TurQUaz at GenAI Detection Task 1: Dr. Perplexity or: How I Learned to Stop Worrying and Love the Finetuning

Kaan Efe Keleş
TOBB University of Economics and Technology
kaanefekeles@etu.edu.tr

Mucahid Kutlu
Qatar University
mucahidkutlu@qu.edu.qa

Abstract

This paper details our methods for addressing Task 1 of the GenAI Content Detection shared tasks, which focus on distinguishing AI-generated text from human-written content. The task comprises two subtasks: Subtask A, centered on English-only datasets, and Subtask B, which extends the challenge to multilingual data. Our approach uses a fine-tuned XLM-RoBERTa model for classification, complemented by features including perplexity and TF-IDF. While perplexity is commonly regarded as a useful indicator for identifying machine-generated text, our findings suggest its limitations in multi-model and multilingual contexts. Our approach ranked 6th in Subtask A, but a submission issue left our Subtask B unranked, where it would have placed 23rd.

1 Introduction

The rapid proliferation of large language models (LLMs) has brought both remarkable advancements and significant challenges to the field of natural language processing (NLP). While these models enable unprecedented levels of fluency and coherence in generated text, their potential for misuse—ranging from generating misleading information to creating plagiarized content—necessitates robust detection methods. Distinguishing between human-written and machine-generated text has thus become a critical area of research, especially in multilingual and multi-model contexts.

Task 1 of the GenAI Content Detection shared tasks (Wang et al., 2025), addresses these challenges by focusing on the development of robust classifiers capable of identifying AI-generated text. This task is divided into two subtasks: Subtask A, which deals with English-only datasets, and Subtask B, which extends detection to multilingual datasets.

In this paper, we present our approaches for Subtask A and Subtask B. For Subtask A, we relied on a fine-tuned version of the XLM-RoBERTa model, achieving competitive performance. For Subtask B, we explored the integration of additional features such as perplexity and term frequency-inverse document frequency (TF-IDF). While perplexity has traditionally been considered a valuable metric for identifying machine-generated text (Varshney et al., 2020), we found its effectiveness limited in complex, multilingual scenarios.

Our approaches were ranked 6th place out of 35 participants in Subtask A. Unfortunately, due to a submission-related issue, our entry for Subtask B was not officially ranked. However, it would have placed 23rd out of 26 participants.

2 Related Work

The advance of LLMs has necessitated the development of robust methods for detecting machine-generated text. This has become a critical research area due to the potential misuse of LLMs for generating misleading or plagiarized content (Adelani et al., 2019; Pan et al., 2023). Early zero-shot detection methods, relying on statistical features like log probability and rank (Solaiman et al., 2019; Gehrmann et al., 2019; Ippolito et al., 2020), are computationally efficient but often lack robustness as models improve in generating text that closely resembles human writing (He et al., 2024).

More sophisticated zero-shot approaches have been proposed to address these shortcomings. DetectGPT (Mitchell et al., 2023) use the concept of probability curvature, comparing the log probability of a text with perturbed versions to identify machine-generated content.

Supervised detection, training classifiers on labeled data, has also been investigated (Uchendu et al., 2020; Guo et al., 2023). However, these methods suffer from limitations related to data requirements and generalization across domains and LLMs. Similarly, watermarking techniques

Proceedings of the 1st Workshop on GenAI Content Detection (GenAIDetect), pages 225–229
January 19, 2025.

(Kirchenbauer et al., 2023; Keleş et al., 2023), while promising, have shown vulnerability to paraphrasing attacks (Krishna et al., 2024; Sadasivan et al., 2023).

3 Proposed Approach

For both Subtask A and Subtask B, we fine-tuned the Facebook XLM-RoBERTa base model[1][2][3] (Conneau et al., 2020).

In Subtask A, we submitted the labels from the bare fine-tuned classifier, as we were unable to improve its performance using additional features including perplexity. For Subtask B, we aimed to enhance the classifier's performance by incorporating additional features including TF-IDF, the source language as a one-hot encoded feature, and perplexity values derived from the Llama 3.2 1B model (Dubey et al., 2024). These combined features were fed into an XGBoost classifier to improve overall performance. While our goal was to evaluate the effectiveness of perplexity as a predictive metric, we found it to be a suboptimal measure in this multi-source, multi-language context.

3.1 Perplexity

Perplexity (PPL) is a key metric used to evaluate how well a language model predicts a sequence of text, making it particularly useful for detecting machine-generated content. For a tokenized sequence $X = (x_1, x_2, \ldots, x_N)$, perplexity is calculated as:

$$\text{PPL}(X) = \exp\left(-\frac{1}{N}\sum_{i=1}^{N}\log P_\theta(x_i \mid x_{<i})\right)$$

where:

- N is the total number of tokens in the sequence.
- $P_\theta(x_i \mid x_{<i})$ is the model's predicted probability of token x_i given all preceding tokens $x_{<i}$.

Lower perplexity scores indicate that the text is more predictable according to the model. In the context of machine-generated text, since the generation process often samples tokens based on higher probability from the model's vocabulary, the output tends to include more likely tokens. This results in lower perplexity scores, which can suggest machine generation. On the other hand human-generated text often exhibits higher perplexity because humans do not always use the most statistically probable words; instead, they may choose words that are less predictable, adding creativity and nuance to the text. This higher perplexity reflects the diversity and unpredictability inherent in human language. Figure 1 shows the distribution of perplexity scores for AI-generated and human-written texts on the shared-task data.

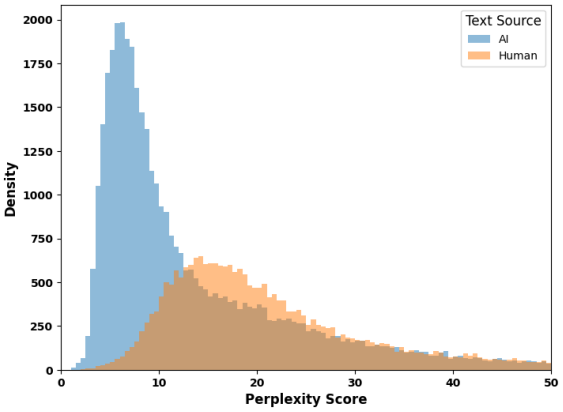

Figure 1: Distribution of perplexity scores for AI-generated (orange) and human-written (blue) texts for Subtask B's validation and test set combined.

Perplexity scores are influenced by various factors, including the specific language model, tokenization methods, and the language of the text. Models may assign higher perplexity to less common languages due to their under-representation in training data. While perplexity offers insights into text predictability, it has limitations as a sole indicator of text quality. Research indicates that perplexity is unreliable for evaluating text quality, as it can be affected by text length, repetition, and punctuation (Wang et al., 2022). These limitations also apply when using perplexity to detect machine-generated text.

4 Experiments

4.1 Perplexity as a Predictive Metric

As shown in Figure 1, Table 1, and Table 2, AI-generated texts generally exhibit lower perplexity scores compared to human-written texts. However, this trend varies significantly depending on

[1] https://huggingface.co/FacebookAI/xlm-roberta-base

[2] https://huggingface.co/keles/fine_tuned_xlm_roberta_for_en

[3] https://huggingface.co/keles/fine_tuned_xlm_roberta_for_mgtd2

the language and the source model. Although the distributions differ, there is no clear distinction or clustering between the two types of texts.

Lang	Type	Min	Max	Median	Mean
en	AI	1.20	31589.12	9.71	17.39
	Human	2.35	2235.15	19.29	28.17
zh	AI	1.35	1e5	11.74	240.60
	Human	3.69	1.4e6	57.74	1622.03
it	AI	4.39	53.35	9.56	12.54
	Human	-	-	-	-
de	AI	4.47	50.41	11.43	13.15
	Human	4.05	17.68	11.14	11.36
ur	AI	1.60	6.93	3.51	3.67
	Human	3.63	23.06	9.17	9.44
bg	AI	4.71	25.12	10.09	10.66
	Human	4.43	65.97	14.40	15.58
id	AI	4.46	17.07	7.07	7.39
	Human	5.58	41.96	13.04	13.80
ar	AI	7.23	53.82	15.29	18.04
	Human	14.05	46.83	21.08	23.27
ru	AI	4.44	42.49	11.39	12.89
	Human	4.00	31.47	8.27	9.73

Table 1: Perplexity statistics from the Llama-3.2-1B model across various languages, comparing AI-generated and human-written texts. Notably, Italian (it) lacks human-written text statistics due to the absence of such instances in the training set. Overall, AI-generated texts generally exhibit lower mean perplexity scores than human-written texts in most languages, with Russian (ru) and German (de) being notable exceptions.

Model	Min	Max	Median	Mean
human	2.35	1.4e6	19.79	148.00
gpt-3.5-turbo	1.60	132.66	7.35	9.27
gpt-35	2.15	1e5	6.03	132.77
davinci	1.35	4671.52	9.77	13.76
cohere	1.61	32.47	5.58	6.14
bloomz	1.52	140.70	12.06	13.05
text-davinci-003	2.36	348.87	8.42	10.90
mixtral-8x7b	2.37	12074.41	6.14	17.99
text-davinci-002	2.70	320.81	8.31	12.65
llama3-70b	2.56	102.10	6.63	9.40
gemma-7b-it	4.02	115.08	11.03	13.68
llama3-8b	1.99	55.92	7.01	9.93
gpt4o	2.68	21.80	7.33	7.52
gpt4	2.95	18.88	5.96	6.31

Table 2: Perplexity statistics from the Llama-3.2-1B model across different text generation models and human writing.

Due to space constraints, we have not detailed all the models tested for perplexity calculations, but our experiments revealed surprising findings about model size. Larger models, including those with up to 9 billion parameters, did not demonstrate any meaningful improvement in discrimination ability. These results suggest that model size may not be a critical factor in the effectiveness of perplexity-based detection methods.

4.2 Implementation Details

4.2.1 Subtask A

As can be inferred from Table 3, perplexity as an additional did not measurably improve the overall performance.

Methodology	F1 Score
XGBoost (FTC + TF-IDF + Perplexity)	0.974
XGBoost (FTC + TF-IDF)	0.973
FTC	0.969

Table 3: F1 Scores for Various Feature Combinations on the Validation Set for Subtask A. FTC refers to the labels produced by the fine-tuned classifier. When used with XGBoost, the probability assigned by the fine-tuned classifier to the positive label is included as a feature.

4.2.2 Subtask B

As can be inferred from Table 4, perplexity as an additional feature did not measurably improve the overall performance.

Methodology	F1 Score
XGBoost (FTC + TF-IDF + Perplexity)	0.972
XGBoost (FTC + TF-IDF)	0.972
FTC	0.966

Table 4: F1 Scores for Various Feature Combinations on the Validation Set for Subtask B. FTC refers to the labels produced by the fine-tuned classifier. When used with XGBoost, the probability assigned by the fine-tuned classifier to the positive label is included as a feature.

4.3 Dataset

4.3.1 Subtask A

This subtask focuses solely on English data sourced from various origins. Of this data, 62.5% is labeled as AI-generated, while 37.5% is labeled as human-written. The datapoints were randomly shuffled and then divided into 90% for training, 8% for testing, and 2% for validation.

The given train set was randomly shuffled and then divided into 90% for training, 8% for testing, and 2% for validation. We observed that the number of positive cases in the test-development set is much less than the one in the trainset. So we set the decision boundary to 0.97.

4.3.2 Subtask B

The train and the leaderboard test datasets for subtask B, as shown in Figures 2 and 3 respectively,

both exhibit significant language imbalances, albeit with notably different distributions. While the training set is dominated by English texts, the leaderboard test set presents an entirely different skew. We determined the language distribution of the test dataset using the langdetect library (Shuyo, 2010), which employs Naive Bayesian filtering for automatic language identification. The detection revealed Chinese (zh) as the dominant language with approximately 60,000 samples (39.8% of the dataset), followed by Urdu (ur) with 30,504 samples (20.1%) and Russian (ru) with 29,036 samples (19.2%). This substantial shift in language distribution between train and test sets, particularly the pivot from Indo-European languages in training to a test set dominated by Asian languages, presents an interesting challenge for fairly evaluating the model's cross-lingual generalization capabilities and robustness to language distribution shifts.

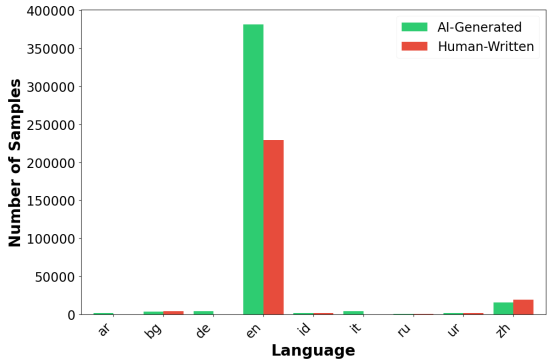

Figure 2: Distribution of AI-generated and human-written texts across different languages in the train dataset. The plot shows a clear imbalance in the dataset, with English (en) having the highest number of samples.

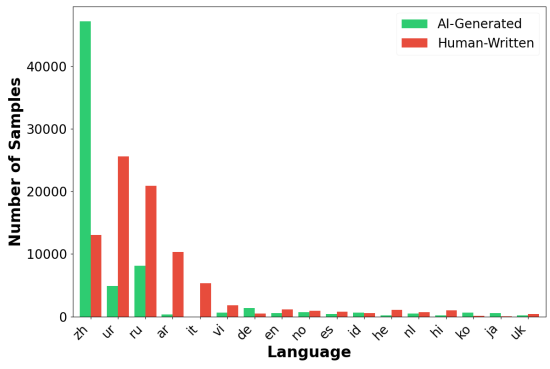

Figure 3: Distribution of AI-generated and human-written texts across different languages in the test dataset. The plot shows a clear imbalance in the dataset, with Chinese (zh) having the highest number of samples.

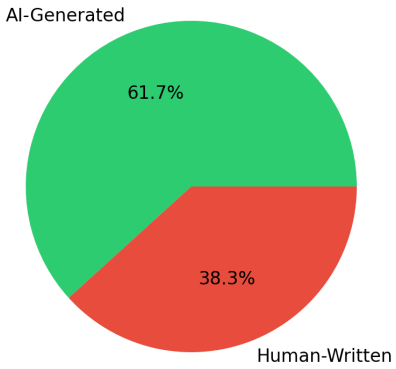

Figure 4: Distribution of AI-Generated vs Human-Written Samples for Subtask B

4.4 Results

The results of our submissions are summarized in Table 5. For Subtask A, which focuses on English-only datasets, our approach achieved an F1 score of 0.85 on the development set and 0.80 on the test set, reflecting a competitive performance. For Subtask B, encompassing multilingual data, the F1 scores were 0.65 and 0.64 on the development and test sets, respectively.

Subtask	Development (F1)	Test (F1)
Subtask A	0.85	0.80
Subtask B	0.65	0.64

Table 5: F1 scores for Subtask A and Subtask B on the development and test datasets, demonstrating the model's performance in detecting AI-generated versus human-written text.

5 Conclusions

In this paper, we presented our approach to the GenAI Content Detection shared tasks, focusing on distinguishing AI-generated from human-written text in both monolingual and multilingual contexts. Our primary findings indicate that while fine-tuned XLM-RoBERTa models can achieve competitive performance, the incorporation of additional features such as perplexity scores did not yield significant improvements in detection accuracy. This was particularly evident in the multilingual context, where perplexity's effectiveness varied considerably across different languages and source models.

References

David Ifeoluwa Adelani, Haotian Mai, Fuming Fang, Huy H. Nguyen, Junichi Yamagishi, and Isao

Echizen. 2019. Generating sentiment-preserving fake online reviews using neural language models and their human- and machine-based detection. *Preprint*, arXiv:1907.09177.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. *Preprint*, arXiv:1911.02116.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander M. Rush. 2019. GLTR: statistical detection and visualization of generated text. *CoRR*, abs/1906.04043.

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*.

Xinlei He, Xinyue Shen, Zeyuan Chen, Michael Backes, and Yang Zhang. 2024. Mgtbench: Benchmarking machine-generated text detection. *Preprint*, arXiv:2303.14822.

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, Online. Association for Computational Linguistics.

Kaan Efe Keleş, Ömer Kaan Gürbüz, and Mucahid Kutlu. 2023. I know you did not write that! a sampling based watermarking method for identifying machine generated text. *Preprint*, arXiv:2311.18054.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. 2023. A watermark for large language models. In *International Conference on Machine Learning*, pages 17061–17084. PMLR.

Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. 2024. Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense. *Advances in Neural Information Processing Systems*, 36.

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

Yikang Pan, Liangming Pan, Wenhu Chen, Preslav Nakov, Min-Yen Kan, and William Yang Wang. 2023. On the risk of misinformation pollution with large language models. *Preprint*, arXiv:2305.13661.

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

Nakatani Shuyo. 2010. Language detection library for java.

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. 2019. Release strategies and the social impacts of language models. *Preprint*, arXiv:1908.09203.

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

Lav R Varshney, Nitish Shirish Keskar, and Richard Socher. 2020. Limits of detecting text generated by large-scale language models. In *2020 Information Theory and Applications Workshop (ITA)*, pages 1–5. IEEE.

Yequan Wang, Jiawen Deng, Aixin Sun, and Xuying Meng. 2022. Perplexity from plm is unreliable for evaluating text quality. *arXiv preprint arXiv:2210.05892*.

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.