Grape at GenAI Detection Task 1: Leveraging Compact Models and Linguistic Features for Robust Machine-Generated Text Detection

Nhi Hoai Doan and **Kentaro Inui**
Mohamed bin Zayed University of Artificial Intelligence (MBZUAI)
{nhi.doan, kentaro.inui}@mbzuai.ac.ae

Abstract

In this project, we aim to solve two Subtasks of Task 1: Binary Multilingual Machine-Generated Text(MGT) Detection (Human vs. Machine) as part of the COLING 2025 Workshop on MGT Detection(Wang et al., 2025) by different approaches. The first method is separate fine-tuned small language models on the specific subtask. The second approach enhances this methodology by incorporating linguistic, syntactic, and semantic features, using ensemble learning to combine these features with model predictions for a more robust classification. By evaluating and comparing these approaches, we want to identify the most effective techniques for detecting machine-generated content across languages, offering insights into improving automated verification tools amid the rapid growth of LLM-generated text in digital spaces. The code of this project is available at here.

1 Introduction

The rapid development of large language models (LLMs) such as GPT-4o, Claude3.5, and Gemini1.5-pro has led to an explosion of machine-generated text across various channels, including news, social media, and academic publications. Khalifa and Albadawy (2024), based on 24 studies of academic domains, points out that using artificial intelligence enhances the productivity of researchers. While this advancement is promising, it has raised significant concerns about misuse, including spreading misinformation and potential disruptions in educational contexts due to the unpredictable nuance of these language models. To address these issues, it is crucial to develop effective systems for distinguishing between human-written and machine-generated content. There are two subtasks in the Task 1:

- Subtask A: English-only machine-generated text(MGT) detection.
- Subtask B: Multilingual MGT detection with nine languages.

The primary goal of this project is to develop an automatic detection system capable of distinguishing machine-generated text from human-written text using small-sized language models. By integrating models with fewer parameters—thus lower computational demands—we aim to demonstrate that effective detection does not require large, resource-intensive models. Specifically, our objectives are to:

- Explore important linguistic, syntactic, and semantic features for human and machine text-generated differentiation.
- Implement and evaluate **newly released language models** in small sizes for text classification.
- Assess model performance and provide insights on machine-generated content detection effectiveness.

2 Related Work

Recent research has focused on detecting machine-generated text using various techniques. For the traditional methodologies, GLTR uses statistical methods to detect generated text with an improvement in human detection of fake text from 54% to 72%(Gehrmann et al., 2019). With the explosive growth of Transformers and Large Language Models(LLMs), Uchendu et al. (2021) shows that FAIR_wmt20 and GPT-3 excel at generating human-like text. Recently, in the *SemEval-2024 Task 8: Multidomain, Multimodel and Multilingual Machine-Generated Text Detection*(Wang et al., 2024), many researchers have tried to apply different approaches, such as statistical, language models, and LLMs to solve the *Subtask A: Human vs. Machine Classification*. Sarvazyan et al. (2024)

Proceedings of the 1st Workshop on GenAI Content Detection (GenAIDetect), pages 209–217
January 19, 2025. ©2025 International Conference on Computational Linguistics

study mixing Llama-2 features, achieved top accuracy. Their performance relies on multiple LLMs and features, focusing on the last tokens. Other teams also attempt to use language models, such as RoBERTa or XLM-RoBERTa(Sarvazyan et al., 2024; Petukhova et al., 2024; Tran et al., 2024).

Regarding our hardware limitations, we want to try to evaluate newly released language models in small sizes. In addition, all previous works on the SemEval-2024 Task 8 mostly work with LLMs. Based on the success of Sarvazyan et al. (2024) with Llama-2, we consider using Llama3(Dubey et al., 2024). While Spiegel and Macko (2024) proposed combined fine-tuned LLMs with zero-shot statistical methods, employing a two-step majority voting system for predictions, Petukhova et al. (2024) utilized a fine-tuned baseline - RoBERTa augmented with diverse linguistic features. All these methods surpass the baseline and achieve good results, supporting our approach, which is a potential way to mix LLMs with traditional linguistic features.

3 Proposed Approach

Recent released large language models, such as Llama 3 (Dubey et al., 2024) or Gemma 2 (Team et al., 2024), are now available in smaller configurations. These smaller models still perform well on popular benchmarks while being more compatible with hardware constraints. Therefore, we decided to fine-tune these models for our task, utilizing their smaller versions to match our hardware limitations.

3.1 Subtask A: Monolingual - English

This subtask focuses on detecting machine-generated text in English generated by *hc3, m4gt, and mage*. We want to use fine-tuned language models and traditional linguistic features as their potential performance from previous research on the same task (Spiegel and Macko, 2024; Tran et al., 2024). The methodology integrates neural network-based approaches with gradient boosting and combines the outputs through a majority voting mechanism. The strategy is outlined in Figure 1:

Fine-Tuning of Small Language Models: The recent availability of smaller, efficient language models, such as Llama3.2-1B and Gemma-2-2B, makes them suitable candidates for fine-tuning on this task. Despite their compact size, these models maintain competitive performance, comparable to larger counterparts like Mixtral 8x7B and

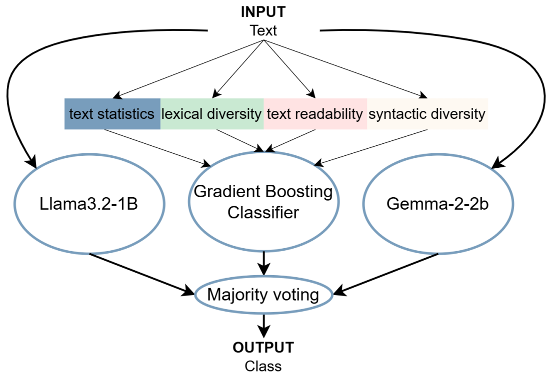


Figure 1: Approach for monolingual task

GPT-3.5. Fine-tuning these models on the task-specific dataset enables them to capture intricate patterns indicative of machine-generated content. Their efficient architecture ensures compatibility with hardware constraints, allowing faster training and inference.

Gradient Boosting Classifier with Linguistic Features: In addition to fine-tuning Llama3.2-1b and Gemma-2-2B, a gradient-boosting classifier will be trained using a comprehensive set of linguistic features extracted from the text. These metrics can provide helpful information when the input to language models is limited. Therefore, adding other linguistic features will allow them to gain information from the truncation part. As desired from the work of (Petukhova et al., 2024), we use four metrics with updated features as follows:

- *Syntactic Complexity*: Metrics obtained from *spaCy*[1] such as average sentence length, average number of noun phrases per sentence, and average number of verbs per sentence capture syntactic patterns and variations within the text.

- *Readability Metrics*: include Flesch Reading Ease, Flesch-Kincaid Grade Level, Gunning Fog Index, SMOG Index, and others assess the ease or difficulty of reading the text. We get these metrics by using the *textstat* package[2].

- *Lexical Diversity*: Metrics such as Type-Token Ratio (TTR), Maas TTR, Hypergeometric Distribution Diversity (HDD), and Mean Length of Textual Diversity (MLTD)[3] provide

[1] https://pypi.org/project/spacy/
[2] https://pypi.org/project/textstat/
[3] https://pypi.org/project/lexical-diversity/

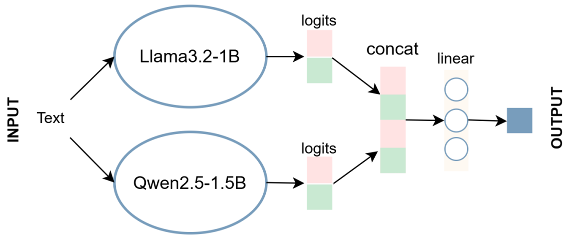


Figure 2: Ensemble model for multilingual task

insights into the lexical richness of the text. Machine-generated texts may exhibit unusual lexical patterns, making these metrics valuable.

- *Text Statistics*: Basic statistics such as the number of difficult words, unique word count, and sentence count offer additional context about the text structure. These can reveal inconsistencies or unnatural pasts often present in machine-generated content.

By combining these diverse features, the gradient-boosting classifier can capture non-linear relationships between linguistic characteristics and the target variable, complementing the capabilities of the fine-tuned language models.

Majority Voting Ensemble: To enhance the robustness and accuracy of the system, a majority voting mechanism will be employed to combine the outputs of the three models: fine-tuned versions of Llama3.2-1b and Gemma-2-2B and the gradient boosting classifier. Each model will provide its prediction, and the final decision will be determined by the majority vote among the three. This ensemble approach leverages each component's strengths, balancing the fine-tuned models' deep contextual understanding with the interpretability and feature-driven analysis of the gradient-boosting classifier.

3.2 Subtask B: Multilingual

This subtask extends the detection of machine-generated text to a multilingual setting. Given the time constraints and resource limitations, the approach will leverage two fine-tuned multilingual models, Llama-3 1B and Qwen-2.5 1.5B (Hui et al., 2024). These models have been selected for their efficient architectures and ability to handle multiple languages effectively. The methodology for Subtask B follows a similar ensemble-based strategy as outlined in Subtask A, with modifications to accommodate multilingual data.

As we see in Figure 2, an ensemble architecture has been developed to combine the strengths of Llama-3 and Qwen-2.5. Each model is fine-tuned separately on the training dataset, and their outputs are then combined through learnable weights. We do not use text linguistic features here because of inconsistent and unavailable support tools for non-English languages. Therefore, we choose to create ensemble models based on fine-tuned multilingual models.

4 General settings

4.1 Experiments

For each subtask, we fine-tune and measure the results of each small model individually before applying majority and ensemble learning methods. The hyperparameters of Llama3.2-1b, Gemma-2-2B, and Qwen2.5-1.5B are learning rate = 2e-5, batch size = 16, max token length = 256, lora = 16, and epoch = 5. Our hardware computation resource is 1× NVIDIA GeForce RTX4090 24GB and is limited to 24 hours.

4.2 Evaluation metrics and baselines

The official evaluation metric is the macro f1-score. Another metric is micro-F1. The task also provided a baseline result for the English track using RoBERTa, which is 81.63. The result for the multilingual track using XLM-R is 65.46.

5 Results

5.1 Subtask A: English-only MGT detection

Overall, the results of the methods for Subtask A, as shown in Table 1, confirm our intuition. The *2SLMs* is a combination of 2 small language models by averaging logits of them. While *2SLMs* combination

Model	Macro F1	Micro F1
Linguistic Features	0.7094	0.7148
Llama3.2-1b	0.8798	0.8843
Gemma-2-2B	0.9070	0.9100
2SLMs	0.9088	0.9117
Majority Voting	**0.9225**	**0.9248**

Table 1: Performance of models on Subtask A dev_test set

improves slightly, adding linguistic features generally increases both metrics. This can be explained by the fact that we had to limit the maximum token length with the two small language models due to hardware constraints. It allowed the model to consider information from the truncated part of the text. For example, one case when *Majority Voting* successfully recognizes the text generated by *human* but *2SLMs* fails:

`<260 tokens>. . . Fill the bowl with enough cool tap water to cover the rice by an inch or two. Use your hand to gently stir the rice, then lift the strainer from the bowl. The water in the bowl will be cloudy from the rice starch. Empty the water, set the strainer in the bowl again, and repeat the process until the water is, more or less, clear. You'll probably have to change the water two or three times. Drain the rice. Pour enough wate . . . < 400 tokens>`

We can see that this text has more than 700 tokens which exceeds our max token length = 256. Previous part of the text describe step by step to prepare a dish but only after the considered context, we see the colloquial phrases like *"you'll probably have to change the water two or three times"* which align with human authorship. In addition, in the rest 400 tokens, it also contains natural and diversity words. Therefore, *Linguistic Features* could have the decision making power in such these cases.

5.2 Subtask B: Multilingual MGT detection

We employ two small language models for multilingual machine-generated text detection in this subtask, as illustrated in Table 2. The ensemble model achieved the best Macro F1 score at 0.7388, indicating its effectiveness in balancing the accuracy across different classes. Combining both models, the ensemble approach enhances generalization across multiple languages, which is beneficial in multilingual settings. However, the Micro F1 score (0.8829) slightly declined compared to Qwen2.5-1.5B, suggesting that while the ensemble model captures class balance well, it may sacrifice a bit of precision on individual sample classifications.

Model	Macro F1	Micro F1
Llama3.2-1b	0.6878	0.8619
Qwen2.5-1.5B	0.7292	**0.8869**
Ensemble	**0.7388**	0.8829

Table 2: Performance of models on Subtask B dev_test set

5.3 Results on the test set

Based on the results of the development dataset, we selected the Majority model and the Ensemble model to submit as the final results in Table 3. Since the golden labels are not publicly available, we cannot definitively conclude which approach is the most effective. However, for Subtask A, our result was 0.8188 — approximately a 0.09-point improvement over the baseline and 0.05-point improvement in Subtask B — indicating that these are promising approaches.

Subtask	Model	Macro F1	Rank
A	Baseline	0.7342	4/35
	Majority Voting	**0.8188**	
B	Baseline	0.7416	1/25
	Ensemble	**0.7916**	

Table 3: Our performance on the test set with **Score** is as *Macro F1*

Generally, ensemble learning is a potential approach, especially when each component has its own strength. In our study, small language models can solve our hardware limitation while maintaining good performance; their disadvantage is that they do not fully capture all information of the text. These models, even when combined, usually give similar results. Our intuition is that in the case of conflicts, the result of the model with more parameters is favored. However, additional linguistic features can handle these cases by looking for the whole text. Although this paper does not evaluate the individual contribution of each feature, we believe that further exploration could yield improvements in model performance.

The organization describes more details about other team methods in subtask A in Table 4. The top-ranking team, Advacheck, utilized a multi-task system with a shared Transformer encoder (DeBERTa-v3-base) and multiple classification heads, leveraging multi-task learning to optimize performance. Unibuc-NLP ranked 2nd with a combination of masked (XLM-RoBERTa) and causal (Qwen 2.5-0.5B) language models, enhanced by LoRA fine-tuning. At the same time, Fraunhofer SIT used adapters for task-specific optimization on RoBERTa-base. While more complex than the top teams' methods, our ensemble-based strategy demonstrates the value of integrating diverse model outputs to achieve competitive performance. Future enhancements could include incorporating multi-task learning or adapter-based approaches for further gains.

Team Name	Ranking	Small PLM	LLM	Feature Combination	Ensemble
Advacheck	1	✓			
Unibuc-NLP	2		✓		
FraunhoferSIT	3	✓			
Our team	4	✓	✓		✓
TechExperts(IPN)	5		✓		

Table 4: English subtask participants overview

Regarding the multilingual test set, we have the result analysis from the organization (Wang et al., 2025) as in Table 5 and Table 6.

When comparing our proposed approach with other teams as described in Table 5 (Wang et al., 2025), our method demonstrates a clear focus on efficiency and robustness by leveraging Small PLMs and ensemble techniques, achieving the top ranking. Unlike teams such as Nota AI and Lux Veri, who utilized broader combinations of techniques, including LLMs and feature engineering, our streamlined approach highlights the effectiveness of simplicity combined with targeted ensemble learning.

Team Name	Ranking	Small PLM	LLM	Feature Combination	Ensemble
Our team	1	✓			✓
Nota AI	3	✓	✓	✓	✓
Lux Veri	4	✓			✓
TechExperts (IPN)	5		✓		

Table 5: Multilingual subtask participants overview.

From Table 6, we could see that our team result for Subtask B Multilingual surpassed the baseline by around 5 percent, and our gap with the second team is 4 percent overall. In the test set, six hidden languages were not present in the training set of this task: Kazakh (KK), Vietnamese(VI), Hindi(HI), Hebrew(HE), Norwegian(NO), and Japanese(JA). Because models are not exposed to many linguistic patterns, structures, and features during training, it is difficult for them to generalize to unknown languages. For example, we achieved only 51.8 on Hindi.

6 Conclusion

We successfully addressed both subtasks using majority voting and ensemble methods. Our approach comprised fine-tuned small language models and linguistic features, contributing to robust task performance. Specifically, fine-tuning small language models allowed us to capture critical nuances in the data while maintaining computational efficiency. Meanwhile, incorporating linguistic features, such as syntactic complexity, readability metrics, and lexical diversity, added a complementary layer of information that enhanced the ensemble's overall effectiveness.

Limitations

Although the result in Section 5.1 shows that using linguistic features improves the model's performance, we have not investigated each feature. Furthermore, no additional linguistic features specific to each language are analyzed regarding the multilingual track. Future work could be research such features, including syntax-specific markers, morphological distinctions, and domain-specific language idiosyncrasies for each language to provide valuable insights and boost classification accuracy.

References

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar,

Table 6: Multilingual subtask detection accuracy across 15 languages. *Underlined languages were not present in the training data.*

Rank	**All**	**ZH**	**UR**	**RU**	**AR**	**IT**	**KK**	**VI**	**DE**	**NO**	**ID**	**NL**	**ES**	**HI**	**HE**	**JA**
Size	151,425	63,009	30,505	27,158	10,670	5,296	2,471	2,326	1,865	1,544	1,200	1,200	1,200	1,199	1,182	600
Our team	79.6	94.2	68.7	67.1	71.2	52.9	55.5	90.5	88.3	80.3	89.6	82.2	89.5	51.8	86.7	77.0
2	75.6	84.7	64.6	74.2	57.9	52.9	83.8	83.5	96.4	76.0	51.7	90.6	91.2	69.6	96.8	95.3
3	75.9	90.2	67.2	58.9	66.8	52.9	92.5	74.7	88.8	72.2	87.4	68.9	47.1	70.6	96.4	72.2
4	75.3	87.6	64.6	63.9	61.3	52.9	75.8	83.4	94.9	88.5	53.5	92.2	90.4	73.0	97.3	92.2
BL	74.8	87.3	68.4	55.3	68.4	52.9	82.8	85.3	85.2	69.8	68.2	92.5	90.5	71.3	89.3	90.0
5	74.7	90.1	64.1	56.0	69.1	52.9	62.9	87.6	59.6	69.8	93.8	81.0	90.4	69.1	96.5	95.0
6	74.5	84.2	65.0	67.9	66.8	52.9	47.5	81.8	93.5	83.2	83.9	85.9	88.9	69.1	89.8	78.2

Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander 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, Florence, Italy. Association for Computational Linguistics.

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Kai Dang, An Yang, Rui Men, Fei Huang, Xingzhang Ren, Xuancheng Ren, Jingren Zhou, and Junyang Lin. 2024. Qwen2.5-coder technical report. *Preprint*, arXiv:2409.12186.

Mohamed Khalifa and Mona Albadawy. 2024. Using artificial intelligence in academic writing and research: An essential productivity tool. *Computer Methods and Programs in Biomedicine Update*, 5:100145.

Kseniia Petukhova, Roman Kazakov, and Ekaterina Kochmar. 2024. PetKaz at SemEval-2024 task 8: Can linguistics capture the specifics of LLM-generated text? In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1140–1147, Mexico City, Mexico. Association for Computational Linguistics.

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.

Michal Spiegel and Dominik Macko. 2024. KInIT at SemEval-2024 task 8: Fine-tuned LLMs for multilingual machine-generated text detection. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 558–564, Mexico City, Mexico. Association for Computational Linguistics.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena

Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Hanh Thi Hong Tran, Tien Nam Nguyen, Antoine Doucet, and Senja Pollak. 2024. L3i++ at SemEval-2024 task 8: Can fine-tuned large language model detect multigenerator, multidomain, and multilingual black-box machine-generated text? In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 13–21, Mexico City, Mexico. Association for Computational Linguistics.

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. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2001–2016, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Osama Mohammed Afzal, Tarek Mahmoud, Giovanni Puccetti, and Thomas Arnold. 2024. 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.

Table 7: Summary of monolingual training dataset for subtask A

Src	*Train*		*Dev*	
	Human	Machine	Human	Machine
hc3	39140	18091	16855	7917
m4gt	86782	181081	37220	71197
mage	103000	182673	44253	84316
total	228922	381845	98328	163430

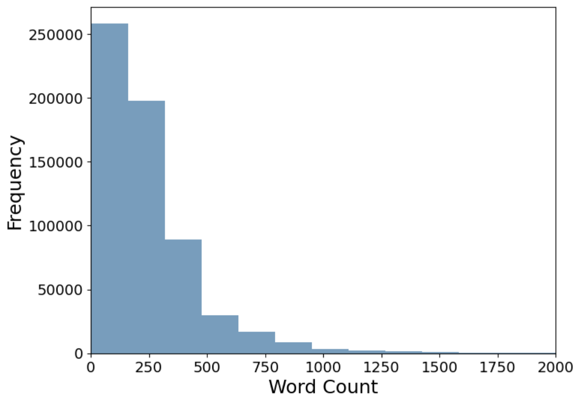
Table 8: Summary of multilingual training dataset for subtask B

Lan	*Train*		*Dev*	
	Human	Machine	Human	Machine
ar	344	1770	150	756
bg	4205	3886	1795	1694
en	223911	386877	98041	163808
de	231	4462	102	1957
id	1895	2081	886	917
it	0	4174	0	1843
ru	684	630	316	284
ur	2085	1676	853	720
zh	19315	15969	8023	6749
total	257968	416115	110166	178728

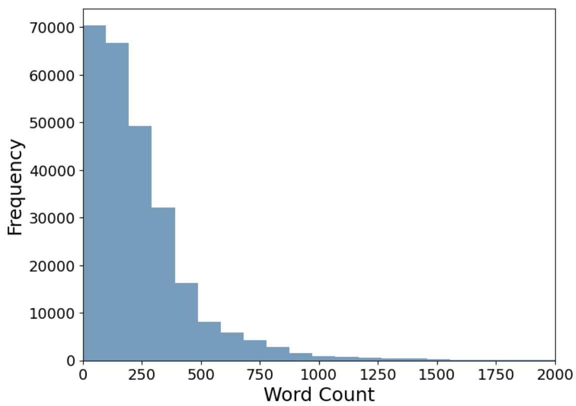
A. Data Analysis

The task provides datasets in multiple domains and multi-model and multilingual text. The organizer extends this dataset from the one provided in *SemEval-2024 Task 8*. Details of the English-only subtask are in Table 7. The ratio of text for each class *human* or *machine* is consistent in both the train and dev set, around 37 %. Table 8 illustrates the distribution of the number of each class per language in two datasets. We see an imbalance across languages that more than 90 % of the text in the training dataset is English. This could cause the model to find it hard to identify each language because using an external dataset is not allowed by the organizer.

Regarding text's length, as we see in Figure 3, while around 70% of the text has a length of 250, the rest are in a range from there to more than 20000 words per text. Using linguistic features can gain valuable information from the truncated part, which small language models ignore.



(a) Train Dataset



(b) Dev Dataset

Figure 3: Distribution of number of words per text in English datasets