AAIG at GenAI Detection Task 1: Exploring Syntactically-Aware, Resource-Efficient Small Autoregressive Decoders for AI Content Detection

Avanti Bhandarkar, Ronald Wilson and Damon Woodard
Florida Institute for National Security
University of Florida, Gainesville, FL, USA
{avantibhandarkar, ronaldwilson, dwoodard}@ufl.edu

Abstract

This paper presents a lightweight and efficient approach to AI-generated content detection using small autoregressive fine-tuned decoders (AFDs) for secure, on-device deployment. Motivated by resource-efficiency, syntactic awareness, and bias mitigation, our model employs small language models (SLMs) with autoregressive pre-training and loss fusion to accurately distinguish between human and AI-generated content while significantly reducing computational demands. The system achieved highest macro-F1 score of 0.8186, with the submitted model scoring 0.7874–both significantly outperforming the task baseline while reducing model parameters by ≈60%. Notably, our approach mitigates biases, improving recall for human-authored text by over 60%. Ranking 8th out of 36 participants, these results confirm the feasibility and competitiveness of small AFDs in challenging, adversarial settings, making them ideal for privacy-preserving, on-device deployment suitable for real-world applications.

1 Introduction

Advancements in Generative AI (GenAI) powered by large language models (LLMs) have significantly improved natural language generation (NLG) capabilities. AI-generated text, often indistinguishable from human writing, presents risks to information integrity, trust, and security (Gehrmann et al., 2019; Ippolito et al., 2020; Wu et al., 2024). The widespread availability of open-source, user-friendly LLMs enables individuals with minimal expertise to conduct misinformation, disinformation, and phishing campaigns, highlighting the need for accurate AI content detection. However, most research rallies behind sophisticated, resource-intensive solutions, often overlooking the security and privacy aspect of AI content detection. Most solutions require cloud connectivity or extensive computational resources, making them impractical for secure, on-device deployment.

Many existing approaches depend on complex architectures, including large pre-trained models (PLMs) like RoBERTa-large and Longformer (Li et al., 2024), or leverage larger LLMs like LLaMA (Hans et al., 2024) through techniques such as instruction-tuning (Wang et al., 2024a). Others employ ensemble methods that combine multiple LLMs (Sheykhlan et al., 2024; Abburi et al., 2023; Lai et al., 2024; El-Sayed and Nasr, 2023; Sarvazyan et al., 2024), that can significantly increase latency. In contrast, we focus on lightweight models that are optimized for secure, on-device deployment. On-device processing supports real-time analysis, essential for fast-paced environments, while also enhancing privacy by retaining sensitive data locally and reducing risks associated with transmitting information to external servers (Xu et al., 2024). This is particularly important for security-sensitive fields such as defense, healthcare, finance, and personal communications, where protecting data from unauthorized access is critical. This study evaluates the feasibility of small autoregressive fine-tuned decoders (AFDs) for efficient and secure AI-generated content detection.

Our approach is guided by three motivations: First, using SLMs ($\leq$135M parameters) suitable for on-device deployment, enabling privacy, real-time processing, and accessibility; Second, leveraging an autoregressive pre-training objective that mirrors the sequential nature of language production, enhancing syntactic awareness essential for detecting structural nuances to differentiate human and machine language; and third, employing a loss fusion strategy to learn from difficult examples and encourage class separation for bias mitigation.

The proposed system, using SmolLM, achieved the highest macro-F1 score of 0.8186 on the test set, significantly outperforming the task baseline (macro-F1 of 0.7568) under similar settings, while also reducing the number of parameters by ≈60%. Our submitted model (selected based on validation

Proceedings of the 1st Workshop on GenAI Content Detection (GenAIDetect), pages 218–224

performance) attained a macro-F1 score of 0.7874, ranking 8th among 36 participants. These results underscore the potential of small AFDs for effective on-device AI-content detection.

2 Task Description

In the COLING Workshop on Detection AI Generated Content, Task 1 posed a binary classification problem: determining whether a given text is machine- or human-authored (Wang et al., 2025). Our investigation focused on Subtask A, which targets English-only MGT detection and extends the SemEval Shared Task 8 (Subtask A)(Wang et al., 2024b). Table 1 shows summary statistics for the provided datasets. From our analysis, this task presented three key challenges to test the model generalizability involving unfamiliar data sources, unknown LLMs as well as adversarially modified texts from Mixset (Zhang et al., 2024), LLM-DetectAIve (Abassy et al., 2024), and CUDRT (Tao et al., 2024).

Table 1: Summary statistics of shared task subsets

Property		Train	Val	Dev	Test
#Sources		3	3	2	6
#LLMs		40	40	5	14
Human	#Samples	228,922	98,328	13,371	34,675
	Avg Len	302	303	339	270
Machine	#Samples	381,845	163,430	19,186	39,266
	Avg Len	273	272	417	411

3 System Overview

Our approach centers on fine-tuning small autoregressive decoders combined with a loss fusion strategy to enhance classification performance. Anticipating the presence of surprise domains and LLMs in the test set, we experimented with two distinct loss functions to optimize performance, particularly on challenging examples.

3.1 Model Selection

The selection of AFDs is driven by two core reasons: bias mitigation and syntactic awareness.

Related research indicates that models like RoBERTa, though powerful, tend to display a bias toward synthetic text, resulting in a high rate of false negatives when classifying human-generated text (Ciccarelli et al., 2024). This suggests that machine-generated content has an identifiable structural pattern that models such as RoBERTa -which are not specifically optimized for language generation- might misinterpret as "non-human". In contrast, autoregressive language models are trained with a next-token prediction objective, which naturally aligns with human language composition, making them more attuned to syntactic patterns typical of human writing. This syntactic awareness is particularly valuable for distinguishing subtle linguistic cues that differentiate human from machine-generated text. Additionally, the "LLM race" has led to the development of increasingly compact SLMs, such as MobileLLM, SmolLM, and GPT-Neo, that achieve high performance on various text generation and reasoning tasks with fewer parameters, enabling efficient, on-device deployment (Xu et al., 2024). These qualities make small AFDs effective and practical for real-world AI-content detection.

3.2 Loss Fusion

As an additional measure for bias mitigation, we employ a loss fusion strategy. We experiment with two primary loss functions: **Cross-Entropy** (Mao et al., 2023), which minimizes classification errors by measuring the divergence between predicted probabilities and true labels, and **Focal Loss** (Mukhoti et al., 2020), which addresses class imbalance by penalizing misclassifications on harder-to-classify examples or ambiguous cases. Additionally, following Ai et al. (2022), we incorporate **Contrastive Loss** as an auxiliary loss that structures the embedding space by pulling similar samples closer and pushing dissimilar ones apart (Dipta and Shahriar, 2024). The final loss is a linear combination of the primary and auxiliary losses.

To evaluate the impact of the auxiliary loss, we also conduct ablation-like experiments using primary loss alone. Thus, our experiments include four different loss configurations: two primary losses (cross-entropy and focal) applied alone and also combined with contrastive loss.

4 Experimental Setup

We evaluate five different SLMs ($\leq$135M parameters), selected for their suitability for on-device deployment (Lu et al., 2024). These include SmolLM (Allal et al., 2024), GPT2 (Radford et al., 2019), GPT-Neo (Black et al., 2021), OPT (Zhang et al., 2022) and MobileLLM (Liu et al., 2024). Since the task baseline (RoBERTa-Large) has significantly more parameters than our selected models, we also include RoBERTa-Base as a baseline for fair per-

Table 2: Summary of results: Best-performing systems per phase are **highlighted**, and second-best underlined.

	System Details			Macro-F1				Test-Recall	
	Model	#Param.	Loss	Val	Dev	Test	%Gain	Human	LLM
Task Baseline	Roberta-L[3]	406M	CE[1]	0.9489	0.8017	0.7139	2.20↑	0.4951	0.9465
			CE + Con[2]	0.9473	0.7823	0.7475	5.57↑	0.5437	0.9581
			Focal	0.9699	0.8677	0.7568	4.94↑	0.5863	0.9293
			Focal + Con	0.9611	0.8122	0.7516	1.59↑	0.5697	0.9371
Our Baseline	Roberta-B[4]	125M	CE	0.9362	0.7898	0.6919	0.00	0.4507	0.9558
			CE + Con	0.9324	0.7653	0.6918	0.00	0.4540	0.9514
			Focal	0.9408	0.7795	0.7074	0.00	0.4675	0.9670
			Focal + Con	0.9489	0.8166	0.7358	0.00	0.5455	0.9323
Proposed System	SmolLM	135M	CE	0.9683	0.8642	0.8104	11.58↑	0.7218	0.8953
			CE + Con	**0.9793**	0.8679	0.7874*	9.56↑	0.6669	0.9051
			Focal	0.9627	0.8678	**0.8186**	11.12↑	0.7509	0.8830
			Focal + Con	0.9690	**0.8777**	0.8135	7.78↑	0.7281	0.8953
System Alternatives	GPT2	117M	CE	0.9354	0.7904	0.7085	1.67↑	0.4765	0.9581
			CE + Con	0.9361	0.7889	0.6670	2.48↓	0.4092	0.9574
			Focal	0.9235	0.7754	0.6867	2.07↓	0.4369	0.9626
			Focal + Con	0.9580	0.8210	0.7314	0.43↓	0.5420	0.9274
	GPT-Neo	125M	CE	0.9665	0.8062	0.7886	9.68↑	0.7603	0.8160
			CE + Con	0.9461	0.8240	0.7464	5.46↑	0.6832	0.8079
			Focal	0.9583	0.8166	0.7812	7.38↑	0.7401	0.8207
			Focal + Con	0.9656	0.8277	0.8070	7.13↑	0.7935	0.8204
	OPT	125M	CE	0.9647	0.8121	0.7024	1.05↑	0.5252	0.8872
			CE + Con	0.9622	0.7946	0.7115	1.97↑	0.5906	0.8332
			Focal	0.9588	0.8346	0.7243	1.69↑	0.5262	0.9311
			Focal + Con	0.9649	0.8400	0.7041	3.17↓	0.5614	0.8498
	MobileLLM	125M	CE	0.9608	0.8187	0.7225	3.06↑	0.6007	0.8446
			CE + Con	0.9593	0.8131	0.7164	2.46↑	0.6002	0.8328
			Focal	0.9620	0.8187	0.7276	2.02↑	0.6067	0.8485
			Focal + Con	0.9622	0.8246	0.7078	2.80↓	0.6121	0.8030

Abbrev: [1] Cross-Entropy Loss; [2] Contrastive Loss; [3] Roberta-Large; [4] Roberta-Base. Note: %Gain represents the performance improvement over our baseline (RoBERTa-base), with arrows (↑/↓) indicating increase or decrease. Test performance of the submitted system is marked with (*).

formance comparison. To align with our objective of testing the feasibility of small AFDs, we adopt a simple architecture: a single linear layer added on top of the frozen AFDs for classification, with a dropout layer (dropout rate = 0.3) applied before classification. Each text sample is represented using mean pooling of all token embeddings. Maximum length is set to 512 tokens, with shorter samples padded and longer samples truncated to max length.

We use a 50:50 split of the provided validation dataset for validation and testing. Training employs early stopping (with patience = 2), retaining the model with the lowest validation loss. Optimization is performed using the AdamW optimizer with a linear warmup scheduler (10% warmup steps). The learning rate is set to 2×10^{-5}, with a batch size of 32. Although the maximum training epochs are set to 15, early stopping is usually triggered within 4 epochs in practice. Mixed-precision training with gradient scaling is used to speed-up training. No further hyperparameter tuning is performed. We release our data and source code for reproducibility[1].

5 Results

Table 2 presents the results from our comprehensive evaluation of five recent AFDs. Three key insights emerge: First, most of the AFDs outperform baseline models by a significant margin, confirming our hypothesis. Second, the best-performing system demonstrates effective bias mitigation. Finally, contrastive learning appears unnecessary, as AFDs demonstrate inherent class separability.

We evaluated 20 models across four loss configurations and five AFDs, with *15 achieving a positive gain over the baseline—a **75%** success rate*, validating our hypothesis and highlighting the effectiveness of small AFDs for AI-generated content detection. The highest performance was achieved with the SmolLM model, likely due to its training on the high-quality SmolLM-Corpus (Ben Allal et al., 2024), which includes a mix of

[1] https://github.com/AvantiB/AAIG-at-GenAI-Detection-Task-1

human and synthetically generated text.

To analyze performance imbalance between the human and LLM classes, we examined their Recall scores on the test set. While baseline models achieve high recall for the LLM class, they perform poorly on the human class - confirming the bias reported in previous studies. In contrast, AFD models show a slight decrease in LLM recall but deliver a substantial improvement in recall for the human class - effectively mitigating bias. *Our best model increases human recall by 60% and 28% over our and task baselines, with minor LLM recall reductions of 9% and 5%, respectively.*

Focal loss consistently outperforms cross-entropy, but the addition of contrastive loss does not always lead to performance improvements. While contrastive loss occasionally enhances the performance of PLMs, AFDs generally show a decline when it is applied. This can be attributed to the already well-separated nature of AFD embeddings where further enforcing separation may over-penalize instances near class boundaries. This observation reinforces the suitability of AFDs, as their inherent syntactic awareness provides strong separability, making additional loss optimization redundant and reducing computational overhead.

Can ensembling AFDs improve performance?

Although the primary goal of this paper is to demonstrate that small LMs are as capable, if not superior, to LLMs, the variability in performance between models raises the question of whether ensembling them could improve results. For instance, GPT-2 performs better than SmolLM in detecting the LLM class, while SmolLM outperforms GPT-2 for the Human class. This suggests that if the models leverage different aspects of the text for detecting AI-generated content, ensembling them might lead to performance gains.

We test this hypothesis by experimenting with all combinations (N=1,2,3,4,5) of AFDs using majority voting on the test set. Figure 1 presents a box-and-whisker plot of performance across different ensemble configurations, where a wider spread reflects greater variability based on the combination of AFDs in the ensemble. Although no significant improvement over the proposed single AFD system is observed, some performance gains are evident, particularly in the Focal+Contrastive loss setup, which achieves the highest macro-F1 score of 0.8295– 9.37 %Gain over the baseline.

Two main findings from the ensemble testing arise: First, performance tends to degrade as the number of models in the ensemble increases, with the best performance achieved at N=2, and second, the combination of SmolLM and GPT-2 consistently delivers strong results. As previously noted, their complementary strengths make this ensemble particularly robust, with each model enhancing the other's performance. Additionally, some improvements are observed when SmolLM is combined with GPT-Neo or OPT.

This suggests that *the success of the ensemble relies heavily on the proposed system using SmolLM, whose complementary strengths enhance other models, highlighting its potential for refinement in low-resource setups*.

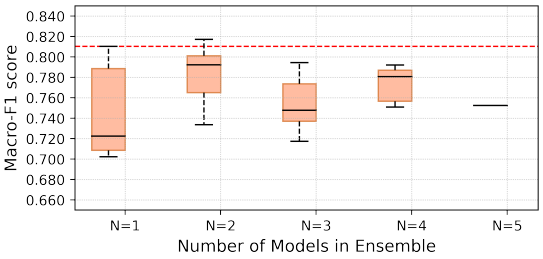
6 Discussion

Given the test set's significant differences from the training set, namely, unfamiliar data sources, adversarially modified text, and unknown LLMs, it is imperative to analyze their impact on model performance.

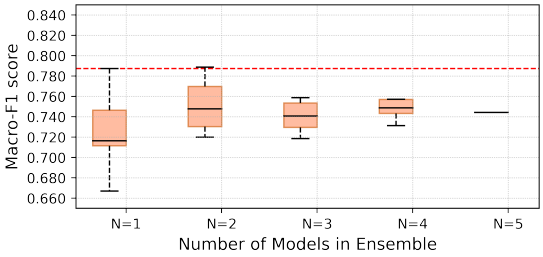
Figure 2 presents a box-and-whisker plot of model performance by data source, with greater spread indicating variability. Overall, the *Human class shows lower performance compared to the LLM class*, with variations across sources. Mixset and CUDRT have the highest misclassification rate for the LLM class, likely due to the inclusion of adversarially perturbed text samples. For the Human class, Mixset, DetectAIve, and ieltsduck show the lowest performance. DetectAIve and ieltsduck contain IELTS test takers' data, likely written by non-native English speakers, contributing to misclassifications. In Mixset, LLM-modified samples labeled as human may also lead to errors.

The perturbation operations performed on each dataset are described in Table 3 with the performance per operation depicted in Figure 3. Overall, *adversarial perturbations increase susceptibility to misclassifications*. Notably, the "summary" operation from the CUDRT dataset results in highest misclassification rate, likely due to concise nature of text, providing insufficient information for accurate classification. Similarly, the "polish" and "complete" operations also degrade performance. However, recent studies suggest that incorporating small amounts of adversarial examples in training can improve detectors' ability to handle perturbed AI-generated content (Zhang et al., 2024).

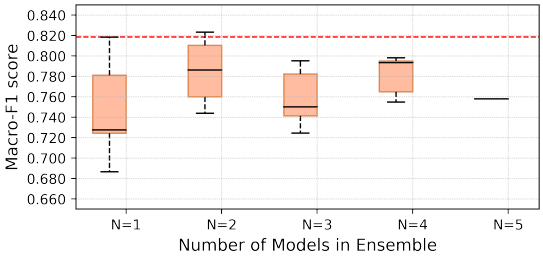
Detection accuracy for each LLM in the test set is depicted in Figure 4. Variations of the Chat-



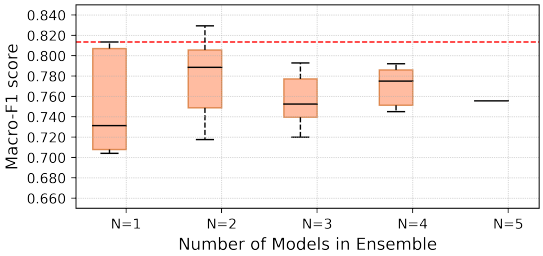
(a) Loss: Cross-Entropy



(b) Loss: Cross-Entropy+Contrastive



(c) Loss: Focal



(d) Loss: Focal+Contrastive

Figure 1: Ensemble of AFDs using majority voting across different loss function configurations. Red dashed line represents the performance of proposed (single AFD) system. Spread of boxes represents performance variability due to choice of AFDs in ensemble.

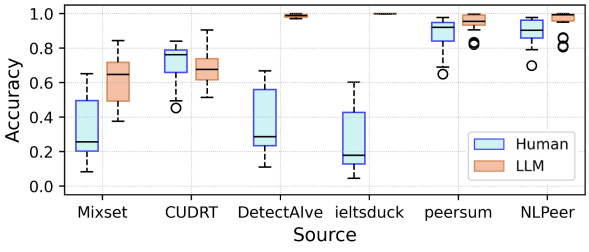


Figure 2: Model performance by source of data

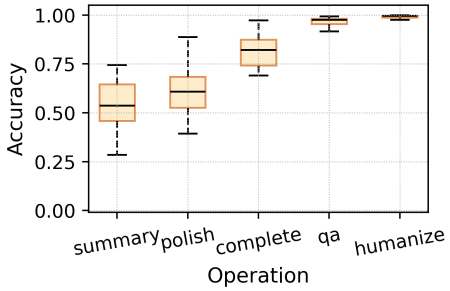


Figure 3: Performance across adversarial perturbations

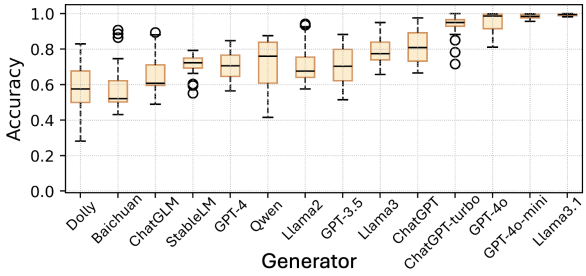


Figure 4: Model performance per LLM in test set

GPT family demonstrate stronger detection, likely due to the inclusion of related models from the same family in the training set. Nevertheless, the generalization of performance across data sources underscores our model's effectiveness. Moreover, ChatGPT's detectability aligns with existing research (Bhandarkar et al., 2024), ensuring safety against misuse of most widely used chatbot. In contrast, lesser-known models like Dolly and Baichuan show lower detection rates, highlighting areas for improvement.

7 Conclusion

This paper addresses the challenge of secure on-device AI content detection by proposing a simple yet effective solution leveraging small AFDs. With their resource-efficient design and syntactic alignment enabled by autoregressive pre-training, the proposed approach-combining AFDs with loss fusion, particularly focal loss-outperforms larger, resource-intensive models by a large margin while reducing model size. Our proposed approach mitigates bias, maintains generalization, and handles challenging data, including unknown domains, unseen LLMs, and adversarially modified text. These results highlight the potential of small AFDs as efficient backbones or ensemble components, especially in scenarios requiring data privacy and faster AI-content detection.

References

Mervat Abassy, Kareem Elozeiri, Alexander Aziz, Minh Ta, Raj Tomar, Bimarsha Adhikari, Saad Ahmed, Yuxia Wang, Osama Mohammed Afzal, Zhuohan Xie, et al. 2024. Llm-detectaive: a tool for fine-grained machine-generated text detection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 336–343.

Harika Abburi, Michael Suesserman, Nirmala Pudota, Balaji Veeramani, Edward Bowen, and Sanmitra Bhattacharya. 2023. Generative ai text classification using ensemble llm approaches. *arXiv preprint arXiv:2309.07755*.

Bo Ai, Yuchen Wang, Yugin Tan, and Samson Tan. 2022. Whodunit? learning to contrast for authorship attribution. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1142–1157, Online only. Association for Computational Linguistics.

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Leandro von Werra, and Thomas Wolf. 2024. Smollm - blazingly fast and remarkably powerful.

Loubna Ben Allal, Anton Lozhkov, Guilherme Penedo, Thomas Wolf, and Leandro von Werra. 2024. Smollm-corpus.

Avanti Bhandarkar, Ronald Wilson, Anushka Swarup, Mengdi Zhu, and Damon Woodard. 2024. Is the digital forensics and incident response pipeline ready for text-based threats in llm era? *Preprint*, arXiv:2407.17870.

Sid Black, Gao Leo, Phil Wang, Connor Leahy, and Stella Biderman. 2021. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow. If you use this software, please cite it using these metadata.

Vittorio Ciccarelli, Cornelia Genz, Nele Mastracchio, Wiebke Petersen, Anna Stein, and Hanxin Xia. 2024. Team art-nat-hhu at semeval-2024 task 8: Stylistically informed fusion model for mgt-detection. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1690–1697.

Shubhashis Roy Dipta and Sadat Shahriar. 2024. Hu at semeval-2024 task 8a: Can contrastive learning learn embeddings to detect machine-generated text? *arXiv preprint arXiv:2402.11815*.

Ahmed El-Sayed and Omar Nasr. 2023. An ensemble based approach to detecting LLM-generated texts. In *Proceedings of the 21st Annual Workshop of the Australasian Language Technology Association*, pages 164–168, Melbourne, Australia. Association for Computational Linguistics.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander M Rush. 2019. Gltr: Statistical detection and visualization of generated text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 111–116.

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. *arXiv preprint arXiv:2401.12070*.

Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. 2020. Automatic detection of generated text is easiest when humans are fooled. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1808–1822.

Zhixin Lai, Xuesheng Zhang, and Suiyao Chen. 2024. Adaptive ensembles of fine-tuned transformers for llm-generated text detection. *arXiv preprint arXiv:2403.13335*.

Yafu Li, Qintong Li, Leyang Cui, Wei Bi, Zhilin Wang, Longyue Wang, Linyi Yang, Shuming Shi, and Yue Zhang. 2024. MAGE: Machine-generated text detection in the wild. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 36–53, Bangkok, Thailand. Association for Computational Linguistics.

Zechun Liu, Changsheng Zhao, Forrest Iandola, Chen Lai, Yuandong Tian, Igor Fedorov, Yunyang Xiong, Ernie Chang, Yangyang Shi, Raghuraman Krishnamoorthi, et al. 2024. Mobilellm: Optimizing sub-billion parameter language models for on-device use cases. In *Forty-first International Conference on Machine Learning*.

Zhenyan Lu, Xiang Li, Dongqi Cai, Rongjie Yi, Fangming Liu, Xiwen Zhang, Nicholas D Lane, and Mengwei Xu. 2024. Small language models: Survey, measurements, and insights. *arXiv preprint arXiv:2409.15790*.

Anqi Mao, Mehryar Mohri, and Yutao Zhong. 2023. Cross-entropy loss functions: Theoretical analysis and applications. In *International conference on Machine learning*, pages 23803–23828. PMLR.

Jishnu Mukhoti, Viveka Kulharia, Amartya Sanyal, Stuart Golodetz, Philip Torr, and Puneet Dokania. 2020. Calibrating deep neural networks using focal loss. *Advances in Neural Information Processing Systems*, 33:15288–15299.

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

Areg Mikael Sarvazyan, José Ángel González, and Marc Franco-salvador. 2024. Genaios at SemEval-2024

task 8: Detecting machine-generated text by mixing language model probabilistic features. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 101–107, Mexico City, Mexico. Association for Computational Linguistics.

M Sheykhlan, S Abdoljabbar, and M Mahmoudabad. 2024. Team karami-kheiri at pan: enhancing machine-generated text detection with ensemble learning based on transformer models. *Working Notes of CLEF*.

Zhen Tao, Zhiyu Li, Dinghao Xi, and Wei Xu. 2024. Cudrt: Benchmarking the detection of human vs. large language models generated texts. *arXiv preprint arXiv:2406.09056*.

Rongsheng Wang, Haoming Chen, Ruizhe Zhou, Han Ma, Yaofei Duan, Yanlan Kang, Songhua Yang, Baoyu Fan, and Tao Tan. 2024a. Llm-detector: Improving ai-generated chinese text detection with open-source llm instruction tuning. *arXiv preprint arXiv:2402.01158*.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Osama Mohammed Afzal, Tarek Mahmoud, Giovanni Puccetti, and Thomas Arnold. 2024b. SemEval-2024 task 8: Multidomain, multimodel and multilingual machine-generated text detection. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 2057–2079, Mexico City, Mexico. Association for Computational Linguistics.

Yuxia Wang, Artem Shelmanov, Jonibek Mansurov, Akim Tsvigun, Vladislav Mikhailov, Rui Xing, Zhuohan Xie, Jiahui Geng, Giovanni Puccetti, Ekaterina Artemova, Jinyan Su, Minh Ngoc Ta, Mervat Abassy, Kareem Elozeiri, Saad El Dine Ahmed, Maiya Goloburda, Tarek Mahmoud, Raj Vardhan Tomar, Alexander Aziz, Nurkhan Laiyk, Osama Mohammed Afzal, Ryuto Koike, Masahiro Kaneko, Alham Fikri Aji, Nizar Habash, Iryna Gurevych, and Preslav Nakov. 2025. GenAI content detection task 1: English and multilingual machine-generated text detection: AI vs. human. In *Proceedings of the 1st Workshop on GenAI Content Detection (GenAIDetect)*, Abu Dhabi, UAE. International Conference on Computational Linguistics.

Junchao Wu, Shu Yang, Runzhe Zhan, Yulin Yuan, Derek F. Wong, and Lidia S. Chao. 2024. A survey on llm-generated text detection: Necessity, methods, and future directions. *Preprint*, arXiv:2310.14724.

Jiajun Xu, Zhiyuan Li, Wei Chen, Qun Wang, Xin Gao, Qi Cai, and Ziyuan Ling. 2024. On-device language models: A comprehensive review. *Preprint*, arXiv:2409.00088.

Qihui Zhang, Chujie Gao, Dongping Chen, Yue Huang, Yixin Huang, Zhenyang Sun, Shilin Zhang, Weiye Li, Zhengyan Fu, Yao Wan, et al. 2024. Llm-as-a-coauthor: Can mixed human-written and machine-generated text be detected? In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 409–436.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

A Appendix

Table 3: Perturbation Operations from Various Datasets. (*) denotes LLM-generated text labeled as human

Operation	Description	Source		
		Mixset	DetectAIve	CUDRT
Polish	Improve quality, fluency, accuracy; includes refine, rewrite, paraphrase, etc.	✓	✓	✓
Complete	Generate part LLM, part human text by completing a given text portion	✓		✓
Q/A	LLMs act as expert to provide detailed answers to questions.			✓
Summary	Generate concise summary, highlighting main points and key information.			✓
Humanize	Add human-like noise (e.g., typos, grammatical errors, tags).	✓*	✓	