A Simple yet Efficient Ensemble Approach for AI-generated Text Detection

Harika Abburi[1,*] **Kalyani Roy**[1], **Michael Suesserman**[2], **Nirmala Pudota**[1],
Balaji Veeramani[2], **Edward Bowen**[2], **Sanmitra Bhattacharya**[2]
[1]Deloitte & Touche Assurance & Enterprise Risk Services India Private Limited India
[2]Deloitte & Touche LLP, USA
{abharika, kalyroy, msuesserman, npudota, bveeramani, edbowen, sanmbhattacharya}@deloitte.com

Abstract

Recent Large Language Models (LLMs) have demonstrated remarkable capabilities in generating text that closely resembles human writing across wide range of styles and genres. However, such capabilities are prone to potential abuse, such as fake news generation, spam email creation, and misuse in academic assignments. Hence, it is essential to build automated approaches capable of distinguishing between artificially generated text and human-authored text. In this paper, we propose a simple yet efficient solution to this problem by ensembling predictions from multiple constituent LLMs. Compared to previous state-of-the-art approaches, which are perplexity-based or uses ensembles with a number of LLMs, our condensed ensembling approach uses only two constituent LLMs to achieve comparable performance. Experiments conducted on four benchmark datasets for generative text classification show performance improvements in the range of 0.5 to 100% compared to previous state-of-the-art approaches. We also study the influence that the training data from individual LLMs have on model performance. We found that substituting commercially-restrictive Generative Pre-trained Transformer (GPT) data with data generated from other open language models such as Falcon, Large Language Model Meta AI (LLaMA2), and Mosaic Pretrained Transformers (MPT) is a feasible alternative when developing generative text detectors. Furthermore, to demonstrate zero-shot generalization, we experimented with an English essays dataset, and results suggest that our ensembling approach can handle new data effectively.

1 Introduction

The domain of Natural Language Generation (NLG) is witnessing a remarkable transformation with the emergence of Large Language Models (LLMs) such as Generative Pre-trained Transformer (GPT-4) (OpenAI, 2023), Large Language

Model Meta AI (LLaMA-2) (Touvron et al., 2023), Pathways Language Model (PaLM) (Chowdhery et al., 2022), Bard[1], and Text-to-Text Transfer Transformer (T5) (Raffel et al., 2020). LLMs, characterized by their large parameter size, have shown state-of-the-art capabilities in generating text that closely mirrors the verbosity and style of human language. They have shown exceptional performance across a wide range of applications, such as story generation (Fan et al., 2018), Artificial Intelligence (AI)-assisted writing (Hutson, 2021), medical question answering (Kung et al., 2023), conversational response generation (Mousavi et al., 2023), radiology report generation (Mallio et al., 2023), and code auto-completion (Tang et al., 2023). Moreover, their capacity to generalize across tasks without the need for explicit training (referred to as zero-shot learning) or conditioning on only a few examples (referred to as few-shot learning) have substantially reduced the need for extensive, task-specific training efforts. These capabilities have significantly lowered the barrier of integrating LLMs into various language generation applications.

With the ability to generate coherent human-like text, the LLMs can also be misused for unethical purposes, such as fake news generation (Uchendu et al., 2021), phishing or spamming (Weiss, 2019), and fabrication of product reviews (Gambetti and Han, 2023). It has become increasingly crucial for both humans and automated systems to be able to detect and distinguish AI-generated text, particularly when this text is employed for disseminating misinformation or propaganda (Weidinger et al., 2021). To address these challenges, automatic detection of AI-generated text has recently become an active area of research.

Diverse modeling strategies, ranging from simple statistical techniques to cutting-edge Transformer-based architectures, have been investi-

---

*Corresponding author

[1]https://bard.google.com/

gated to develop solutions capable of distinguishing AI-generated text from those written by humans. Gehrmann et al. (2019) proposed straightforward statistical methods for identifying model-generated text that could be integrated into a visual tool to aid in their detection process. The authors assumed that AI systems produce text from a limited set of language patterns for which they have a high level of confidence. Wu et al. (2023) and Yang et al. (2023) explored entropy, n-gram frequency, and perplexity to distinguish between human-authored and AI-generated texts. Advanced deep-learning frameworks, such as Transformer-based models have also been explored to improve the precision and reliability of AI-generated text detection techniques. DetectGPT (Mitchell et al., 2023) focused on generating minor perturbations of a text passage using a generic pre-trained T5 model. It then compared the log probability of both the original text and the perturbed versions to determine if the text is authored by a human or generated by AI. Liu et al. (2022) proposed a Coherence-based Contrastive learning (CoCo) model where the input text is represented as a coherence graph to capture its entity consistency. Robustly optimized Bidirectional Encoder Representations from Transformers (BERT) approach (RoBERTa) embeddings are extracted and concatenated with sentence level graphical representations. In order to improve the model's performance, it is trained using a combination of contrastive loss and cross-entropy loss. Most recently, Abburi et al. (2023a,b) proposed an ensemble modeling approach for detecting AI-generated text where the probabilities from various constituent pre-trained LLMs are concatenated and passed as a feature vector to machine learning classifiers. The ensemble modeling approach resulted in improved predictions compared to what any individual classifier could achieve independently.

Although the primary purpose of AI-generated text detectors is to mitigate risks associated with harmful AI-generated content, erroneously classifying genuine, human-authored work as AI-generated can, conversely, lead to significant harm. Recently, there has been growing apprehension regarding the accuracy and reliability of these generative AI text detectors (Liang et al., 2023; Sadasivan et al., 2023; He et al., 2023). Liang et al. (2023) highlighted potential bias observed with several GPT detectors. The authors showed that a majority of existing detectors incorrectly classified English

writing samples from non-native English speakers as AI-generated. Surprisingly, altering language created by non-native speakers with prompts like *"Enhance it to sound more like that of a native speaker"* resulted in a significant drop in misclassification. This highlights that a majority of detectors prioritize low perplexity as a primary criterion for identifying text as AI-generated. Since the potential bias in detectors is tied to perplexity scores, the authors propose a more robust and equitable redesign of these detectors. In addition, they propose thorough evaluation of these detectors that takes other important metrics such as bias and fairness into consideration.

In this paper, we extend the work of Abburi et al. (2023a,b) by proposing an architecture that is simpler in design, while maintaining model performance. We validate the effectiveness of our model by benchmarking it on various publicly available datasets, including the Automated Text Identification (AuTexTification) (Sarvazyan et al., 2023) dataset. We also study how inclusion of data generated by various LLMs in the training corpus affects the model performance and generalizability. In order to examine if our approach suffers from similar drawbacks and biases as other perplexity-based approaches, we evaluate the zero-shot performance of our trained model on the aforementioned English essays dataset (Liang et al., 2023) and report the corresponding results.

In summary, our key contributions in this paper are: 1) we propose a simpler non-perplexity based AI-text detector model that extends prior work 2) we demonstrate the robustness of our approach across multiple benchmark datasets, including the one that examines potential biases in model predictions, 3) we analyze the influence that training data from individual LLMs have on model performance and 4) we find that excluding GPT data from training sets improves the accuracy of detecting human-authored samples.

2 Datasets

In this section, we provide a brief description of various publicly available benchmark datasets for AI-generated text detection. We also describe a number of datasets that we crafted and used in our experiments (henceforth referred to as curated datasets). Table 1 shows the number of human-authored (Human) and AI-generated (AI) samples available for train and test splits of each dataset.

	Dataset	*Train*		*Test*	
		Human	*AI*	*Human*	*AI*
Benchmark Datasets	AuText	17,046	16,799	10,642	11,190
	AA	213	1,706	853	6,822
	TB	5,964	10,6240	1,915	35,442
	AP	82	78	18	12
	GPT-OD	250,000	250,000	5,000	5,000
	EWEssays	–	–	394	352
Curated Datasets	D1	17,046	8,263	10,642	11,190
	D2	17,046	16,799	10,642	11,190
	D3	17,046	16,799	10,642	11,190

Table 1: Dataset statistics

2.1 Publicly available benchmark datasets

For the first set of experiments, which demonstrate the robustness of our approach, we use multiple benchmark datasets described below.

2.1.1 AuTexTification (AuText):

The AuText dataset (Sarvazyan et al., 2023) consists of human-authored and AI-generated texts from five domains, where three domains (legal, wiki, tweets) are represented in the training corpus, and two different domains (reviews, news) are represented in the testing corpus. The generated text is created using six LLMs of varying parameter sizes ranging from 2B to 175B. Three of them are BigScience Large Open-science Open-access Multilingual Language Model (BLOOM) models and the other three are GPT variants: (i). bloom-1b7 (A), (ii). bloom-3b (B), (iii). bloom-7b1 (C), (iv). babbage (D), (v). curie (E), and (vi). text-davinci-003 (F).

2.1.2 Author Attribution (AA):

The AA dataset (Uchendu et al., 2020) consists of nine categories: human and eight LLMs generated texts. Political news articles from CNN, New York Times, and Washington Post represent the human-authored text. The titles of these news articles written by human journalists are used as the prompts to generate the AI-generated text from eight LLMs such as Conditional Transformer Language Model (CTRL), Cross-Lingual Language Model (XLM), eXtreme Multi-Label Multi-Task Learning with a Language Model (XLNet), GPT, GPT2, Grover, Meta's Fair, and Plug and Play Language Model (PPLM).

2.1.3 Turing Bench (TB):

The TB dataset (Uchendu et al., 2021) is created by gathering around 10k news articles written by journalists in various media channels. The title of each article is used as a prompt to generate the text from 19 LLMs, such as GPT, GPT2, GPT3, PPLM, Transformer-XL, XLM, XLNet, and various versions of these models. After preprocessing, the dataset comprises 168,612 articles with around 8k samples in each LLM category, including human-authored.

2.1.4 Academic Publications (AP):

The AP dataset (Liyanage et al., 2022) is composed of 100 papers selected from ArXiv in computation and language domain and labeled as human-authored. GPT-2 is used to generate the 100 equivalent research papers and labeled as AI-generated. From both human-authored and GPT2-generated text, the sections such as methodology, results, and discussion which contain diagrams, tables, equations are ignored.

2.1.5 Gpt-2-Output-Dataset (GPT-OD):

The GPT-OD (Radford et al., 2019) dataset contains data from WebText test set as well as samples generated by four GPT-2 variants (with parameters 117M, 345M, 762M, and 1542M) trained on the WebText training set. More details about the dataset can be obtained here [2]. In this study, we consider 255k samples from the WebText test set as human-authored and 255k samples generated using the GPT-2 XL-1542M model (temperature 1, no truncation) as AI-generated samples.

2.1.6 Essays from native and non-native English writers:

This dataset is primarily comprised of essays authored by native and non-native English speakers Liang et al. (2023). US 8-th grade student essays represent essays authored by native English speakers, while Test of English as a Foreign Language (TOEFL) essays obtained from a Chinese educational forum represent essays authored by non-native English speakers. ChatGPT 3.5 with simple prompts was used on the aforementioned essays, as well as Stanford CS224n final project

---

[2]https://github.com/openai/gpt-2-output-dataset

abstracts and US Common App college admission essays, to generate artificial essays. We refer to this dataset as *EWEssays* hereafter. In this paper, we evaluate the performance of our model on this dataset using a zero-shot approach, utilizing the complete dataset for inference.

Model	Train	Test
human-authored	17046	10642
bloom-1b7 (A)	2,750	1,704
bloom-3b (B)	2,705	1,782
bloom-7b1 (C)	2,808	1,831
babbage (D)	2,834	1,960
curie (E)	2,843	1,958
text-davinci-003 (F)	2,859	1,955

Table 2: AuText dataset statistics

2.2 Curated training datasets

We created a number of curated datasets motivated by the following factors: 1) demonstrate the influence of training data from individual LLMs on model performance, 2) explore whether model performance is affected in out-of-domain testing, i.e., the model is tested on a dataset generated by a LLM that is not used in training data creation, and 3) specifically, analyze whether a model trained without GPT data can achieve similar performance to a model trained using GPT data, which is subject to specific restrictions regarding commercial usage. While we focus primarily on the AuText dataset to derive these curated datasets, our analysis is broadly applicable to other datasets mentioned in Section 2.1.

The distribution of train and test splits for both human-authored and AI-generated data in the AuText dataset are shown in Table 2. Around half of the AI-generated data is produced by BLOOM-based models (Scao et al., 2022), while the rest are generated by GPT-based models. Given the restrictions on commercial usage of data generated by GPT-based models[3], we wanted to explore whether replacing GPT data with data from other recent open LLMs (LLaMA2[4], Falcon[5], and MPT[6]) is a feasible alternative for training generative text detectors. We selected LLaMA2, because it outperformed other open LLMs on various external benchmarks, including reading comprehension,

reasoning, coding, and knowledge tests (Touvron et al., 2023). The LLaMA2 chat models have additionally been trained on over 1 million human annotations compared to its previous version. Prior to LLaMA2, Falcon and MPT were outperforming other open LLMs on the open LLM leaderboard[7].

Using the three selected open LLMs, we created the following variants of the AuText dataset:
1. In the first variant (D1), we removed all GPT-based data (categories D, E, and F in Table 2) from the AuText training data.
2. In the second variant (D2), we replaced the training data from the GPT-based models (categories D, E, and F) with that from LLaMA2-13b-chat model. The prompts we used with the LLaMA2-13b-chat model were the same ones used by the developers of the AuText dataset.
3. In the third variant (D3), we substituted training data from categories D, E, and F (Table 2) with data generated from Falcon-40b-chat, MPT-30b-instruct, and LLaMA2-13b-chat, respectively. As before, we used the same prompts as those used by the developers of the AuText dataset.

In the D1 dataset, the number of training samples are reduced from 33845 to 25309 as we removed GPT-based data. In the D2 and D3 datasets, the number of samples in training data is same as the AuText training samples since we just replaced the GPT samples with the same number of samples generated using open LLMs. In all these curated training datasets, text from human-authored and BLOOM-based models (categories A-C in Table 2) remain unchanged. No changes were made to the test datasets from AuText.

3 Approach

We used an ensemble modeling approach similar to the one proposed by (Abburi et al., 2023a,b), where each input is passed through five pre-trained models, namely: 1. Decoding-enhanced BERT with disentangled attention (DeBERTa) large[8] (He et al., 2021), 2. cross-lingual language model RoBERTa (XLM-RoBERTa) with Cross-lingual Natural Language Inference (XNLI)[9], 3. RoBERTa large[10] (Liu et al., 2019), 4. RoBERTa base OpenAI de-

Dataset	Baseline model	Acc	F_{macro}	Pre	Rec
AA	RoBERTa-base (Uchendu et al., 2020)	0.970	0.923	0.932	0.914
TB	RoBERTa-large-MNLI (Uchendu et al., 2021)	0.997	0.985	0.976	0.995
AP	DistilBERT (Liyanage et al., 2022)	0.250	0.242	0.242	0.242
GPT-OD	COCO (Liu et al., 2022)	0.943	0.941	–	–

Table 3: Performance on various benchmark datasets with state-of-the-art models.

Dataset	Acc		F_{macro}		Pre		Rec	
	E	SE	E	SE	E	SE	E	SE
AA	0.994 (+2.5%)	0.990 (+2.1%)	0.986 (+6.8%)	0.975 (+5.6%)	0.993 (+6.5%)	0.988 (+6.0%)	0.979 (+7.1%)	0.962) (+5.2%)
TB	0.998 (+0.1%)	0.998 (+0.1%)	0.990 (+0.5%)	0.989 (+0.4%)	0.997 (+2.2%)	0.993 (+1.7%)	0.983 (-1.2%)	0.986 (-0.9%)
AP	0.500 (+100.0%)	0.475 (+90.0%)	0.479 (+97.9%)	0.475 (+96.3%)	0.484 (+100.0%)	0.475 (+96.3%)	0.485 (+100.4%)	0.475 (+96.3%)
GPT-OD	0.990 (+5.0%)	0.983 (+4.2%)	0.989 (+5.1%)	0.983 (+4.4%)	0.990 –	0.983 –	0.990 –	0.983 –

Table 4: Performance of *Ensemble* (E) and *Short Ensemble* (SE) models on four datasets. Numbers in the parenthesis indicate percentage changes compared to baselines.

tector[11] (Solaiman et al., 2019), and 5. XLM-RoBERTa NLI[13] (Reimers and Gurevych, 2019). In the training phase, these models are fine-tuned using the training data for each dataset shown in Table 1 (except EWEssays). For inference and testing, each model independently generates classification probabilities. In order to maximize the advantage of each model, each of these probabilities are concatenated to create a feature vector and passed as an input to train a voting classifier (Logistic Regression (LR), Random Forest (RF), Gaussian Naive Bayes (NB), Support Vector machines (SVM) (Mahabub, 2020)) to produce final predictions. Hereafter, we refer to this architecture as *Ensemble*.

In addition to experimenting with ensembling five models proposed by (Abburi et al., 2023a,b) , we also conducted experiments with various combinations of these models using the same architecture. We observed that an ensemble of only the RoBERTa base OpenAI detector and the XLM-RoBERTa NLI model, along with the voting classifier, delivers performance comparable to the *Ensemble* architecture. Henceforth, we refer to this architecture as *Short Ensemble*. For both architectures, the experimental setup and hyperparameter choices are similar to Abburi et al. (2023b).

---

[11]A finetuned sequence classifier based on RoBERTa-base (125 million parameters)[12] and RoBERTa-large (356 million parameters)

[13]https://huggingface.co/sentence-transformers/xlm-r-100langs-bert-base-nli-stsb-mean-tokens

4 Experiments

In this section, we present an evaluation of our AI-generated text detection experiments. Results are presented for multiple models on both benchmark and curated datasets. Results from a zero-shot evaluation on the EWEssays dataset is also presented. Traditional classification metrics, namely, accuracy (Acc), macro F1 score (F_{macro}), precision ($Prec$), and recall (Rec) are reported for each of the experiments.

4.1 Performance of proposed architectures across various benchmark datasets

As baselines, we use four Transformer-based architectures, which, to our knowledge, are the current state-of-the-art models on the corresponding benchmark datasets: 1. AA dataset: RoBERTa-base (Uchendu et al., 2020), 2. TB dataset: RoBERTa-large-MNLI (Uchendu et al., 2021), 3. AP dataset: DistilBERT (Liyanage et al., 2022), and 4. GPT-OD dataset: COCO (Liu et al., 2022). RoBERTa-large-MNLI is a RoBERTa-large model fine-tuned on the Multi-Genre Natural Language Inference (MNLI) corpus. The DistilBERT model uses knowledge distillation during the pre-training phase. Both RoBERTa and DistilBERT were fine-tuned for this experiment. COCO is a coherence-based contrastive learning model that detects AI-generated texts in low-resource settings.

Table 3 shows the results of baseline models on four publicly available benchmark datasets. The results from the *Ensemble* and *Short Ensemble*

models on the same four benchmark datasets are shown in Table 4. The results show both the *Ensemble* and *Short Ensemble* architectures perform well when compared to the other architectures across datasets. In both ensemble architectures the RoBERTa base OpenAI detector model shows stronger performance compared to other constituent models. When compared to the state-of-the-art, we find that our models deliver performance improvements in the range of 0.5-97.9% for F_{macro}, across the benchmark datasets. The highest improvement is on the AP dataset (97.9%), followed by the AA dataset (6.8%). Notably, the results from the *Short Ensemble* model closely approximate those from the *Ensemble* architecture. This indicates that an ensemble of just two models, as seen in the *Short Ensemble* architecture is adequate to achieve state-of-the-art performance. Importantly, this approach significantly simplifies the model's complexity compared to the larger ensemble model. As a result, we choose the *Short Ensemble* architecture for presenting the remaining results in the paper.

4.2 Analysis of model performance on curated training datasets

Each of the curated training datasets (D1, D2 and D3) are variants of the AuText dataset, and comprises data from different combinations of LLMs. Table 5 illustrates the influence these different datasets have on model performance compared to the unmodified AuText training data. The *Short Ensemble* architecture is trained on each of the curated datasets independently and evaluated on the AuText test set. Based on the results from D1, D2, and D3, it is evident that, despite the absence of text generated by GPT models in the training data, the *Short Ensemble* model is able to effectively detect GPT-generated text. In D1, by simply removing the GPT text from the training data the F_{mac} score improved to 0.769 from its baseline F_{mac} 0.732. The model performance is further improved with a F_{mac} score of 0.774 when GPT text is replaced with LLaMA2 data. An interesting observation we made regarding the D3 dataset (which uses Falcon, MPT, and LLaMA2 data) is that, even though the model demonstrated improved performance compared to AuText, it did not outperform the results achieved by exclusively utilizing text from LLaMA2. Across these metrics, the *Short Ensemble* model, fine-tuned on D2 dataset with

LLaMA2 data, outperformed those trained on the AuText data and other variants.

Dataset	Acc	F_{macro}	Pre	Rec
AuText	0.750	0.732	0.822	0.744
D1	0.775	0.769	0.796	0.771
D2	**0.784**	**0.774**	**0.828**	**0.779**
D3	0.760	0.747	0.812	0.755

Table 5: Result on AuText test data with *Short Ensemble* model trained on four different training sets.
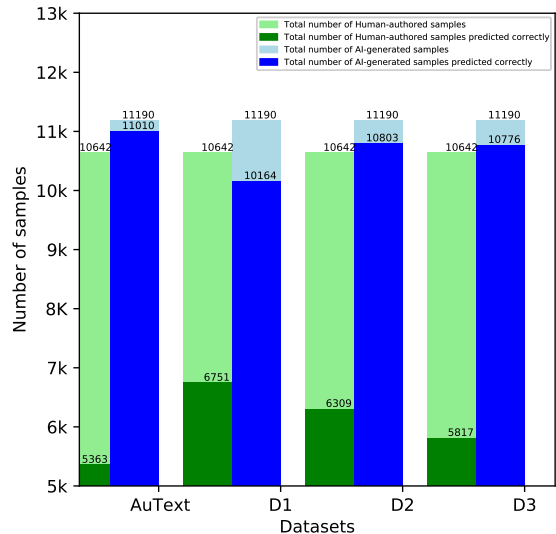

Figure 1: Performance of *Short Ensemble* model on AuText and our curated datasets.

Figure 1 presents the distribution of samples within human-authored and AI-generated classes, along with the number of correct predictions across the four datasets. The results indicate that the model trained on the AuText data correctly predicts the highest number of AI-generated samples, followed by D2, D3, and D1, with only minor differences between them. In the case of human-authored class predictions, D1 showed higher performance followed by D2, D3 and AuText. Interestingly, the model not trained using GPT-generated text, i.e., D1, performed better in detecting human-authored text. We note, however, further investigation is required to understand why certain combinations of LLM training data underperform others.

Overall, the three motivating factors behind the creation of these curated datasets (outlined in Section 2.2) were addressed with these experiments. Our experiments demonstrate that using recent open LLMs over commercially-restrictive GPT-based data is a feasible alternative in developing

generative text detectors.

Dataset	Acc	F_{macro}	Pre	Rec
AuText	0.684	0.683	0.694	**0.689**
D1	**0.693**	**0.690**	0.693	**0.689**
D2	0.655	0.639	0.720	0.670
D3	0.633	0.601	**0.744**	0.651

Table 6: Zero-shot results on EWEssays dataset with *Short Ensemble* approach.
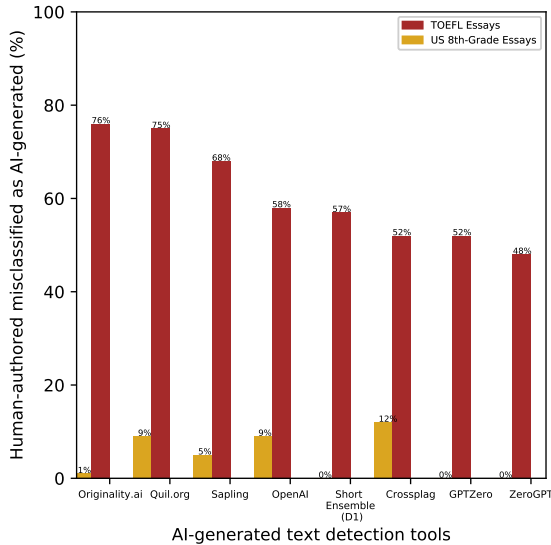

Figure 2: Performance of AI-generated text detection tools along with *Short Ensemble* (D1) model on EWEssays dataset (TOEFL and US 8th grade essays categories)

4.3 Zero-shot generalization

To assess the robustness and generalizability of our *Short Ensemble* model, we tested the models outlined in Table 5 on the EWEssays dataset in a zero-shot setting. The results are shown in Table 6. We find that D1 outperforms other models achieving F_{mac} score of 0.69. Furthermore, we performed an analysis to assess the precision of our model in detecting two distinct human-authored classes: US 8-th grade essays, and TOEFL essays, similar to Li et al. (2023). Figure 2 depicts the performance of various generative text detectors along with *Short Ensemble* (D1). Along with GPTZero[14] and ZeroGPT[15], our *Short Ensemble* model with D1 dataset also classified all the US 8-th grade essays correctly as human-authored, whereas it misclassified 57.14% of TOEFL essays as AI-generated,

---

[14] https://gptzero.me/
[15] https://www.zerogpt.com/

achieving overall accuracy of 42.86%. The performance of our model in the zero-shot setting is not as promising, and highlights the need for further improvements in terms of bias and fairness evaluation. Nevertheless, we note that our model outperformed Originality.ai, Sapling.ai, Quil.org, and OpenAI text detectors as shown in Figure 2.

5 Conclusion

In this research, we proposed a simple yet effective *Short Ensemble* model for distinguishing between AI-generated and human written text. We investigated the robustness of our proposed model across various benchmark datasets and observed that our model performs better compared to several state-of-the-art baselines. In addition, we crafted a set of datasets using open LLMs and examined their impact on model performance. Our study shows that fine-tuning models with text generated from open LLMs performs comparable or better when compared to models fine-tuned on GPT-generated text. Furthermore, we investigated the zero-shot generalization capabilities of our model on the EWEssays dataset. We observed that our model outperformed several text detection tools in correctly classifying English essays authored by non-native English writers. However, it is important to note that our model with the highest accuracy on EWEssays achieved a score of 42.8%, emphasizing the need for ongoing efforts to enhance both generalization and robustness in our approach. We also acknowledge that our approach should further be evaluated on detecting text generated from more advanced LLMs (LLMs with more than 175B parameters such as GPT-4 (OpenAI, 2023)).

References

Harika Abburi, Michael Suesserman, Nirmala Pudota, Balaji Veeramani, Edward Bowen, and Sanmitra Bhattacharya. 2023a. An ensemble-based approach for generative language model attribution. In *International Conference on Web Information Systems Engineering*, pages 699–709. Springer.

Harika Abburi, Michael Suesserman, Nirmala Pudota, Balaji Veeramani, Edward Bowen, and Sanmitra Bhattacharya. 2023b. Generative ai text classification using ensemble llm approaches. In *IberLEF@ SEPLN*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling

language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. *arXiv preprint arXiv:1805.04833*.

Alessandro Gambetti and Qiwei Han. 2023. Combat ai with ai: Counteract machine-generated fake restaurant reviews on social media.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander M Rush. 2019. Gltr: Statistical detection and visualization of generated text. *arXiv preprint arXiv:1906.04043*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Xinlei He, Xinyue Shen, Zeyuan Chen, Michael Backes, and Yang Zhang. 2023. Mgtbench: Benchmarking machine-generated text detection. *arXiv preprint arXiv:2303.14822*.

Matthew Hutson. 2021. Robo-writers: the rise and risks of language-generating ai. *Nature*, 591(7848):22–25.

Tiffany H Kung, Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño, Maria Madriaga, Rimel Aggabao, Giezel Diaz-Candido, James Maningo, et al. 2023. Performance of chatgpt on usmle: Potential for ai-assisted medical education using large language models. *PLoS digital health*, 2(2):e0000198.

Hanzhou Li, John T Moon, Saptarshi Purkayastha, Leo Anthony Celi, Hari Trivedi, and Judy W Gichoya. 2023. Ethics of large language models in medicine and medical research. *The Lancet Digital Health*.

Weixin Liang, Mert Yuksekgonul, Yining Mao, Eric Wu, and James Zou. 2023. Gpt detectors are biased against non-native english writers. *arXiv preprint arXiv:2304.02819*.

Xiaoming Liu, Zhaohan Zhang, Yichen Wang, Yu Lan, and Chao Shen. 2022. Coco: Coherence-enhanced machine-generated text detection under data limitation with contrastive learning. *arXiv preprint arXiv:2212.10341*.

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

Vijini Liyanage, Davide Buscaldi, and Adeline Nazarenko. 2022. A benchmark corpus for the detection of automatically generated text in academic publications. *arXiv preprint arXiv:2202.02013*.

Atik Mahabub. 2020. A robust technique of fake news detection using ensemble voting classifier and comparison with other classifiers. *SN Applied Sciences*, 2(4):525.

Carlo A Mallio, Andrea C Sertorio, Caterina Bernetti, and Bruno Beomonte Zobel. 2023. Large language models for structured reporting in radiology: performance of gpt-4, chatgpt-3.5, perplexity and bing. *La radiologia medica*, pages 1–5.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature. *arXiv preprint arXiv:2301.11305*.

Seyed Mahed Mousavi, Simone Caldarella, and Giuseppe Riccardi. 2023. Response generation in longitudinal dialogues: Which knowledge representation helps?

OpenAI. 2023. Gpt-4 technical report.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

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*. Association for Computational Linguistics.

Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. 2023. Can ai-generated text be reliably detected? *arXiv preprint arXiv:2303.11156*.

Areg Mikael Sarvazyan, José Ángel González, Marc Franco Salvador, Francisco Rangel, Berta Chulvi, and Paolo Rosso. 2023. Autextification: Automatic text identification. In *Procesamiento del Lenguaje Natural*, Jaén, Spain.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, et al. 2019. Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203*.

Ze Tang, Jidong Ge, Shangqing Liu, Tingwei Zhu, Tong-tong Xu, Liguo Huang, and Bin Luo. 2023. Domain adaptive code completion via language models and decoupled domain databases. *arXiv preprint arXiv:2308.09313*.

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

Adaku Uchendu, Thai Le, Kai Shu, and Dongwon Lee. 2020. Authorship attribution for neural text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8384–8395.

Adaku Uchendu, Zeyu Ma, Thai Le, Rui Zhang, and Dongwon Lee. 2021. Turingbench: A benchmark environment for turing test in the age of neural text generation. *arXiv preprint arXiv:2109.13296*.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. 2021. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.

Max Weiss. 2019. Deepfake bot submissions to federal public comment websites cannot be distinguished from human submissions. *Technology Science*, 2019121801.

Kangxi Wu, Liang Pang, Huawei Shen, Xueqi Cheng, and Tat-Seng Chua. 2023. Llmdet: A large language models detection tool.

Xianjun Yang, Wei Cheng, Linda Petzold, William Yang Wang, and Haifeng Chen. 2023. Dna-gpt: Divergent n-gram analysis for training-free detection of gpt-generated text.