0)	42.3	(-6.0)	99.2	(6.6)
	similarity	44.3		96.7		11.5		**100**		65.0		99.9	
	+demographic attributes	45.9	(1.5)	**100**	(3.3)	0.140	(2.5)	99.8	(-0.2)	**65.2**	(0.1)	0.924	(-7.5)
	diversity	44.4		**100**		0.9		**100**		33.5		**100**	
	+demographic attributes	44.4	(0.0)	**100**	(0.0)	1.3	(0.3)	**100**	(0.0)	33.4	(-0.1)	0.999	(-0.1)
	within	44.4		**100**		2.4		**100**		47.8		83.6	
	+demographic attributes	44.4	(0.0)	**100**	(0.0)	2.2	(-0.2)	**100**	(0.0)	48.9	(1.0)	0.791	(-4.5)
	representative	44.3		96.8		2.4		**100**		47.9		94.1	
	+demographic attributes	44.4	(0.1)	**100**	(3.2)	3.1	(0.7)	**100**	(0.0)	41.4	(-6.4)	0.936	(-0.5)
Flan-UL2	zero-shot	60.9		71.0		86.7		**92.8**		69.5		99.6	
	+demographic attributes	49.7	(-11.2)	82.2	(11.2)	**86.7**	(0.1)	92.0	(-0.8)	69.4	(-0.1)	98.7	(-0.9)
	random	68.4		83.8		84.2		84.6		69.7		99.1	
	+demographic attributes	65.9	(-2.5)	88.8	(5.0)	82.8	(-1.4)	81.0	(-3.6)	69.3	(-0.4)	98.8	(-0.3)
	similarity	68.6		85.6		85.3		87.4		70.0		**99.9**	
	+demographic attributes	64.9	(-3.7)	88.5	(2.9)	84.6	(-0.7)	89.6	(2.2)	**70.2**	(0.2)	99.1	(-0.8)
	diversity	68.3		83.5		85.4		83.0		69.6		98.8	
	+demographic attributes	67.6	(-0.8)	88.4	(5.0)	85.1	(-0.3)	86.2	(3.2)	70.2	(0.6)	97.4	(-1.4)
	within	68.9		82.3		84.5		85.0		69.8		98.8	
	+demographic attributes	67.7	(-1.2)	**89.1**	(6.8)	84.8	(0.3)	89.0	(4.0)	69.8	(0.0)	98.6	(-0.2)
	representative	**69.1**		82.6		84.5		84.4		69.8		98.6	
	+demographic attributes	66.3	(-2.8)	88.1	(5.6)	83.6	(-0.9)	80.6	(-3.8)	70.2	(0.3)	96.1	(-2.5)
LLaMA-13B	zero-shot	22.3		77.5		11.5		99.8		36.9		0.978	
	+demographic attributes	5.2	(-17.1)	**91.1**	(13.5)	12.9	(1.4)	**100**	(0.2)	28.6	(-8.3)	98.0	(0.2)
	random	31.3		69.1		74.2		82.0		55.8		0.970	
	+demographic attributes	46.9	(15.6)	68.2	(-0.9)	79.1	(4.9)	81.4	(-0.6)	50.6	(-5.2)	97.3	(0.3)
	similarity	48.5		52.6		78.7		95.6		**64.5**		**0.989**	
	+demographic attributes	**55.6**	(7.1)	42.8	(-9.8)	**83.0**	(4.3)	83.0	(-12.6)	62.1	(-2.4)	95.2	(-3.8)
	diversity	23.5		75.7		78.3		83.0		51.6		0.978	
	+demographic attributes	35.4	(11.8)	51.8	(-23.9)	81.5	(3.2)	82.6	(-0.4)	60.2	(8.6)	95.8	(-2.0)
	within	36.0		48.7		73.0		78.4		56.0		0.935	
	+demographic attributes	44.7	(8.7)	55.4	(6.7)	78.8	(5.8)	78.0	(-0.4)	53.4	(-2.6)	91.4	(-2.1)
	representative	32.0		78.2		73.6		81.8		54.8		0.956	
	+demographic attributes	46.1	(14.1)	66.9	(-11.3)	79.9	(6.3)	77.8	(-4.0)	49.0	(-5.8)	97.1	(1.5)
LLaMA-65B	zero-shot	40.5		**84.6**		8.0		**99.4**		0.4		**99.8**	
	+demographic attributes	41.0	(0.4)	75.8	(-8.8)	13.1	(5.1)	**99.4**	(0.0)	0.7	(0.4)	99.6	(-0.2)
	random	44.7		76.4		73.7		86.0		54.7		96.4	
	+demographic attributes	48.3	(3.5)	53.5	(-23.0)	75.6	(1.9)	84.4	(-1.6)	52.0	(-2.7)	99.6	(3.2)
	similarity	52.2		79.6		74.1		83.6		61.2		93.6	
	+demographic attributes	54.7	(2.5)	71.2	(-8.4)	71.4	(-2.7)	85.4	(1.8)	59.1	(-2.1)	95.1	(1.5)
	diversity	49.6		60.7		82.1		84.6		49.9		93.4	
	+demographic attributes	**63.7**	(14.1)	34.4	(-26.3)	**83.1**	(1.0)	83.6	(-1.0)	**62.0**	(12.2)	96.8	(3.4)
	within	47.2		71.3		73.2		85.2		54.6		92.5	
	+demographic attributes	47.5	(0.3)	59.1	(-12.2)	73.1	(-0.1)	81.8	(-3.4)	50.3	(-4.3)	93.0	(0.4)
	representative	48.8		68.7		74.7		88.4		54.3		94.5	
	+demographic attributes	50.4	(1.6)	57.6	(-11.1)	75.8	(1.0)	82.6	(-5.8)	50.0	(-4.4)	89.6	(-4.9)
Alpaca-7B	zero-shot	28.7		87.9		2.3		99.8		35.9		92.0	
	+demographic attributes	45.6	(16.9)	87.2	(-0.7)	13.1	(10.8)	**100**	(0.2)	**57.9**	(22.0)	86.5	(-5.6)
	random	48.8		66.1		76.7		78.2		2.2		**100**	
	+demographic attributes	58.2	(9.4)	46.7	(-19.4)	74.4	(-2.3)	82.4	(4.2)	30.8	(28.6)	94.4	(-5.6)
	similarity	52.2		82.9		**82.1**		79.8		10.2		98.9	
	+demographic attributes	57.9	(5.7)	77.4	(-5.5)	76.2	(-6.0)	87.8	(8.0)	49.6	(39.5)	97.3	(-1.7)
	diversity	45.6		78.6		80.6		83.4		0.0		**100**	
	+demographic attributes	**62.0**	(16.4)	35.7	(-42.9)	0.757	(-5.0)	81.2	(-2.2)	30.5	(30.5)	97.3	(-2.7)
	within	45.7		80.2		76.3		78.4		2.5		99.5	
	+demographic attributes	53.2	(7.5)	79.8	(-0.4)	74.9	(-1.4)	85.0	(6.6)	27.7	(25.2)	97.6	(-2.0)
	representative	48.9		**92.8**		76.1		79.6		2.1		99.9	
	+demographic attributes	58.5	(9.6)	61.7	(-31.1)	72.5	(-3.6)	84.0	(4.4)	34.5	(32.4)	94.4	(-5.5)
Alpaca-13B	zero-shot	27.7		85.7		29.0		96.0		21.9		97.2	
	+demographic attributes	44.2	(16.5)	**98.1**	(12.4)	52.4	(23.4)	99.4	(3.4)	49.5	(27.6)	70.0	(-27.2)
	random	34.9		84.8		18.2		99.2		35.7		98.8	
	+demographic attributes	60.9	(26.0)	59.5	(-25.3)	78.2	(59.9)	79.2	(-20.0)	35.3	(-0.4)	85.4	(-13.4)
	similarity	38.3		78.5		34.0		95.0		36.5		**99.4**	
	+demographic attributes	60.6	(22.3)	68.4	(-10.1)	78.3	(44.3)	82.8	(-12.2)	**53.8**	(17.3)	97.4	(-2.1)
	diversity	37.1		74.7		1.7		**100**		24.6		97.4	
	+demographic attributes	**64.7**	(27.5)	62.6	(-12.1)	**80.0**	(78.3)	81.0	(-19.0)	47.7	(23.1)	85.7	(-11.8)
	within	35.5		76.9		18.4		98.4		35.6		95.4	
	+demographic attributes	57.7	(22.2)	74.4	(-2.4)	77.4	(59.0)	76.8	(-21.6)	37.9	(2.3)	92.3	(-3.2)
	representative	36.6		77.1		17.7		98.4		36.7		98.0	
	+demographic attributes	62.9	(26.3)	65.1	(-12.0)	78.3	(60.6)	76.8	(-21.6)	37.2	(0.5)	86.3	(-11.7)

Table 5: Performance of open source models across datasets when adding demographic attributes to the demonstrations and prompt. Results without demographic attributes are shown as comparison, as well as a difference between them. **Bold** is best per model×dataset and underlined is best per dataset (above a random baseline). We shade results that have an F1 score below a random baseline.

model	selection method	HateXplain race F1	HateXplain race 1-GAP	Bias in Bios F1	Bias in Bios 1-GAP	Twitter Sentiment F1	Twitter Sentiment 1-GAP
UL2	zero-shot	**53.5**	92.7	**19.2**	99.6	58.1	98.6
	random	44.3	99.1	2.5	**100**	48.2	92.6
	similarity	44.3	96.7	11.5	**100**	**65.0**	99.9
	diversity	44.4	**100**	0.9	**100**	33.5	**100**
	representative	44.3	96.8	2.4	**100**	47.9	94.1
	within	44.4	**100**	2.4	**100**	47.8	83.6
	+similarity	44.3	96.8	2.1	**100**	48.5	97.6
	+diverse	44.4	**100**	1.9	**100**	50.6	02.4
Flan-UL2	zero-shot	60.9	71.0	**86.7**	92.8	69.5	99.6
	random	68.4	83.8	84.2	84.6	69.7	99.1
	similarity	68.6	85.6	85.3	87.4	**70.0**	**99.9**
	diversity	68.3	83.5	85.4	83.0	69.6	98.8
	representative	**69.1**	82.6	84.5	84.4	69.8	98.6
	within	68.9	82.3	84.5	85.0	69.8	98.8
	+similarity	50.3	**87.2**	31.9	100	59.4	96.4
	+diverse	68.6	86.3	85.2	88.0	69.4	93.5
LLaMA-13B	zero-shot	22.3	77.5	11.5	99.8	36.9	97.8
	random	31.3	69.1	74.2	82.0	55.8	97.0
	similarity	**48.5**	52.6	**78.7**	95.6	**64.5**	98.9
	diversity	23.5	75.7	78.3	83.0	51.6	97.8
	representative	32.0	78.2	73.6	81.8	54.8	95.6
	within	36.0	48.7	73.0	78.4	56.0	93.5
	+similarity	37.3	**81.8**	11.3	**100**	47.0	**99.5**
	+diverse	25.5	29.0	77.0	91.8	63.9	75.0
LLaMA-65B	zero-shot	40.5	**84.6**	8.0	99.4	00.4	99.8
	random	44.7	76.4	73.7	86.0	54.7	96.4
	similarity	**52.2**	79.6	74.1	83.6	61.2	93.6
	diversity	49.6	60.7	**82.1**	84.6	49.9	93.4
	representative	48.8	68.7	74.7	88.4	54.3	94.5
	within	47.2	71.3	73.2	85.2	54.6	92.5
	+similarity	41.0	81.5	8.6	**100**	44.1	**99.8**
	+diverse	48.0	73.6	79.9	96.6	**62.0**	73.0
Alpaca-7B	zero-shot	28.7	87.9	2.3	99.8	35.9	92.0
	random	48.8	66.1	76.7	78.2	2.2	100
	similarity	**52.2**	82.9	**82.1**	79.8	10.2	98.9
	diversity	45.6	78.6	80.6	83.4	0.0	**100**
	representative	48.9	**92.8**	76.1	79.6	2.1	99.9
	within	45.7	80.2	76.3	78.4	2.5	99.5
	+similarity	49.3	80.4	8.7	**100**	36.2	99.5
	+diverse	50.3	71.0	76.8	93.2	**58.9**	96.7
Alpaca-13B	zero-shot	27.7	**85.7**	29.0	96.0	21.9	97.2
	random	34.9	84.8	18.2	99.2	35.7	98.8
	similarity	38.3	78.5	34.0	95.0	36.5	**99.4**
	diversity	37.1	74.7	1.7	**100**	24.6	97.4
	representative	36.6	77.1	17.7	98.4	36.7	98.0
	within	35.5	76.9	18.4	98.4	35.6	95.4
	+similarity	44.3	74.6	11.4	**100**	**37.3**	98.0
	+diverse	**59.1**	66.9	**79.9**	82.6	33.6	76.9

Table 6: Performance of open source models across datasets for demonstration selection methods that select based on semantic similarity within the same demographic category (*within* + *similarity*) and semantic diversity within the same demographic (*within* + *diversity*). We show results for other selection methods for context. **Bold** is best per model×dataset and underlined is best per dataset (above a random classifier baseline). We shade results that have an F1 score below a random class. baseline.

$$\min_{\theta} f(X;\theta) \triangleq \frac{1}{N}\sum_{i=1}^{N}\mathcal{L}(x_i;\theta)+\lambda[\max(0,\epsilon(X;\theta)-\epsilon_t)] \quad (2)$$

where $\epsilon(X;\theta)$ is the ϵ-DEO measure, eq. (1), for the classifier, ϵ_t is the desired base fairness (in our experiments 0), and λ is a hyper-parameter that trades between prediction loss and fairness (Foulds et al., 2020). Since the fairness term is differentiable, the model can be trained using stochastic gradient descent on the objective via backpropagation and automatic differentiation. A *burn-in* period and stochastic approximation-based update are adopted following Foulds et al. (2020).

To obtain the best performing model, we use a grid search for each task, with a learning rate= $[1e^{-4}, 1e^{-5}, 1e^{-6}]$ with Adam optimizer (Kingma and Ba, 2014), batch size= $[16, 32, 48]$, warmup= $[.1, .05, .005]$, epsilon= $[1e-7, 1e-8, 1e-9]$, *burn-in*= $[.5, 1]$, $\lambda = [.01, .1]$ and $\rho = [.9, .1, .01]$. We select the best performing model on development data and report test data results.

C Hyperparameter Experiments

When considering the performance of LLMs for classification it may be important finetune the hyperparameters for generation. In this section, we report the result of experiments when varying the temperature parameter across datasets. Since we evaluate on 12 models across 3 datasets and 6 demonstration selection methods (total of 216 settings), varying the temperature for all settings is not practical. Thus, we select the best performing open-source model, FLAN-UL2 for this experiment.

Figure 4 shows the results for performance (F1) and fairness (1-GAP) for FLAN-UL2 across all three datasets. We observe little difference when varying temperature in the classification performance and the fairness of the model across demonstration selection strategies.

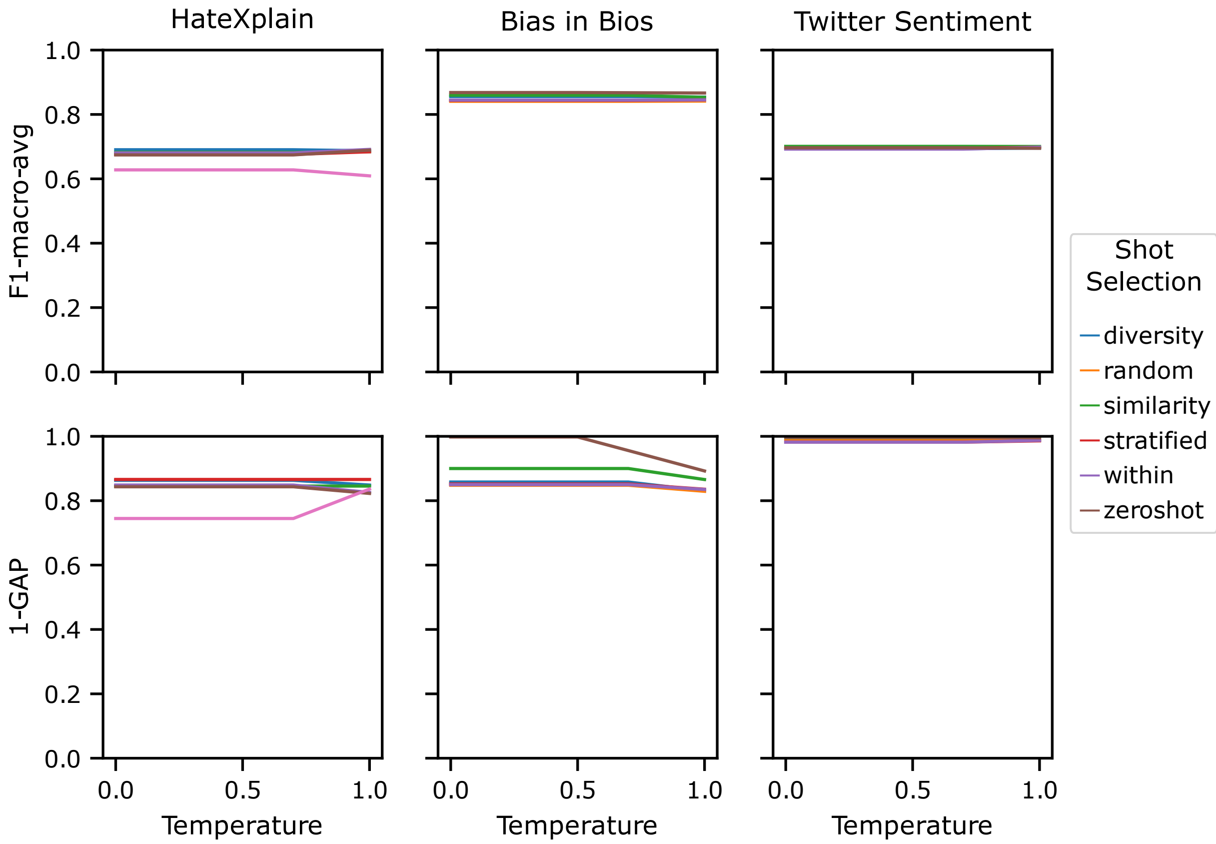


Figure 4: Results of varying temperature across datasets for Flan-UL2. No meaningful difference found.