

Identifying Bias in Machine-generated Text Detection

Kevin Stowe, Svetlana Afanaseva, Rodolfo Raimundo, Yitao Sun, Kailash Patil
Pindrop

{kevin.stowe, safanaseva, rraimundo, ysun, kpatil}@pindrop.com

Abstract

The meteoric rise in text generation capability has been accompanied by parallel growth in interest in machine-generated text detection: the capability to identify whether a given text was generated using a model or written by a person. While detection models show strong performance, they have the capacity to cause significant negative impacts. We explore potential biases in English machine-generated text detection systems. We curate a dataset of student essays and assess 16 different detection systems for bias across four attributes: gender, race/ethnicity, English-language learner (ELL) status, and economic status. We evaluate these attributes using regression-based models to determine the significance and power of the effects, as well as performing subgroup analysis. We find that while biases are generally inconsistent across systems, there are several key issues: several models tend to classify disadvantaged groups as machine-generated, ELL essays are more likely to be classified as machine-generated, economically disadvantaged students' essays are less likely to be classified as machine-generated, and non-White ELL essays are disproportionately classified as machine-generated relative to their White counterparts. Finally, we perform human annotation and find that while humans perform generally poorly at the detection task, they show no significant biases on the studied attributes.

1 Introduction

With the rise in the usage of generative AI systems, there is a growing need to distinguish content generated by a model from human-written content. To this end, there has been an explosion of research into machine-generated text detection¹: identifying content that has been automatically generated by generative systems (e.g., large language models).

¹Also referred to as "deepfake" or "LLM-generated" text detection.

These systems span detection of video, audio, and text-based generation, and are becoming essential tools for many practical scenarios where guidelines require human-written content.

It is important to consider the practical implications of machine-generated text detection systems. In this work, we assess the potential for bias in these detection systems. There is substantial potential for harm in systems that erroneously flag content as automatically generated, especially if these systems exhibit bias toward disadvantaged populations. This could lead to rejection of genuine work, reduction of visibility, and unfair allocation of resources. Consider student essays, where inaccurate false positives could lead to harmful consequences for students, or content moderation tools, where human perspectives and representation could be unfairly filtered or silenced.

To assess bias in machine-generated text detection systems, we systematically explore publicly available systems, analyzing their potential to unfairly classify human-written text as machine-generated. We curate a dataset of human-written texts and explore a series of publicly available models for potential biases across several dimensions: race, gender, age, ELL status, and economic status. We pursue the following research questions:

1. **Do machine-generated text detection systems exhibit bias?** We are particularly interested in bias across dimensions of gender, race, age, ELL status, and economic status. We perform regression analysis, evaluating each attribute along with potential confounds.
2. **Which subgroups are likely to be impacted?** We analyze each of the 16 possible subgroups, evaluating performance compared to the general population.
3. **Do humans exhibit the same biases?** We have expert human annotators attempt to classify text as machine-generated or human-

written, similarly evaluating their results for potential biases.

We find that while many English language learners texts are classified at a significantly higher rate than native speakers, these results are inconsistent, and models vary greatly in their respective biases. Several models tend to misclassify texts from disadvantaged groups as machine-generated, while other trained and zero-shot models vary. Across all attributes, we find that economic disadvantage serves as a negative indicator: essays from students with no economic disadvantage are classified as machine-generated more often across most systems. Subgroup analysis reveals more significant issues: non-White ELL essays are much more likely to be classified as machine-generated than White ELL essays, with higher incidence for men than for women. Finally, we find that human experts, while generally poor at the task (with accuracy ranging from 0.449 to 0.526), exhibit no significant biases based on the attributes studied.

2 Background

Our goal is to identify situations in which machine-generated text detection models make consistent mistakes with regard to certain attributes, violating group fairness (Czarnowska et al., 2021), which can lead to representational harms (e.g., protected groups misrepresented as abusing generative tools) and allocational harms (e.g., writing of protected groups being disqualified, censored, or minimized due to unfair model performance). We adopt their terminology: we analyze four **sensitive attributes**: gender, race/ethnicity, ELL status, and economic status. Within these attributes we define **advantaged/disadvantaged** groups (see Section 3.1).² Our evaluation framework is based on Dayanik et al. (2022), who outline a method for identifying bias across multiple attributes in NLP problems.

2.1 Machine-generated Text Detection

Recent advances in generative AI have had many benefits, but understanding whether text has been written by a large language model (LLM) or a human is often essential. Domains such as news, where generative models can be used to spread misinformation (Hanley and Durumeric, 2024; Pan et al., 2023), education, where the use of generative

²We use "disadvantaged" rather than "protected" as only two of our attributes, gender and race/ethnicity, are considered protected in the United States.

AI in assisting students is under scrutiny (Meyer et al., 2023), and fraud, where generative models are being used to perpetrate scams (Romero-Moreno, 2025) highlight the need for accurate detection of machine-generated text.

The power of generative models has been accompanied by increased interest in detection systems (including a workshop focused specifically on this task (Alam et al., 2025)). There have been a wide variety of systems proposed, including feature-based methods, fine-tuned models, and zero-shot systems. For a comprehensive survey of recent datasets and methods, see Wu et al. (2025).

2.2 Bias in Models

Bias has been extensively studied in deepfake detection systems across domains of video and audio (Liu et al., 2025; Katamneni et al., 2024; Ju et al., 2024), and as a ubiquitous aspect of natural language processing (Bartl et al., 2025; Stanczak and Augenstein, 2021; Blodgett et al., 2020). Recent work has shown that humans exhibit significant biases when attempting to identify deepfake social media profiles (Mink et al., 2024). However, there remains relatively little work concerning the bias of machine-generated text detection systems.

Jung (2025) explores this topic in considerable depth, but focuses only on text length and stylistic personality as biases, which excludes disadvantaged groups. Liang et al. (2023) explore bias in machine-generated text detection against non-native English speakers, claiming that seven major detectors perform significantly worse on non-native English speakers than native speakers. However, they do not indicate which models were used, and they report results on a limited dataset of only 179 student essays. Verma et al. (2024) echo this result for English learner data, but note that it was unclear whether the differences were due to language or the length of the documents.

To address this gap in our understanding, we evaluate a suite of machine-generated text detection systems on a large corpus of student essays, evaluating performance for bias across four attributes: gender, race, ELL status, and economic status. We provide a thorough analysis of possible confounding factors as well as exploring subgroup differences. To our knowledge, this is the first work to perform a rigorous analysis of bias in machine-generated text detection.

	Total	Gender		Race/Ethnicity		ELL		Economic Disadvantage	
		Male	Female	White	Non-White	No	Yes	No	Yes
PERSUADE V2.0	24695	12074	12621	11282	13413	22451	2244	11003	13692
ASAP V2.0	24728	12498	12230	9841	14887	20991	3737	7933	16795
ELLIPSE	6482	3636	2846	471	6011	0	6482	1974	4508
Combined (clean)	41743	21277	20466	15078	26665	31079	10664	18188	23555

Table 1: Counts for each attribute/group in our combined corpus.

3 Evaluation Datasets

To evaluate potential bias in machine-generated detection systems, we use three datasets. These are PERSUADE-V2.0 (Crossley et al., 2024), ASAP-V2.0 (Crossley et al., 2025), and ELLIPSE (Crossley, 2024). These datasets all consist of persuasive essays written by 6th to 12th grade students in the United States, containing demographic information about race, gender, English-language-learner status, and economic status of the writers. The ASAP-V2.0 and ELLIPSE datasets extend the PERSUADE-V2.0 dataset: ASAP-V2.0 adds 12k new samples and fills in demographic gaps, while the ELLIPSE dataset adds new samples focusing on English language learners. We combine these datasets into a single, cleaned version, removing duplicate texts as well as instances where any demographic information is missing; details are in Table 7.

3.1 Sensitive Attributes

Gender: The corpora contain gender as a binary attribute (male/female). We use these labels, with male considered the advantaged class and female as the disadvantaged class: machine learning models are known for biased performance on female data (Bartl et al., 2025). We recognize this binary labeling prohibits proper study and representation of other potential gender labels. Our setup mirrors previous work in which datasets contain only binary labels out of necessity (Biester, 2025; Plaza-del Arco et al., 2024; Savoldi et al., 2021), and we continue with the understanding that this labeling system contains an inherent risk of misrepresenting bias, erasure, and other representational and allocational harms with regard to non-binary genders (Stanczak and Augenstein, 2021; Dev et al., 2021).

Race/ethnicity: The dataset contains six different labels for race/ethnicity. For our initial analysis, we compress these into two groups: White and non-White: for this attribute, this reflects the majority group (White) and the minority group (non-White) (Economic Policy Institute, 2022). This

posits White as the advantaged class and non-White disadvantaged, but there are substantial differences between subgroups: we provide further exploration of the differences between these in Section 8. We note such fixed categorization schemes can serve to entrench inequalities, and that racial divisions are a product of social contexts (Field et al., 2021; Hanna et al., 2020). Our goal is to examine potential biases in these models, but we stress our analysis inherits potential pitfalls from this labeling system.

English-language learner (ELL) status: The corpora make a binary distinction between ELL and non-ELL students; we consider ELL the disadvantaged attribute and non-ELL the advantaged.

Economic status: The corpora define two economic statuses: not disadvantaged (the "advantaged" group) and disadvantaged.

4 Models

There are many options for machine-generated text detection models: we focus on an array of zero-shot and pretrained models.³

4.1 Zero-shot Models

We split zero-shot models into GPT-based and non-GPT-based systems. All are suitable for zero-shot detection: they can score an independent text without training or other context. While not necessarily tuned to specifically detect GPT-based generation, the GPT-based models are clustered together as their dependence on OpenAI models makes them somewhat harder to inspect, and their behavior may change as access to these models changes. The models used are Ghostbuster (Verma et al., 2024) and Glimpse (Bao et al., 2025). For non-GPT models, we utilize Fast-DetectGPT (Bao et al., 2024) and Binoculars (Hans et al., 2024), which rely on extracting features from an underlying transformer-based language model, as well as Zippy (Thinkst, 2023), which uses compression-based methods.

³Model implementation details provided in Appendix B.

4.2 Trained Models

These are publicly available models that have been trained through varying methods on various datasets. We use BiScope (Guo et al., 2024a), which has four variants (Yelp, Arxiv, Essay, and Creative) based on training data, and DeTeCtive (Guo et al., 2024b), which has four variants of which we use three (MAGE, M4GT, TuringBench), as the fourth OUTFOX variant overlaps with our evaluation data.

Another subset consists of fine-tuned versions of transformer models. They have either been tuned for a specific task (Desklib (Desklib, 2025) and e5-lora (Dugan et al., 2024) are optimized for the RAID benchmark⁴), or are designed to be generally applicable for machine-generated text detection (RADAR (Hu et al., 2023)).

5 Benchmarking

We start by benchmarking the models to better understand their overall performance, and then explore potential biases on human-written corpora. To benchmark our models, we utilize the OUTFOX dataset (Koike et al., 2024). This dataset combines human-written texts from the PERSUADE-V2.0 corpus with three machine-generated samples for each human-written text. This corpus comes from the same source, matching the domain, style, and tone of our human-written evaluation corpus.

We evaluate each model, reporting precision, recall, F1 score, and area under the receiver operating characteristics (AUROC) in Table 2. These metrics provide a broad overview of performance: they have different implications for different use cases, with precision minimizing false positives, recall maximizing coverage, and AUROC providing a balanced view across thresholds. We convert model scores into binary classification by identifying the threshold that optimizes equal error rate (EER) over a validation set of 1000 samples. We then use this threshold to make predictions, considering a sample machine-generated if the score provided by the model exceeds this threshold. This result is strictly improved F1 scores while keeping constant AUROC.

We find that model performance is fairly disparate: the zero-shot models are mostly strong except for Zippy. The trained models depend heavily

⁴<https://raid-bench.xyz/leaderboard>; at the time of writing, these are the two top-performing, publicly available systems.

Model	Prec.	Rec.	F1	AUROC
Ghostbuster	0.638	0.606	0.622	0.667
Glimpse	0.899	0.861	0.880	0.948
Binoculars	0.869	0.825	0.846	0.907
FDG (falcon-7b)	0.670	0.635	0.652	0.708
FDG (gpt-neo)	0.781	0.733	0.756	0.829
Zippy (LZMA)	0.359	0.327	0.343	0.262
BiScope (Yelp)	0.706	0.691	0.699	0.726
BiScope (Arxiv)	0.404	0.381	0.392	0.327
BiScope (Essay)	0.841	0.761	0.799	0.805
BiScope (Creative)	0.388	0.302	0.339	0.362
DeTeCtive (MAGE)	0.470	0.127	0.200	0.477
DeTeCtive (M4GT)	0.863	0.472	0.610	0.696
DeTeCtive (TuringBench)	0.508	0.870	0.641	0.450
RADAR	0.700	0.613	0.653	0.706
Desklib	0.976	0.960	0.968	0.994
E5-lora	0.417	0.361	0.387	0.362

Table 2: Benchmark model performance for the investigated systems on a balanced corpus of human-written and LLM-generated texts.

on the dataset: BiScope performance ranges in AUROC from 0.362 to 0.805, with the Essay variant performing best, likely because it best matches the evaluation data domain. These models tend to struggle when applied to new domains, and while this is important to note, our primary goal is not to compare or evaluate the exact performance of these models, but rather to assess whether the mistakes they are making significantly favor certain groups. For this reason, we proceed with our bias analysis using all models, with the understanding that some may be better or worse suited to this task.

We also aim for relative model independence to cover a broad range of potential systems. We calculate Pearson correlations between all models: only 5.5% of pairs have correlation over 0.6, and none with correlation over 0.8, indicating models have weak to moderate correlation. The primary correlations are between BiScope models, where the Creative variant overlaps with the Essay and Yelp variants, and between the two FDG variants.⁵

6 Logistic Regression Analysis

To study bias, we need a methodology that can account for the relationships and confounds present in the data. The attributes we study are unlikely to be independent, and additional factors may influence model performance. To handle this, we perform our analysis based on the methodology of Dayanik et al. (2022), who outline procedures for identifying bias in natural language processing

⁵For more, see Appendix C.

Model	Gender			Race/Ethnicity			ELL Status			Econ. Status		
	Diff.	Coef.	DA	Diff.	Coef.	DA	Diff.	Coef.	DA	Diff.	Coef.	DA
Ghostbuster	-.102	-.139‡	2.98	.041	-.244‡	1.01	-.143	-.945‡	6.50 ↓	-.120	-.338‡	5.87 ↓
Glimpse	-.046	-.338‡	1.07	-.016	-.012‡	0.32	-.047	-.199‡	0.86	-.036	-.444‡	0.87
Binoculars	-.042	-.168‡	1.00	-.002	-.548‡	0.09	-.010	-.475‡	0.53	-.059	-.280‡	2.25
FDG (falcon-7b)	.068	.167‡	2.80	.007	.361‡	0.14	.003	.195‡	1.13	.080	.185‡	4.09
FDG (gpt-neo)	.052	.182‡	4.65	.018	.267‡	0.85	-.019	-.081	3.87	.047	.115‡	3.39
Zippy (LZMA)	-.036	-.146‡	1.39	.009	.471‡	0.26	-.135	-.286‡	14.84 ↓	.048	.177‡	1.34
BiScope (Yelp)	.052	.085	3.36	-.001	.141‡	0.01	.022	.090	0.52	.074	.224‡	10.10 ↑
BiScope (Arxiv)	.028	.088‡	0.90	.010	.301‡	0.40	-.035	.042	5.01	.058	.184‡	4.06
BiScope (Essay)	.020	.082	2.74	.011	.048	0.71	-.016	-.280‡	5.25 ↓	.024	.166‡	6.03 ↑
BiScope (Creative)	.073	.255‡	2.95	-.016	.311‡	0.14	-.009	.162‡	1.81	.102	.344‡	6.16 ↑
DeTeCtive (MAGE)	-.056	-.185‡	5.53 ↓	.008	.038	0.18	-.180	-.249‡	79.12 ↓	-.004	.105	0.54
DeTeCtive (M4GT)	.008	-.058	0.25	-.008	-.109	1.49	.009	.083	0.26	.017	.288‡	4.39
DeTeCtive (TuringBench)	.004	.010	0.15	-.004	.238‡	0.13	-.008	-.344‡	1.96	.014	.129	1.31
RADAR	.007	.037	0.70	.049	-.111‡	6.48 ↓	.006	-.470‡	0.84	-.035	-.310‡	3.72
Desklib	-.002	.001	0.04	-.002	-.396‡	1.02	-.001	-.274‡	0.04	-.007	.027	0.73
E5-lora	-.044	.074	0.37	.103	.186‡	5.15 ↑	-.029	-.212‡	0.21	-.125	-.257‡	5.49 ↓

Table 3: Model performance differences (Diff.), attribute coefficients (Coef., $p < 1.56e - 4$) and Dominance Analysis scores (DA) for each model. ↑ indicates the advantaged class is more likely classified as machine-generated; ↓ indicates the disadvantaged class is more likely machine-generated.

systems where there may be many factors involved.

We train a logistic regression model over relevant features as well as confounds to predict the error of the model, and use feature coefficients as well as dominance analysis to assess the impact of each attribute. For bias variables, we use the sensitive attributes: gender (male/female), race/ethnicity (White/non-White), English language-learner status (no/yes), and economic status (not disadvantaged/disadvantaged). We then define covariates, which are other potential predictors of error: we use perplexity via the opt-iml-1.3b model (Iyer et al., 2023) and length of the text in words (defined by splitting on whitespace). We calculate variance inflation factor (VIF) for each bias variable and covariate, finding the values to all be below 4, therefore suitable with minimal multicollinearity. We report the following:

Performance Difference (Diff.): The difference in model performance for each attribute. This is calculated as the mean score for the advantaged class minus that of the disadvantaged class. Higher scores indicate that the advantaged class is more likely to be classified as machine-generated; lower scores indicate the disadvantaged class is more likely to be classified as machine-generated.

Coefficient (Coef.): The coefficient for the given attribute in the regression model. Positive scores indicate the advantaged class predicts higher

machine-generated scores; negative scores indicate the disadvantaged class. For all experiments, we use a significance threshold of 0.01 with Bonferroni correction, using the number of models (16) and the number of categories, yielding a p value of $0.01/(categories \times models)$; we note the specific values in each experiment.

Dominance analysis (DA): This indicates the strength of this attribute’s contribution in the logistic regression model. We report dominance as percentage of relative importance (e.g., 4.73 indicates that 4.73% of the prediction comes from this attribute). We consider dominance scores over 5% to be meaningful.

Significance tests are useful for detecting the presence of systematic differences, but do not capture the magnitude of difference (Dayanik et al., 2022; Stanczak and Augenstein, 2021). Hence, we are most interested in cases where both the coefficient from the model is significant (indicating a significant relationship between the attribute and model performance) and the dominance factor is strong (> 5 , indicating that at least 5% of the difference in performance is due to this attribute): these cases are marked with ↑ (the advantaged class is more likely to be classified as machine-generated) and ↓ (the disadvantaged class is more likely to be classified as machine-generated). Table 3 shows these results over all models and categories.

6.1 General Analysis

We observe relatively few instances where model coefficients are significant and dominance analysis indicates a strong influence of the corresponding attribute (12 out of 64 total observations). These cases also tend to be inconsistent, showing no systematic preference for either advantaged or disadvantaged groups. The primary exception is ELL status, where most models incorrectly classify essays written by ELL students as machine-generated; four models exhibit both significant coefficients and dominance analysis results. We further analyze differences by model and attribute.

6.2 Model Analysis

Most models exhibit inconsistent and minor biases across all categories. The two GPT-based models (Ghostbuster and Glimpse), as well as Binoculars, RADAR, and the DeTeCtive (MAGE) variant, tend to misclassify the disadvantaged population as machine-generated, although the associated dominance is often minimal: while the effects are statistically significant, the attribute plays only a minor role in the resulting classification. FDG models are relatively inconsistent, while most trained models tend to misclassify essays from both ELL students and those without economic disadvantages as machine-generated.

6.3 Attribute Analysis

We observe relatively limited impact from gender and race in this analysis, though subgroup analysis may provide more insight. While many models exhibit significant coefficients, these are not reflected in the dominance analysis, suggesting a limited role in classification decisions. The affected groups also vary: different models and variants exhibit minor biases in both directions.

ELL status shows consistent negative effects: ELL essays are more likely to be classified as machine-generated by most models. While this general trend aligns with prior research indicating that ELL students are unfairly treated by detection systems, we note that the magnitude of this effect is typically small.

Economic status shows a relatively strong but mixed effect. Essays from students without economic disadvantage are more likely to be classified as machine-generated by BiScope and FDG models. However, other zero-shot and fine-tuned models present potential risks: the GPT-based Ghostbuster

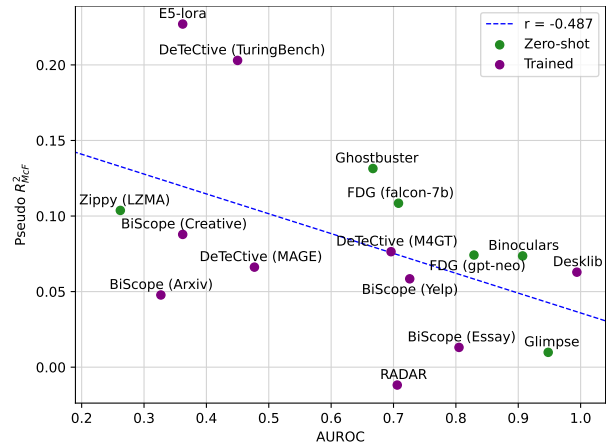


Figure 1: Pseudo- R^2 values from the regression analysis plotted against AUROC scores for each model.

and Glimpse models, as well as all three trained variants, exhibit negative coefficients, which in two cases correspond with higher dominance values.

6.4 Overall Results

From the regression analysis, we conclude the following: (1) ELL status appears to be a major contributing factor, with ELL student essays more likely to be classified as machine-generated, consistent with prior research; (2) economic status is an important variable, although results vary across model types; and (3) there is relatively little evidence of bias related to race or gender.

We further examine model performance in relation to overall bias. In Figure 1, we plot each model’s AUROC score against its McFadden’s pseudo- R^2 value from the regression analysis. While R^2 typically measures variance explained in linear regression, this approximation is used here as a general indicator of bias in logistic regression models. We observe a general negative correlation: as model performance (AUROC) increases, the estimated bias (R^2_{McF}) decreases ($r = -0.486$). This trend suggests that higher-performing models may exhibit lower bias and thus reduce potential harms.

7 Subgroup Analysis

While regression analysis offers a broad overview of biases by attribute, it may obscure subgroup-specific effects. To explore these, we partition the dataset into 16 subgroups representing all combinations of the four attributes under study. We then perform pairwise z-tests comparing each subgroup’s scores with those of the remaining dataset, identifying statistically significant differences in

Econ. Disadvantage	Gender Race ELL	Male								Female							
		White				Non-White				White				Non-White			
		No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Ghostbuster		-.09	-	-	-	-.06	.06	.10	.15	-.16	-	-	-	-.11	-	.12	.13
Glimpse		-.05	-.03	-	-	-	-	-	.03	-.05	-	-	-	-	.02	.05	.07
Binoculars		.05	-	-	-	-	-.04	-	-.03	.05	-	-	-	-	-.04	-	-
FDG (falcon-7b)		.08	-	-	-	-	-.07	-	-	.08	-	-	-	-	-.07	-	-
FDG (gpt-neo)		.06	-	-	-	-	-.05	-	-	.05	-	-	-	-	-.06	-	-
Zippy (LZMA)		-	-.11	-	-	-	-.07	.14	.11	-	-.10	-	-	-	-.08	.14	.12
BiScope (Yelp)		.07	-	-	-	-	-.05	-	-	.10	-	-	-	-	-.06	-	-
BiScope (Arxiv)		.05	-.04	-	-	-	-.05	-	.04	.06	-	-	-	-	-.07	-	-
BiScope (Essay)		.04	-	-	-	-	-	-	-	.04	-	-	-	-	-.04	-	-
BiScope (Creative)		.08	-.04	-	-	-	-.10	-	-	.13	-	-	-	.04	-.10	-	-
DeTeCtive (MAGE)		-.03	-.08	-	.10	-.04	-.06	.15	.16	-	-.08	-	-	-.03	-.07	.15	.16
DeTeCtive (M4GT)		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
DeTeCtive (TuringBench)		-	-	-	-	-	-.02	-	-	.01	-	-	-	-	-.01	-	-
RADAR		-	-.08	-	-	-	-.04	-	-	.07	-	-	-	.05	-	-	-
Desklib		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
E5-lora		-.03	.12	-	-	-	.14	.06	.08	-.18	-	-.14	-	-.16	-	-	-

Table 4: Results for subgroup analysis: we report the differences in error from the given subgroup to those not in that subgroup, with **positive** scores indicating the subgroup is more likely classified as machine generated and **negative** indicating the opposite. We report only statistically significant differences ($p < 3.91e - 5$).

classification. Results are presented in Table 4.

This analysis reveals several notable findings not evident in the overall regression results. Although ELL essays are generally more likely to be classified as machine-generated by a large number of models, this effect is disproportionately concentrated among non-White students. Specifically, non-White ELL essays ($n = 9, 443$) are more frequently misclassified by seven different models, compared to only one model for their White counterparts ($n = 1, 221$). Moreover, this effect is more pronounced among males: three additional models exhibit significant differences for male non-White ELL essays compared to females.

We therefore need to revise our earlier assessment that bias related to race and gender is minimal. Subgroup analysis suggests that both race and gender play a substantial role, highlighting the need for more rigorous intersectional analysis when approaching potential biases.

For non-ELL essays, some differences are significant, but results are inconsistent. Interestingly, essays from ELL students with economic disadvantages are often less likely to be misclassified, though this outcome varies considerably by model.

8 Race/Ethnicity

We initially conducted analysis using a simplified binary race/ethnicity classification (White

Model	W	H/L	A/PI	B/AA	Two+	AI/AN
Ghostbuster	-	-	-	-	-	-
Glimpse	-.05	.03	-	.05	-.02	-.02
Binoculars	-	-	-	-	-	-
FDG (falcon-7b)	.04	-.02	.06	-.07	-.01	.04
FDG (gpt-neo)	.04	-.01	.02	-.06	-.02	.05
Zippy (LZMA)	-.02	.05	.10	-.07	-.07	-.04
BiScope (Yelp)	-	-.00	-	-	-	-
BiScope (Arxiv)	.02	-.01	.02	-.03	-.01	.01
BiScope (Essay)	-	-	-	-	-	-
BiScope (Creative)	-	-	-	-	-	-
DeTeCtive (MAGE)	-.06	.08	.05	-.04	-.06	.02
DeTeCtive (M4GT)	.01	-.01	.02	-.01	.01	-.01
DeTeCtive (TuringBench)	-	.01	.01	-.02	-	.01
RADAR	-	-	.05	-.01	-.02	-.01
Desklib	-	-	-.01	-	.01	.04
E5-lora	-.05	.04	-.09	.08	-.03	-.01

Table 5: Performance differences on race/ethnicity groups. Highlighted values indicate significance based on z-scores between this attribute and the rest of the dataset ($p < 1.04e - 4$).

vs. non-White). Here, we extend our analysis to explore individual race and ethnicity categories. The dataset includes six groups: White (W), Hispanic/Latino (H/L), Asian/Pacific Islander (A/PI), Black/African American (B/AA), Two or More Races/Other (Two+), and American Indian/Alaskan Native (AI/AN). We evaluate model performance on each individual group and compare

it against the full dataset in Table 5.

First, we note that no models exhibited significant performance differences for the AI/AN subgroup, likely due to the small sample size ($n = 208$). This lack of significance should not be interpreted as conclusive evidence of no bias, but rather as an indication that the dataset is underpowered for detecting such effects. Further investigation with more representative data is warranted.

Two groups, however, show consistent disparities: A/PI essays are more likely to be classified as machine-generated by most models, while B/AA essays are less likely to be misclassified in this way. Results for H/L and W essays are inconsistent, while the Two+ category shows significant negative effects in only two models.

9 Human Performance

Identifying machine-generated text remains a challenging task for humans (Dugan et al., 2023; Ethayarajh and Jurafsky, 2022; Clark et al., 2021). Lee et al. (2025) finds that, even with the aid of collaborative tools, human accuracy in this task reaches only 57%. We evaluate human performance with respect to potential bias: given the same dataset, do human annotators exhibit biases comparable to those observed in automated detection systems?

To investigate this, we selected a balanced subsample of our corpus consisting of 800 total texts, with at least 100 examples from each group across the four key attributes: gender (male/female), race (White/non-White), ELL status (ELL/non-ELL), and economic status (disadvantaged/not disadvantaged). For each text, we used Claude Sonnet 3.5 (Anthropic, 2024) to generate a continuation based on the first twenty tokens.⁶ This process resulted in a dataset containing equal numbers of human-written and LLM-generated texts, balanced across all demographic categories.

Three expert annotators were each assigned a subset of these texts ($231 < n < 318$), including a minimum of 25 human-written samples for each sensitive attribute. Annotators were instructed to classify each text as either human-written or machine-generated. We then applied the same logistic regression analysis used in Section 6 to assess potential biases in human predictions. Performance differences are presented in Table 6.⁷

Our results align with previous findings regard-

⁶See Appendix D for full prompt details.

⁷Full annotator details are provided in Appendix E.

Annotator	Accuracy	Gender	Race/Ethnicity	ELL Status	Econ. Status
Ann. 1	0.492	0.060	-0.069	0.013	0.150
Ann. 2	0.449	-0.096	-0.033	0.131	0.011
Ann. 3	0.526	-0.011	-0.004	0.018	0.183

Table 6: Differences in human classifications based on attributes. Positive numbers indicate the advantaged attribute is more likely classified as machine-generated; negative indicate the opposite. No results were indicated as significant for $p < .01$.

ing human performance: annotators performed at approximately chance. However, we found no significant differences in classification based on the studied attributes. While slightly elevated coefficients were observed for economic status (mirroring trends in the system evaluations), these were not statistically significant.

10 Conclusions

This work investigates bias in machine-generated text detection systems across four key attributes: gender, race/ethnicity, ELL status, and economic status. We find that several models tend to disproportionately affect disadvantaged groups; essays written by ELL students are more frequently misclassified as machine-generated, and this effect is particularly pronounced among non-White students. We also observe that while human annotators perform poorly at this task, they do not exhibit significant biases.

The key takeaway for practitioners is the critical importance of understanding the behavior and limitations of machine-generated text detection models. Misclassification presents a substantial risk, not only in this context but in other domains where such models may be applied. Our findings show no singular or consistent bias across all systems, underscoring the need for case-by-case evaluation. To ensure fairness, models and their predictions must be carefully scrutinized for disproportionate impacts on disadvantaged populations. AI developers and regulators can support this goal by creating and adopting datasets and metrics such as those proposed here that allow for the detection and mitigation of bias before real-world deployment.

11 Limitations

This study represents an initial step in analyzing bias in machine-generated text detection systems, but it is necessarily constrained in several ways.

11.1 Models

We examine only a limited subset of models commonly used for detecting machine-generated text. Our selection criteria emphasized public availability, broad use, and general applicability. While we aimed for methodological diversity, many relevant models remain outside the scope of this analysis, and our findings should not be assumed to generalize across all possible systems.

11.2 Dataset

Our evaluation data is similarly constrained. It consists entirely of student essay writing, drawn from three datasets produced by the same organization. All datasets are in English, and written by students in the United States. This choice was driven by three considerations: (1) the datasets are publicly available and include detailed demographic information (an uncommon feature); (2) we had access to a corresponding machine-generated dataset (OUTFOX), facilitating benchmarking; and (3) the education domain represents a high-stakes use case, where misclassification could cause significant harm.

The consequence of these advantages is limited generalizability. Our findings may not extend to other text domains, and the dataset reflects a narrow slice of the broader population. Accordingly, the biases observed here may not reflect those that would occur when systems are evaluated on other demographics or styles of text.

11.3 Categories of Gender and Race

We acknowledge concerns regarding binary gender labels and predefined racial categories, as addressed in Section 3, and this remains a limitation. Prior work warns that such categorization may reinforce essentialist or harmful views of identity. We are constrained here by the demographic labels provided in the datasets. Future research should explore more inclusive and representative identity categorizations.

11.4 Statistical Methods

Numerous statistical approaches exist for evaluating model fairness, and this remains an active

area of research in bias in NLP. We sought to minimize methodological inconsistencies by following the statistical framework proposed by Dayanik et al. (2022), whose work closely aligns with our use case. However, we recognize that alternative methodologies could yield different insights.

11.5 Human Annotation

Our human annotation effort involved only three expert annotators. This limited scope was a function of prioritizing expertise over general human judgment. Although preliminary, our results suggest that human evaluators may introduce less bias than automated systems. However, the small sample size restricts the generalizability of this finding.

Overall, we acknowledge that this study is constrained by time, resources, and dataset availability. Many of the questions raised here warrant further investigation at larger scales and across more diverse settings.

References

- Firoj Alam, Preslav Nakov, Nizar Habash, Iryna Gurevych, Shammur Chowdhury, Artem Shelmanov, Yuxia Wang, Ekaterina Artemova, Mucahid Kutlu, and George Mikros, editors. 2025. *Proceedings of the 1st Workshop on GenAI Content Detection (GenAIDetect)*. International Conference on Computational Linguistics, Abu Dhabi, UAE.
- Anthropic. 2024. Claude 3.5 sonnet. <https://www.anthropic.com/claude>.
- Guangsheng Bao, Yanbin Zhao, Juncai He, and Yue Zhang. 2025. *Glimpse: Enabling white-box methods to use proprietary models for zero-shot llm-generated text detection*. Preprint, arXiv:2412.11506.
- Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. 2024. *Fast-detectgpt: Efficient zero-shot detection of machine-generated text via conditional probability curvature*. Preprint, arXiv:2310.05130.
- Marion Bartl, Abhishek Mandal, Susan Leavy, and Suzanne Little. 2025. *Gender bias in natural language processing and computer vision: A comparative survey*. *ACM Comput. Surv.*, 57(6).
- Laura Biester. 2025. *Sports and women’s sports: Gender bias in text generation with olympic data*. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 195–205, Albuquerque, New Mexico. Association for Computational Linguistics.

- Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. [Language \(technology\) is power: A critical survey of “bias” in NLP](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.
- Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. [All that’s ‘human’ is not gold: Evaluating human evaluation of generated text](#). 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 7282–7296, Online. Association for Computational Linguistics.
- S.A. Crossley, Y. Tian, P. Baffour, A. Franklin, M. Benner, and U. Boser. 2024. [A large-scale corpus for assessing written argumentation: Persuade 2.0](#). *Assessing Writing*, 61:100865.
- Scott Crossley. 2024. [The english language learner insight, proficiency and skills evaluation \(ellipse\) corpus](#). *International Journal of Learner Corpus Research*, 9(2).
- Scott A. Crossley, Perpetual Baffour, L. Burleigh, and Jules King. 2025. [A large-scale corpus for assessing source-based writing quality: Asap 2.0](#). *Assessing Writing*, 65:100954.
- Paula Czarnowska, Yogarshi Vyas, and Kashif Shah. 2021. [Quantifying social biases in nlp: A generalization and empirical comparison of extrinsic fairness metrics](#). *Transactions of the Association for Computational Linguistics*, 9:1249–1267.
- Erenay Dayanik, Ngoc Thang Vu, and Sebastian Padó. 2022. [Bias identification and attribution in NLP models with regression and effect sizes](#). *Northern European Journal of Language Technology*, 8.
- Desklib. 2025. [ai-text-detector: Ai-generated text detection model](#). Accessed: 2025-07-10.
- Sunipa Dev, Masoud Monajatipoor, Anaelia Ovalle, Arjun Subramonian, Jeff Phillips, and Kai-Wei Chang. 2021. [Harms of gender exclusivity and challenges in non-binary representation in language technologies](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1968–1994, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Liam Dugan, Alyssa Hwang, Filip Trhlfk, Andrew Zhu, Josh Magnus Ludan, Hainiu Xu, Daphne Ippolito, and Chris Callison-Burch. 2024. [RAID: A shared benchmark for robust evaluation of machine-generated text detectors](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12463–12492, Bangkok, Thailand. Association for Computational Linguistics.
- Liam Dugan, Daphne Ippolito, Arun Kirubarajan, Sherry Shi, and Chris Callison-Burch. 2023. [Real or fake text? investigating human ability to detect boundaries between human-written and machine-generated text](#). In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI’23/IAAI’23/EAAI’23. AAAI Press.
- Economic Policy Institute. 2022. [Racial and ethnic disparities in the united states: An interactive chartbook](#). Updated November 2024.
- Kawin Ethayarajh and Dan Jurafsky. 2022. [The authenticity gap in human evaluation](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6056–6070, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Anjalie Field, Su Lin Blodgett, Zeerak Waseem, and Yulia Tsvetkov. 2021. [A survey of race, racism, and anti-racism in NLP](#). 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 1905–1925, Online. Association for Computational Linguistics.
- Hanxi Guo, Siyuan Cheng, Xiaolong Jin, Zhuo Zhang, Kaiyuan Zhang, Guanhong Tao, Guangyu Shen, and Xiangyu Zhang. 2024a. [Biscope: Ai-generated text detection by checking memorization of preceding tokens](#). In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pages 104065–104090.
- Xun Guo, Shan Zhang, Yongxin He, Ting Zhang, Wanquan Feng, Haibin Huang, and Chongyang Ma. 2024b. [Detective: Detecting ai-generated text via multi-level contrastive learning](#). *Preprint*, arXiv:2410.20964.
- Hans W. A. Hanley and Zakir Durumeric. 2024. [Machine-made media: Monitoring the mobilization of machine-generated articles on misinformation and mainstream news websites](#). *Preprint*, arXiv:2305.09820.
- Alex Hanna, Remi Denton, Andrew Smart, and Jamila Smith-Loud. 2020. [Towards a critical race methodology in algorithmic fairness](#). In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, FAT* ’20*, page 501–512, New York, NY, USA. Association for Computing Machinery.
- Abhimanyu Hans, Avi Schwarzschild, Valeriia Cherepanova, Hamid Kazemi, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2024. [Spotting llms with binoculars: zero-shot detection of machine-generated text](#). In *Proceedings of the 41st International Conference on Machine Learning, ICML’24*. JMLR.org.

- Xiaomeng Hu, Pin-Yu Chen, and Tsung-Yi Ho. 2023. [Radar: Robust ai-text detection via adversarial learning](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 15077–15095. Curran Associates, Inc.
- Srinivasan Iyer, Xi Victoria Lin, Ramakanth Pasunuru, Todor Mihaylov, Daniel Simig, Ping Yu, Kurt Shuster, Tianlu Wang, Qing Liu, Punit Singh Koura, Xian Li, Brian O’Horo, Gabriel Pereyra, Jeff Wang, Christopher Dewan, Asli Celikyilmaz, Luke Zettlemoyer, and Ves Stoyanov. 2023. [Opt-impl: Scaling language model instruction meta learning through the lens of generalization](#). *Preprint*, arXiv:2212.12017.
- Yan Ju, Shu Hu, Shan Jia, George H. Chen, and Siwei Lyu. 2024. [Improving fairness in deepfake detection](#). In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4655–4665.
- Minseok Jung. 2025. [Responsible computational text generation: Ai content classification and policy framework](#). Master’s thesis, Massachusetts Institute of Technology, Cambridge, MA, February.
- Vinaya Sree Katamneni, Aakash Varma Nadimpalli, and Ajita Rattani. 2024. [Demographic fairness and accountability of audio- and video-based unimodal and bi-modal deepfake detectors](#). *Face Recognition Across the Imaging Spectrum*, pages 205–231.
- Ryuto Koike, Masahiro Kaneko, and Naoaki Okazaki. 2024. [Outfox: Llm-generated essay detection through in-context learning with adversarially generated examples](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(19):21258–21266.
- Jooyoung Lee, Xiaochen Zhu, Georgi Karadzhov, Tom Stafford, Andreas Vlachos, and Dongwon Lee. 2025. [Collaborative evaluation of deepfake text with deliberation-enhancing dialogue systems](#). *Preprint*, arXiv:2503.04945.
- Weixin Liang, Mert Yuksekgonul, Yining Mao, Eric Wu, and James Zou. 2023. [Gpt detectors are biased against non-native english writers](#). *Patterns*, 4(7):100779. Accessed: 2025-03-12.
- Decheng Liu, Zongqi Wang, Chunlei Peng, Nannan Wang, Ruimin Hu, and Xinbo Gao. 2025. [Thinking racial bias in fair forgery detection: Models, datasets and evaluations](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(5):5379–5387.
- Jesse G Meyer, Ryan J Urbanowicz, Patrick CN Martin, Karen O’Connor, Ruowang Li, Pei-Chen Peng, Tiffani J Bright, Nicholas Tatonetti, Kyoung Jae Won, Graciela Gonzalez-Hernandez, and Jason H Moore. 2023. [Chatgpt and large language models in academia: opportunities and challenges](#). *BioData Mining*, 16(1):20.
- Jaron Mink, Miranda Wei, Collins W. Munyendo, Kurt Hugenberg, Tadayoshi Kohno, Elissa M. Redmiles, and Gang Wang. 2024. [It’s trying too hard to look real: Deepfake moderation mistakes and identity-based bias](#). In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI ’24, New York, NY, USA. Association for Computing Machinery.
- Yikang Pan, Liangming Pan, Wenhui Chen, Preslav Nakov, Min-Yen Kan, and William Wang. 2023. [On the risk of misinformation pollution with large language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1389–1403, Singapore. Association for Computational Linguistics.
- Flor Miriam Plaza-del Arco, Amanda Cercas Curry, Alba Curry, Gavin Abercrombie, and Dirk Hovy. 2024. [Angry men, sad women: Large language models reflect gendered stereotypes in emotion attribution](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7682–7696, Bangkok, Thailand. Association for Computational Linguistics.
- Felipe Romero-Moreno. 2025. [Deepfake detection in generative ai: A legal framework proposal to protect human rights](#). *Computer Law & Security Review*, 58:106162.
- Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. [Gender bias in machine translation](#). *Transactions of the Association for Computational Linguistics*, 9:845–874.
- Karolina Stanczak and Isabelle Augenstein. 2021. [A survey on gender bias in natural language processing](#). *Preprint*, arXiv:2112.14168.
- Thinkst. 2023. [Meet zippy: A fast ai/llm text detector](#). Accessed: 12/03/2025.
- Vivek Verma, Eve Fleisig, Nicholas Tomlin, and Dan Klein. 2024. [Ghostbuster: Detecting text ghostwritten by large language models](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1702–1717, Mexico City, Mexico. Association for Computational Linguistics.
- Junchao Wu, Shu Yang, Runzhe Zhan, Yulin Yuan, Lidia Sam Chao, and Derek Fai Wong. 2025. [A survey on LLM-generated text detection: Necessity, methods, and future directions](#). *Computational Linguistics*, 51(1):275–338.

A Data

Statistics for the human-written dataset that was curated for our experiments are shown in Table 7.

B System Descriptions

B.1 Ghostbuster

We use the implementation provided at <https://github.com/vivek3141/ghostbuster>. This was

	Total	Gender		Race/Ethnicity		ELL		Economic Disadvantage	
		Male	Female	White	Nonwhite	No	Yes	No	Yes
PERSUADE v2.0	24695	12074	12621	11282	13413	22451	2244	11003	13692
ASAP v2.0	24728	12498	12230	9841	14887	20991	3737	7933	16795
ELLIPSE	6482	3636	2846	471	6011	0	6482	1974	4508
Combined (clean)	41743	21277	20466	15078	26665	31079	10664	18188	23555

Table 7: Counts for each attribute in our combined corpus. Note that the ELLIPSE corpus is designed to capture ELL speakers, and thus contains only that group.

modified to fix an issue where outdated OpenAI models were referenced; we use davinci-002 and babbage-002 models.

B.2 Glimpse

We use the implementation provided at <https://github.com/baoguangsheng/glimpse>.

B.3 Binoculars (Hans et al., 2024)

We use the implementation provided at <https://github.com/ahans30/Binoculars>.

B.4 Fast-DetectGPT (Bao et al., 2024)

We use two settings that use different models for scoring: gpt-neo-2.7b for speed and falcon-7b for maximal accuracy.

We use the implementation provided at <https://github.com/baoguangsheng/fast-detect-gpt>.

B.5 Zippy (Thinkst, 2023)

We use the implementation provided at <https://github.com/thinkst/zippy>. We experimented with the LZMA and ensemble versions, and found no significant differences in performance.

B.6 BiScope (Guo et al., 2024a)

We use the implementation provided at <https://github.com/MarkGHX/BiScope>: they do not provide an explicit "best" model for each domain, so we train each of our four variants using all the provided data from the respective domains.

B.7 DeTeCtivE (Guo et al., 2024b)

We use the implementation provided at <https://github.com/heyongxin233/DeTeCtive>.

B.8 RADAR (Hu et al., 2023)

We use the implementation provided at <https://github.com/IBM/RADAR>.

B.9 Desklib (Desklib, 2025)

We use the implementation provided at <https://github.com/desklib/ai-text-detector>.

B.10 E5-lora (Dugan et al., 2024)

We use the implementation provided at <https://github.com/menglinzhou/e5-small-lora-ai-generated-detector>.

The creators indicate the desired citation is for the RAID dataset.

B.11 Architecture/Costs

For model training, inference, and evaluation we use Amazon AWS EC2 instances. We use the g6e.xlarge instance type. This instance type has an NVIDIA L40S Tensor Core GPU with 48 GB of GPU memory, allowing us to experiment with models that have larger GPU memory requirements (notably Binoculars and the FDG systems require significant GPU memory).

Running all models over our dataset requires approximately 6 hours of machine time, costing approximately \$12 USD. We additionally spent approximately \$200 USD for OpenAI model usage, required for the Ghostbuster and Glimpse models.

C Correlations

Figure 2 shows a heatmap of correlations between model predictions.

D Prompts

We utilize the following prompt to interface with the language model (Claude 3.5). The prompt asks for completion of a given student essay. That essay is trimmed to the first 20 tokens, which are provided to the model with the instruction to complete the text. The model is instructed to limit the output to the length of the original essay, while mimicking the style of a student:

Here is the start of a student's essay. Complete the essay. It should be at

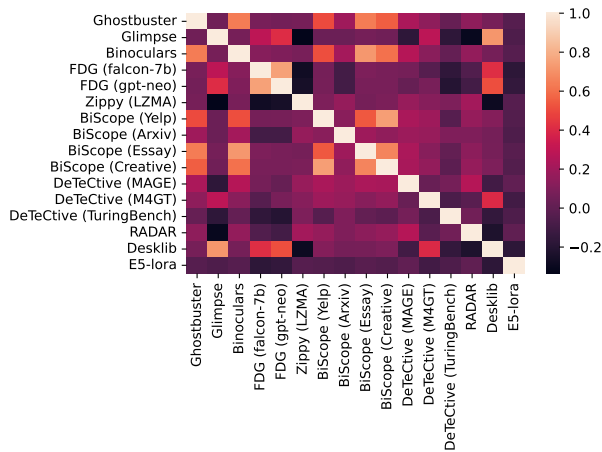


Figure 2: Pearson correlation for the predictions for each model.

annotations were completed. This prevented any type of bias in having authors perform annotation. After annotations were completed, the annotators were able to make significant contributions to the analysis and writing of the paper, and given the amount of work and the level of contribution they made, we believed the best way to credit their work and contribution was to include them as authors.

most `{len(text.split())}` words long.
 Do not go over this requirement.
 Emulate the style of a student between
 6th and 12th grade. You may include
 some common misspellings and
 punctuation errors, so that the text
 looks like a students.

Start:

```
{' '.join(text.split()[:20])}
```

Return only the resulting text as a
 json object:

```
{{\"text\": \"<generation>\"}}
```

Ensure the result is under
`{len(text.split())}` tokens.”

E Annotators

We recruited three annotators through personal requests for our experiments. These annotators are professionals with previous work in the field of deepfake detection, and consented to their results being used individually. All three annotators are post-graduate educated, fluent English speakers. Each annotator was given a batch of samples with the instruction to classify each sample as either human-written or machine-generated. Annotators were compensated as part of salaried work at a rate above minimum wage.

After completion of their individual tasks, we decided to involve the annotators as authors on the paper. Note that the annotators had no information about the project before their annotations were completed (which models, prompts, and datasets were used, what the goal of the project was, etc.): they were kept entirely independent until after their