CMI-AIGCX at GenAI Detection Task 2: Leveraging Multilingual Proxy LLMs for Machine-Generated Text Detection in Academic Essays

Kaijie Jiao[1], **Xingyu Yao**[2], **Shixuan Ma**[2], **Sifan Fang**[2], **Zikang Guo**[1], **Benfeng Xu**[1]
Licheng Zhang[1], **Quan Wang**[2], **Yongdong Zhang**[1] and **Zhendong Mao**[1*]
[1]University of Science and Technology of China, Hefei, China
[2]MOE Key Laboratory of Trustworthy Distributed Computing and Service,
Beijing University of Posts and Telecommunications, Beijing, China
zdmao@ustc.edu.cn

Abstract

This paper presents the approach we proposed for GenAI Detection Task 2, which aims to classify a given text as either machine-generated or human-written, with a particular emphasis on academic essays. We participated in subtasks A and B, which focus on detecting English and Arabic essays, respectively. We propose a simple and efficient method for detecting machine-generated essays, where we use the Llama-3.1-8B as a proxy to capture the essence of each token in the text. These essences are processed and classified using a refined feature classification network. Our approach does not require fine-tuning the LLM. Instead, we leverage its extensive multilingual knowledge acquired during pretraining to significantly enhance detection performance. The results validate the effectiveness of our approach and demonstrate that leveraging a proxy model with diverse multilingual knowledge can significantly enhance the detection of machine-generated text across multiple languages, regardless of model size. In Subtask A, we achieved an F1 score of **99.9%**, ranking **first** out of 26 teams. In Subtask B, we achieved an F1 score of 96.5%, placing fourth out of 22 teams, with the same score as the third-place team.

1 Introduction

The capabilities of large language models (LLMs) are advancing rapidly, with models like, ChatGPT (OpenAI, 2022), GPT-4 (OpenAI et al., 2024), Google Gemini (Team et al., 2024), and Llama3.1 (Dubey et al., 2024) generating increasingly fluent and human-like text. Students can easily leverage these models to produce coherent, logical texts for assignments or essays, which profoundly impacts traditional educational methods of learning and evaluation, leading to issues in academic integrity and a weakening of critical thinking skills. However, humans perform only slightly better than random chance in distinguishing between machine-generated and human-written text (Mitchell et al., 2023), underscoring the urgent need for an automated system to identify machine-generated content. To address this, (Chowdhury et al., 2025) organized the GenAI Detection Task 2, a challenge focused on detecting machine-generated academic essays in English and Arabic to uphold academic authenticity and prevent the misuse of LLMs in educational contexts.

Most current methods for detecting machine-generated text can be generally categorized into two approaches (Taguchi et al., 2024): zero-shot detection and supervised detection. The former is time-consuming and suffers from performance degradation when the generation model is unknown, while the latter like RoBERTa-based detection (Guo et al., 2023) requires fine-tuning large models, which is resource-intensive and often lacks multilingual capabilities. In contrast, we employed a multilingual model, such as Llama-3.1-8B (Dubey et al., 2024), as a proxy. By extracting high-dimensional token essences and classifying them with a convolutional neural network, our model achieves high accuracy even without knowledge of the generation model. Furthermore, it does not require fine-tuning and effectively utilizes the multilingual knowledge embedded in the LLM's pretraining, making it a simple, efficient solution for detecting machine-generated text in both English and Arabic.

In Subtask A, our model achieved an F1 score of **0.999**, ranking **first** among 26 teams. In Subtask B, we obtained an F1 score of 0.965, securing fourth place among 22 teams. **In short, our contributions are as follows**: (1) Utilizing the last-layer essences of proxy LLMs as features enhances detection performance. (2) The scale of the proxy LLMs does not significantly improve detection accuracy. (3) Proxy LLMs with broader multilingual knowledge exhibit higher detection accuracy.

*Corresponding author: Zhendong Mao.

Proceedings of the 1st Workshop on GenAI Content Detection (GenAIDetect), pages 290–298
January 19, 2025.

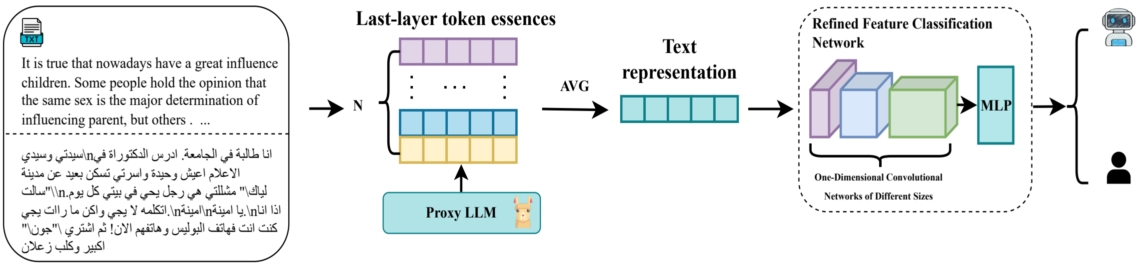

Figure 1: System Architecture

2 Related Work

Machine-generated text detection methods can generally be divided into two categories. The first category is zero-shot detection, where the simplest approach involves calculating the average log-likelihood of a text (Solaiman et al., 2019b), establishing a strong baseline for many zero-shot detection methods. More advanced techniques, such as DetectGPT (Mitchell et al., 2023) and its improved version Fast-DetectGPT (Bao et al., 2023), have shown that machine-generated text tends to fall within regions of negative probability curvature, effectively enabling machine-generated text detection. However, these approaches are often time-intensive and experience a significant performance drop when the generation model is unknown. The second category involves supervised detection methods. For instance, (Zhan et al., 2023) employed a fine-tuned RoBERTa-large (Liu et al., 2019) as a detector, but found it challenging to generalize effectively across different generation models. The T5-sentinel (Chen et al., 2023) addresses text detection by leveraging the next-token prediction capability of T5 (Raffel et al., 2023). Additionally, (Hu et al., 2023) introduced an iterative training process involving both a paraphraser and a detector, aiming to enhance robustness against paraphrasing attacks.

(Bhattacharjee and Liu, 2024) integrated the text to be detected into the prompt and directly asked ChatGPT whether the text is machine-generated or human-written, which is similar to our method, as both approaches leverage LLMs. However, our method does not directly inquire whether a text is machine-generated using LLM, nor does it require fine-tuning the LLM. Instead, it harnesses the high-dimensional multilingual representation capabilities of Llama-3.1-8B and the text is simply input into Llama-3.1-8B to extract token essences (refer to the last layer hidden states) as features, which are then fed into a classifier for final classification.

3 System Overview

To obtain a meaningful representation for the input text, we feed it into a proxy LLM, Llama-3.1-8B (Dubey et al., 2024), to extract essences from the last layer of the proxy LLM and subsequently pass the average of the essences through the Refined Feature Classification Network (RFCN), the overall model structure is shown in Figure 1.

The original text to be detected is first tokenized, with shorter sequences padded and longer ones truncated to a maximum length of 1024 tokens, resulting in the tokenized sequence $\boldsymbol{x} = \{x_1, x_2, \ldots, x_n\}$, the procedure is as follows:

Token essences from the Proxy LLM The tokenized sequence $\boldsymbol{x}$ is input into the Llama-3.1-8B model, which supports text across multiple languages. As $\boldsymbol{x}$ passes through the proxy LLM, it generates hidden states for each token at each layer. We specifically focus on the last-layer token essences (hidden states) of the proxy LLM, which serve as the high-level representations of each token. These token essences encapsulate both their individual meanings and the broader context within the text. Here, the representation quality across different languages is consistent. Supplementary details can be found in the Appendix B. To derive a single representation $\boldsymbol{h}$ of the input text, we take the average of the essences across all n tokens.

Refined Feature Classification Network The averaged representation $\boldsymbol{h}$ is then input into the RFCN for classification. In the first stage, the CNN extracts relevant features from the input through three convolutional and pooling layers, progressively capturing more complex patterns information. In the second stage, the refined features are passed through three fully connected layers, where each layer fine-tunes the representations by learning complex relationships and interactions between

Team	F1
starlight	0.997
saehyunMa	0.993
Fsf	0.993
Team_1-800-SHARED-TASKS	0.990
tesla	0.986
Baseline	0.478
CMI-AIGCX (ours)	**0.999**
w/o LLM	0.673
w/o RFCN	0.982

Table 1: Top: performance on English track. Bottom: ablation study about LLM and RFCN.

Team	F1
msmadi	**0.984**
Team_USTC-BUPT	0.972
starlight	0.965
apricity	0.960
Team_AAST-NLP	0.957
Team_1-800-SHARED-TASKS	0.952
Baseline	0.461
CMI-AIGCX (ours)	0.965
w/o LLM	0.606
w/o RFCN	0.934

Table 2: Top: performance on Arabic track. Bottom: ablation study about LLM and RFCN.

features, ultimately outputting the class probabilities p. The detailed design concept can be found in the Appendix C. The model is trained by minimizing the cross-entropy loss.

4 Experimental setup

4.1 Datesets and Evaluation Metrics

Datasets The dataset consists of essays written by humans and generated by AI, with a specific example shown in Appendix A. The human-written essays were curated from the ETS Corpus of Non-Native Written English (Blanchard et al., 2014). For the AI-generated essays, the organizers used seven models, including GPT-3.5-Turbo (OpenAI, 2022), GPT-4o (OpenAI et al., 2024), GPT-4o-mini (OpenAI et al., 2024), Gemini-1.5 (Team et al., 2024), Llama-3.1 (Dubey et al., 2024), Phi-3.5-mini (Abdin et al., 2024), and Claude-3.5 (Anthropic, 2024), to generate academic essays. The detailed data distribution is provided in Tables 5 and 6 in Appendix E.

Evaluation Metrics For both Subtask A and Subtask B, the primary evaluation metric is macro-F1, calculated as the harmonic mean of precision and recall.

4.2 Training

We utilize Llama as the proxy LLM for obtaining token essences, with the maximum length set to 1024. For the CNN, the input channel is set to 1, where three convolutional layers are employed, with the number of kernels being 32, 64, and 96 respectively. The sizes of their corresponding kernels are 24, 16, and 8. More details are provided in Appendix D.

5 Results

In this section, we present the results of our final submission to demonstrate the effectiveness of our approach, comparing our system's performance with that of several top-performing teams, and highlight key insights from our analysis.

5.1 Subtask A: English track

A total of 26 teams participated in the English track. Due to space constraints, this paper compares and analyzes the systems of several notable teams, including starlight, saehyunMa, Fsf, Team_1-800-SHARED-TASKS, and tesla. The official results are presented in Table 1. Our system achieved an accuracy, recall, and F1 score of 99.9%, securing first place in the official rankings. This outstanding performance underscores the significant superiority and effectiveness of our approach in the detection of machine-generated English texts.

5.2 Subtask B: Arabic track

A total of 22 teams participated in the Arabic track of the competition. This paper only compares and analyzes the systems of selected teams, including msmadi, Team_USTC-BUPT, starlight, CMI-AIGCX (ours), apricity, Team_AAST-NLP, and Team_1-800-SHARED-TASKS. According to the official results (as shown in Table 2), Our system achieved an F1 score of 96.5%, ranking fourth. This result highlights that our approach excels not only in detecting machine-generated English texts but also proves highly effective for Arabic texts, underscoring its robust cross-lingual applicability and efficiency.

5.3 Ablation Study

We conducted a comprehensive ablation experiment to separately assess the effectiveness of LLM token essences and RFCN components within our model. The experimental outcomes, presented in Tables 1 and 2, reveal significant insights. When LLM token essences were excluded and tokens from the XLM-RoBERTa (Solaiman et al., 2019a) were directly input into the RFCN, the F1 scores for Subtasks A and B declined to 67.3% and 60.6%, respectively. This suggests that the multilingual knowledge encoded in LLM token essences during pretraining provides superior feature representations for detecting machine-generated text. Additionally, substituting the RFCN with an MLP resulted in F1 scores of 98.2% and 93.4% for Subtasks A and B, respectively. This underscores the capability of CNNs to capture local dependencies and recognize repetitive patterns across different positions in the text—essential features that enable the RFCN to effectively integrate token essences across entire text sequences. These findings substantiate both the effectiveness and necessity of the components within our proposed approach.

5.4 Scale and Multilingual Knowledge of Proxy Model

We conducted extensive experiments using LLM of varying scales, including 8 billion and 70 billion parameters, and models with different levels of multilingual knowledge, such as Llama-2 and Llama-3.1, as proxy models for subtasks A and B.

The experimental results are presented in Tables 3 and 4. Notably, the Llama-3-8B model, despite being approximately one-tenth the size of Llama-2-70B, achieved F1 scores of 99.2% and 93.8% for Subtasks A and B, respectively, outperforming Llama-2-70B by 7.1% and 1.9%. When comparing Llama-3-8B to Llama-3-70B, despite the latter's larger scale, the performance improvement was marginal, with increases of only 0.2% and 1.4% for Subtasks A and B, respectively. These results suggest that the scale of the proxy model is not the primary determinant of performance in detecting machine-generated text.

Furthermore, when the proxy model was Llama-3.1-8B, the F1 score for subtask A was 99.9%, which was 7.8% higher than Llama-2-70B and 0.5% higher than Llama-3-70B. For subtask B, the F1 score was 96.5%, which was 4.6% more than Llama-2-70B and 1.3% more than Llama-3-70B.

Proxy Model	F1
Llama-2-70B	0.921
Llama-3-8B	0.992
Llama-3-70B	0.994
Llama-3.1-8B (ours)	**0.999**

Table 3: Performance on English track using different scale and multilingual knowledge of proxy model.

Proxy Model	F1
Llama-2-70B	0.919
Llama-3-8B	0.938
Llama-3-70B	0.952
Llama-3.1-8B (ours)	**0.965**

Table 4: Performance on Arabic track using different scale and multilingual knowledge of proxy model.

This indicates that the performance of multilingual machine-generated text detection is not solely dependent on the scale of the model but is significantly influenced by the richness of multilingual knowledge within the LLMs.

Upon further analysis, we found that Llama-2-70B's training data was primarily in English, which limits its multilingual capabilities. While Llama-3-8B and 70B were pre-trained on multilingual data, they were initially intended for English use. In contrast, the Llama-3.1 series was pre-trained on a corpus of 15 trillion multilingual tokens, making it a more effective proxy model for detecting machine-generated essays in both English and Arabic. More details are in Appendices F.1 and F.2.

6 Conclusion

This paper presents our approach and results for the GenAI Detection Task 2, where our system ranked first in the English track and tied for third in the Arabic subtask. We adopted an efficient strategy, using proxy LLM to generate fused token essences, which were then classified via a refined feature classification network. This method capitalizes on the multilingual representational capacity of LLMs without fine-tuning, enhancing performance in detecting machine-generated text. Our findings further underscore that proxy models with extensive multilingual knowledge markedly improve detection in multilingual contexts. Future work will explore the broader application of multilingual LLMs in language generation detection and investigate optimized strategies to leverage LLM token essences.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (No.62222212, 62121002).

Limitations

Given the limited number of languages in the dataset, we validated the effectiveness of our model only on machine-generated texts in English and Arabic. Future experiments will extend this validation to a broader range of languages. Our model has demonstrated outstanding performance on Llama-3.1-8B. Furthermore, an analysis of the results from Llama-3-8B and Llama-3-70B indicates that increasing the model size does not significantly improve performance, which is why we did not conduct experiments on Llama-3.1-70B. Moving forward, we plan to experiment with additional LLMs on more diverse datasets to determine which proxy LLM is most effective for detecting machine-generated texts. Since the official has not released the true labels of the test data, it is impossible to analyze the specific error cases. We will further optimize our results after the true labels of the test dataset are released.

References

Marah I Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Hassan Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Caio César Teodoro Mendes, Weizhu Chen, Vishrav Chaudhary, Parul Chopra, Allie Del Giorno, Gustavo de Rosa, Matthew Dixon, Ronen Eldan, Dan Iter, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Jamie Huynh, Mojan Javaheripi, Xin Jin, Piero Kauffmann, Nikos Karampatziakis, Dongwoo Kim, Mahmoud Khademi, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Chen Liang, Weishung Liu, Xihui (Eric) Lin, Zeqi Lin, Piyush Madan, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Corby Rosset, Sambudha Roy, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Xia Song, Olatunji Ruwase, Xin Wang, Rachel Ward, Guanhua Wang, Philipp Witte, Michael Wyatt, Can Xu, Jiahang Xu, Weijian Xu, Sonali Yadav, Fan Yang, Ziyi Yang, Donghan Yu, Chengruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yunan Zhang, and Xiren Zhou. 2024. Phi-3 technical report: A highly capable language model locally on your phone. Technical Report MSR-TR-2024-12, Microsoft.

Anthropic. 2024. Claude 3.5 sonnet.

Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. 2023. Fast-detectgpt: Efficient zero-shot detection of machine-generated text via conditional probability curvature. *arXiv preprint arXiv:2310.05130*.

Amrita Bhattacharjee and Huan Liu. 2024. Fighting fire with fire: can chatgpt detect ai-generated text? *ACM SIGKDD Explorations Newsletter*, 25(2):14–21.

Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. 2014. Ets corpus of non-native written english.

Yutian Chen, Hao Kang, Vivian Zhai, Liangze Li, Rita Singh, and Bhiksha Raj. 2023. Token prediction as implicit classification to identify LLM-generated text. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13112–13120, Singapore. Association for Computational Linguistics.

Shammur Absar Chowdhury, Hind Al-Merekhi, Mucahid Kutlu, Kaan Efe Keleş, Fatema Ahmad, Tasnim Mohiuddin, Georgios Mikros, and Firoj Alam. 2025. GenAI content detection task 2: AI vs. human – academic essay authenticity challenge. In *Proceedings of the 1st Workshop on GenAI Content Detection (GenAIDetect)*, Abu Dhabi, UAE. International Conference on Computational Linguistics.

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, 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.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *arXiv preprint arXiv:2301.07597*.

Xiaomeng Hu, Pin-Yu Chen, and Tsung-Yi Ho. 2023. Radar: Robust ai-text detection via adversarial learning. *Advances in Neural Information Processing Systems*, 36:15077–15095.

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. *Preprint*, arXiv:1907.11692.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature. In *International Conference on Machine Learning*, pages 24950–24962. PMLR.

OpenAI. 2022. Chatgpt: Optimizing language models for dialogue.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the limits of transfer learning with a unified text-to-text transformer. *Preprint*, arXiv:1910.10683.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, Miles McCain, Alex Newhouse, Jason Blazakis, Kris McGuffie, and Jasmine Wang. 2019a. Release strategies and the social impacts of language models. *Preprint*, arXiv:1908.09203.

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

Kaito Taguchi, Yujie Gu, and Kouichi Sakurai. 2024. The impact of prompts on zero-shot detection of ai-generated text. *arXiv preprint arXiv:2403.20127*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, and James Molloy. 2024. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Haolan Zhan, Xuanli He, Qiongkai Xu, Yuxiang Wu, and Pontus Stenetorp. 2023. G3detector: General gpt-generated text detector. *Preprint*, arXiv:2305.12680.

A Example of English and Arabic essays

We randomly select an essay from the English and Arabic datasets, as shown in Figures 2 and 3.

I disagree with the statement that the development of artificial intelligence will create more jobs than it will eliminate. While it is true that AI has the potential to automate certain tasks and improve efficiency, I believe that its development will ultimately lead to more job losses than gains.\n\nOne of the main reasons for this is that many jobs that are currently done by humans can be easily automated with the help of AI. For example, self-service kiosks have already replaced many cashiers in retail stores, and AI-powered customer service chatbots are becoming increasingly common. Additionally, many manufacturing jobs are being replaced by robots and other machines that can perform tasks faster and more accurately than humans. These job losses will not be offset by the creation of new jobs, as the demand for human workers in these areas will decrease.\n\nFurthermore, while AI may create some new jobs in fields such as AI development and deployment, these jobs will likely require highly specialized skills and education. Many people may not have the necessary skills to compete for these jobs, and therefore will not benefit from the development of AI. This could lead to a widening of the gap between the rich and the poor, as those who have the skills and education to work with AI will be better off than those who do not.\n\nIn conclusion, while AI has the potential to bring many benefits, I believe that its development will ultimately lead to more job losses than gains. As AI becomes more widespread, it is likely to automate many jobs that are currently done by humans, leading to significant unemployment. Therefore, policymakers and educators must take steps to prepare workers for the changing job market and to ensure that the benefits of AI are shared by all.

Figure 2: English machine-generated essay

قبل .حمله مع واشتركنا التصريح واخذنا وذهبنا النظام تخالف لا لكي تصريح يتطلب والحج الحرام البيت في الحج الى ذهبن
طويلا انتزرنا الحمله حافلة الى وصلنا ما نوعاً متعبا الاغراض تجهيز فكان واجهناها التي الصعوبات بعض هناك كان الذهاب
لانتظاره فنضطر المحدده بالمواعيد لايأتي البعض كان تأخرنا في الرئسي السبب كان وهذا متأخرن وصل البعض وكان
كانت سفوريه والاخر القرآن يقرأ البعض وكان جداً طويلا الطريق كان مكه الى وانطلقنا لتأخر أعذار لديه يكون والبعض
نفسي في كبيرة هذه الأجواء آثر كان الممتعه والفعاليات المسابقات بعض الطريق منتصف في وكان جميله ايمانية اجواء
فقط الحج من افضل بإنها قال وسلم عليه الله صلى والرسول العمرة الحجو كانت لان طواف فأخذنا محرمين مكه الى وصلنا
الى توجهنا ثم الغروب حتى فيها ومكثنا عرفه الى التاسع اليوم في ذهبنا ثم ومن الثامن اليوم حتى مكه في ومكثنا .فعله ما وهو
والمغرب العشاء وصلينا مزدلفه

Figure 3: Human-written Arabic essay

B Ensure consistent representation quality across different languages

The Llama-3.1-8B model is pretrained on a large-scale multilingual corpus, which enables it to learn the structures, syntactic patterns, and semantic relationships across a variety of languages. This multilingual training allows the model to generate token embeddings that capture both language-specific and language-independent features. Even though the model encounters tokens from different languages, it maps them into a shared embedding space, ensuring that semantically similar words are represented in a comparable way. This approach ensures consistent representation quality across different languages.

C The detailed design concept of the RFCN

The motivation behind designing the RFCN is to better leverage the local features of the text for classification, which are essential for distinguishing between human and machine-generated text. For the task of AI-generated text detection, the choice of three convolutional layers and specific kernel sizes (24, 16, 8) is aimed at effectively extracting text features. Using three convolutional layers allows for the extraction of progressively complex features from the text. In AI-generated text detection, this is crucial for capturing both simple language patterns and more complex syntactic structures and semantic information. Each layer's features enhance the model's ability to detect subtle differences in AI-generated text. The first kernel (24-sized) has a smaller receptive field, primarily capturing smaller local text patterns. The second kernel (16-sized) provides a medium receptive field, targeting phrase-level structural patterns. The last kernel (8-sized) features the largest receptive field, integrating more contextual information to focus on long-range dependencies. These specific kernel sizes and their corresponding receptive fields enable the model to extract features at multiple levels of granularity.

D Detailed Experimental Setup

We use the AdamW optimizer with a linear warmup decay learning schedule and a dropout of 0.1. The batch size and learning rate are set to 128 and 3e-4, and the model is trained for 20 epochs. During the training of our model, the training and validation datasets for Subtasks A and B were merged at a ratio of 19:1 to form new training and validation sets. We monitored the accuracy on the validation set to select the checkpoint with the best performance. The final training dataset consisted of the complete training and validation sets for each subtask, with the entire validation set evaluated after each training epoch. We selected the model that performed best on the validation set as the final model.

E Datasets

Datasets The detailed distribution of data categories in the dataset is as follows. The proportion of human and AI categories in the test set has not yet been disclosed, and as such, the table only presents the total number of samples in the test set. For a comprehensive breakdown of the data distribution, please refer to (Chowdhury et al., 2025).

	Train	Dev	Test
human	629	1235	
AI	1467	391	
Total	2096	1626	1129

Table 5: Dataset division of subtask A.

	Train	Dev	Test
human	1145	182	
AI	925	299	
Total	2070	481	293

Table 6: Dataset division of subtask B.

F Llama

In this section, we provide an overview of the pre-training corpora of Llama-2, Llama-3, and Llama-3.1, along with their intended purposes, which helps to explain the differences in their performance on multilingual tasks.

F.1 Llama-2

Llama-2 (Touvron et al., 2023), released by Meta in 2023, is an open-source suite of LLMs available in configurations of 7 billion (7B), 13 billion (13B), and 70 billion (70B) parameters. The model's pre-training involved approximately 2 trillion tokens, marking a 40% increase in data volume compared to Llama-1. These tokens were drawn from publicly accessible online sources, explicitly excluding data from the products or services of Meta. In addition to an expanded context window, increasing from 2,048 to 4,096 tokens, the 70B model also implemented Grouped-Query Attention (GQA) to enhance inference capabilities and computational efficiency. However, the pre-training corpus of Llama-2-70B is primarily in English, making it unsuitable for multilingual tasks.

F.2 Llama-3 and Llama-3.1

Llama-3 (Dubey et al., 2024) represents Meta's most recent advancement in LLM technology, launched in 2024 with parameter configurations of 8 billion (8B), 70 billion (70B), and later extended to 405 billion (405B) parameters in the Llama-3.1 series. Although Llama-3-8B and 70B were pre-trained on multilingual data, they were intended for commercial and research use in English, which made them more optimized for English-language tasks. In contrast, the Llama-3.1 series was pre-trained on a significantly larger corpus comprising approximately 15 trillion tokens (Dubey et al., 2024), far exceeding the corpus size of Llama-2. This expanded corpus includes data across a diverse set of over 30 languages, such as English, German, French, Italian, Portuguese, Hindi, Spanish, and Thai. Llama 3.1 is intended for commercial and research use in multiple languages, which we believe significantly enhances its adaptability to multilingual tasks when employed as a proxy model.