Overview of the 2023 ALTA Shared Task: Discriminate between Human-Written and Machine-Generated Text

Diego Molla
Macquarie University
diego.molla-aliod@mq.edu.au

Haolan Zhan
Monash University
haolan.zhan@monash.edu

Xuanli He
University College London
xuanli.he@ucl.ac.uk

Qiongkai Xu
The University of Melbourne
qiongkai.xu@unimelb.edu.au

Abstract

The ALTA shared tasks have been running annually since 2010. In 2023, the purpose of the task is to build automatic detection systems that can discriminate between human-written and synthetic text generated by Large Language Models (LLM). In this paper we present the task, the evaluation criteria, and the results of the systems participating in the shared task.

1 Introduction

The generative abilities of recent Large Language Models (LLMs) such as ChatGPT have shown impressive abilities in generating content with quality close to those generated by humans. Despite the possible advantages of LLMs, the concern about inappropriate utilization of these generated contents, accompanied by social and ethical issues, has been underscored in several preceding studies (Zellers et al., 2019; Aliman and Kester, 2021; Ranade et al., 2021; Xu et al., 2022).

Some of those LLMs are designed with watermarks (He et al., 2022; Kirchenbauer et al., 2023). However, there is also the possibility of deploying LLMs without watermarks. Consequently, effectively distinguishing texts by vanilla language models from the human-written text pieces has become an emerging and challenging task.

The goal of the 2023 ALTA shared task is to build automatic detection systems that can discriminate between human-written and text generated by LLMs. The text comes from a variety of sources and different LLMs.

Formally, this is a binary classification problem, as each candidate sentence can be generated either by human or a LLM. The evaluation metric is accuracy.

Section 2 presents related work. Section 3 details how the data have been gathered and labeled. Section 4 presents the evaluation framework. Section 5 describes a baseline that was made available to the participants. Section 6 lists the details of the participating systems and their results. Finally, Section 7 concludes this paper.

2 Related Work

The preliminary work for identifying machine-generated text involves feature-based approaches, such as utilizing linguistic patterns (Muñoz-Ortiz et al., 2023) and cues (Solaiman et al., 2019), e.g., bag-of-words. More recent work (Zellers et al., 2019) proposes to use detectors based on pre-trained language models. e.g., Liu et al. (2019) use RoBERTa as the basis of the detector. After a fine-tuning process, RoBERTa has been proven its prowess as a detector across multiple domains (Solaiman et al., 2019; Fagni et al., 2021; Rodriguez et al., 2022). To align with our research goals, we depart from the conventional assumption that detailed knowledge of synthetic data origin is readily available, which includes specifics about generative models, decoding strategies, and domains. In reality, such information often remains elusive.

It is worth noting several recent works on discriminating human- and machine-generated texts, e.g., OpenAI GPT-2 Detector (OpenAI, 2023), GPTZero (Tian and Cui, 2023), DetectGPT (Mitchell et al., 2023), DIPPER (Krishna et al., 2023) and G3-Detector (Zhan et al., 2023), which train their detectors on collected datasets with labeled human-written and machine-generated texts. Later on, a training-free detector DNA-GPT (Yang et al., 2023) was proposed to discover n-gram patterns in the machine-generated text.

Although some progress has been made in the corresponding task, its efficacy and reliability largely depend on the task settings, such as the domains of the generative tasks, the structures and scale of the generative models, etc. (Sadasivan et al., 2023) Kumarage et al. (2023) propose an assessment framework using evasive soft prompts,

Proceedings of The 21st Annual Workshop of the Australasian Language Technology Association, pages 148–152
November 29 - December 1, 2023 ©2023 Association for Computational Linguistics

and Chakraborty et al. (2023) further introduce AI detectability index as an evaluation metric for machine-generated text detection.

Related shared tasks include CLIN33[1], AuTextification[2] (Sarvazyan et al., 2023) , Detecting Generated Scientific Papers[3] (robodasha, 2022), and Machine Learning Model Attribution Challenge[4] (Merkhofer et al., 2023).

3 Data Gathering

The data for the 2023 ALTA shared task has been gathered from four generative benchmarks across multiple domains in the data. These comprise machine translation, and specifically the WMT (De-En) benchmark (Bojar et al., 2014), summarization, with CNN-DailyMail (CNNDM) (Nallapati et al., 2016), and language pre-training, including WikiData and the OpenwebText benchmark (Radford et al., 2019).

The human-written text are directly extracted from the ground-truth sentences in the above benchmarks. In contrast, the machine-generated text are produced by several widely-used generative models, all of which are GPT-based models. Specifically, these models contain GPT2-large, GPT3.5-turbo, and GPT4. We have used GPT2 model files through the Huggingface repository [5], and then fine-tuned these models on the aforementioned datasets. For the GPT3.5-turbo and GPT4 models, we use prompt-based text generation through the OpenAI API [6]. Specifically, we use the following prompts for different generative benchmarks:

Translation: `Please translate the following German sentence into English.`

Summarization: `Please summarize the following long paragraph with a short summary.`

Language Pre-training: `Please paraphrase the following sentence.`

The final data used in the 2023 ALTA shared task was selected by random sampling from the gathered data to ensure 50%-50% between human and machine-generated text (Table 1).

[1] https://sites.google.com/view/shared-task-clin33/home

[2] https://sites.google.com/view/autextification/home

[3] https://www.kaggle.com/competitions/detecting-generated-scientific-papers

[4] https://mlmac.io/

[5] https://huggingface.co/

[6] https://chat.openai.com/

Partition	Human (0)	Machine (1)	Total
Training	9,000	9,000	18,000
Development	1,000	1,000	2,000
Test	1,000	1,000	2,000

Table 1: Statistics of the data used in the 2023 ALTA shared task

4 Evaluation Framework

The evaluation framework was implemented as a CodaLab competition[7] with three phases.

In the **development phase**, labelled training and unlabelled development sets were made available. Participant systems could submit their system output on the development set up to 100 times, and the evaluation results were made public to all participating systems via a leaderboard.

In the **test phase**, an additional unlabelled test set was made available, and participating systems could make up to 3 submissions. The results of the test phase form a separate leaderboard and are used for the final ranking reported in this paper.

A third **unofficial submissions** phase has no end date and is available to all participant systems so that they can make additional submissions on the test data. These submissions form a separate leaderboard and are not used for the final ranking.

Table 1 shows the statistics of the three partitions.

5 Baseline

We formulate the detection framework as a binary classification task. Based on previous observations (Fagni et al., 2021; Rodriguez et al., 2022), RoBERTa has proven successful in various detection tasks. Therefore, to provide a starting point for participants, we provide the vanilla `RoBERTa-large` (Liu et al., 2019) as a baseline system[8]. Specifically, we use the corresponding checkpoint presented in Huggingface[9], which contains 354 million parameters. The performance of `RoBERTa-large` on the test set is 0.9765 in terms of accuracy.

[7] https://codalab.lisn.upsaclay.fr/competitions/14327

[8] https://github.com/zhanhl316/ALTA2023_shared_task

[9] https://huggingface.co/roberta-large

System	Category	Accuracy
OD-21	Student	**0.9910**
DetectorBuilder	Student	0.9845
AAST-NLP	Student	0.9835
SamNLP	Student	0.9820
Baseline		*0.9765*
VDetect	Student	0.9715
cantnlp	Student	0.9675
ScaLER	Student	0.9665
SynthDetectives	Student	0.9555

Table 2: Results of the 2023 ALTA shared task

6 Participating Systems and Results

A total of 9 teams submitted runs in the development phase, and 8 submitted in the test phase[10]. Table 2 shows the results of the baseline and the participating systems for the text phase.

The ALTA shared tasks have two categories, a student category where student members are not allowed to have completed a PhD degree and cannot be employed full time (with the exception of student supervisors), and an open category for those who are not eligible for the student category. However, this year (2023) only teams in the student category submitted in the test phase.

Tests of statistical significance[11] indicate that the difference between the first and the second team is statistically significant.

All of the participating systems that submitted a system description to us reported to have used LLMs in different ways, often as part of ensemble approaches, sometimes in addition to other approaches.

Team OD-21 (Gagiano and Tian, 2023) used Falcon-7B and label smoothing. They also used prompting techniques for samples with low confidence scores.

Team DetectorBuilder (Fang, 2023) used an ensemble with majority voting of BERT, RoBERTa, and DeBERTaV3.

Team AAST NLP (El-Sayed and Nasr, 2023) used an ensemble with majority voting of DistillBERT, XLMRoBERTa, and RoBERTa.

[10]Not all teams who submitted in the test phase had submitted in the development phase

[11]We conducted both McNemar's and Bootstrap tests using https://github.com/rtmdrr/testSignificanceNLP

Team SamNLP (Joy and Aishi, 2023) used a feature-level ensemble of DeBERTaV3 and XLM-RoBERTa, where these LLMs are jointly trained by concatenating their last layer and adding subsequent lineal layers.

Team VDetect (Liyanage and Buscaldi, 2023) experimented with various ensemble approaches using a varied range of models including several Transformer models, RNNs, and CNN, plus SVM and Naive Bayes.

Team SynthDetectives (Nguyen et al., 2023) used an ensemble of ALBERT, ELECTRA, RoBERTa, and XLNet, where the predictions of these LLMs are fed to a linear regression classifier.

7 Conclusions

The 2023 ALTA shared task focused on the discrimination between human-written text and machine-generated text. All systems submitting runs to the test phase had accuracy results over 0.95, and the baseline based on RoBERTa had an accuracy result of 0.9765. The top system submitted to the shared task had an accuracy of 0.9910, yet the difference with the second best system was statistically significant.

We were pleased to observe such good performance by the participants. This indicates that the task of identifying machine-generated text can be easy when used as a shared task like the one presented here. This task may become more difficult in the future as technology evolves.

References

Nadisha-Marie Aliman and Leon Kester. 2021. Epistemic defenses against scientific and empirical adversarial ai attacks. In *CEUR Workshop Proceedings, 2021 Workshop on Artificial Intelligence Safety*.

Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA. Association for Computational Linguistics.

Megha Chakraborty, SM Tonmoy, SM Zaman, Krish Sharma, Niyar R Barman, Chandan Gupta, Shreya Gautam, Tanay Kumar, Vinija Jain, Aman Chadha, et al. 2023. Counter turing test CT^2: AI-generated text detection is not as easy as you may think–introducing ai detectability index. *arXiv preprint arXiv:2310.05030*.

Ahmed El-Sayed and Omar Nasr. 2023. An ensemble based approach to detecting synthetic data generated by large language models. In *Proceedings of ALTA 2023, shared task section*.

Tiziano Fagni, Fabrizio Falchi, Margherita Gambini, Antonio Martella, and Maurizio Tesconi. 2021. Tweepfake: About detecting deepfake tweets. *Plos one*, 16(5):e0251415.

Yunhao Fang. 2023. Automatic detection of machine-generated text using pre-trained language models. In *Proceedings of ALTA 2023, shared task section*.

Rinaldo Gagiano and Lin Tian. 2023. A prompt in the right direction: Prompt based classification of machine-generated text detection. In *Proceedings of ALTA 2023, shared task section*.

Xuanli He, Qiongkai Xu, Lingjuan Lyu, Fangzhao Wu, and Chenguang Wang. 2022. Protecting intellectual property of language generation apis with lexical watermark. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10758–10766.

Saman Sarker Joy and Tanusree Das Aishi. 2023. Feature-level ensemble learning for robust synthetic text detection with DeBERTaV3 and XLM-RoBERTa. In *Proceedings of ALTA 2023, shared task section*.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. 2023. A watermark for large language models. *arXiv preprint arXiv:2301.10226*.

Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. 2023. Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense. *arXiv preprint arXiv:2303.13408*.

Tharindu Kumarage, Paras Sheth, Raha Moraffah, Joshua Garland, and Huan Liu. 2023. How reliable are ai-generated-text detectors? an assessment framework using evasive soft prompts. *arXiv preprint arXiv:2310.05095*.

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. *arXiv preprint arXiv:1907.11692*.

Vijini Liyanage and Davide Buscaldi. 2023. An ensemble method based on the combination of transformers with convolutional neural networks to detect artificially generated text. In *Proceedings of ALTA 2023*.

Elizabeth Merkhofer, Deepesh Chaudhari, Hyrum S. Anderson, Keith Manville, Lily Wong, and João Gante. 2023. Machine learning model attribution challenge.

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

Alberto Muñoz-Ortiz, Carlos Gómez-Rodríguez, and David Vilares. 2023. Contrasting linguistic patterns in human and llm-generated text. *arXiv preprint arXiv:2308.09067*.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

Duke Nguyen, Khaing Myat Noe Naing, and Aditya Joshi. 2023. Stacking the odds: Transformer-based ensemble for AI-generated text detection. In *Proceedings of ALTA 2023, shared task section*.

OpenAI. 2023. AI text classifier.

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

Priyanka Ranade, Aritran Piplai, Sudip Mittal, Anupam Joshi, and Tim Finin. 2021. Generating fake cyber threat intelligence using transformer-based models. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–9.

Yury Kashnitsky robodasha. 2022. Detecting generated scientific papers.

Juan Rodriguez, Todd Hay, David Gros, Zain Shamsi, and Ravi Srinivasan. 2022. Cross-domain detection of GPT-2-generated technical text. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1213–1233, Seattle, United States. Association for Computational Linguistics.

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

Areg Mikael Sarvazyan, José Ángel González, Marc Franco-Salvador, Francisco Rangel, Berta Chulvi, and Paolo Rosso. 2023. Overview of AuTexTification at IberLEF 2023: Detection and attribution of machine-generated text in multiple domains. *Procesamiento del Lenguaje Natural*, 71(0):275–288.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, and Jasmine Wang. 2019. Release strategies and the social impacts of language models. *CoRR*, abs/1908.09203.

Edward Tian and Alexander Cui. 2023. GPTZero: Towards detection of ai-generated text using zero-shot and supervised methods.

Qiongkai Xu, Xuanli He, Lingjuan Lyu, Lizhen Qu, and Gholamreza Haffari. 2022. Student surpasses teacher: Imitation attack for black-box NLP APIs. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2849–2860.

Xianjun Yang, Wei Cheng, Linda Petzold, William Yang Wang, and Haifeng Chen. 2023. DNA-GPT: Divergent n-gram analysis for training-free detection of GPT-generated text. *arXiv preprint arXiv:2305.17359*.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. In *Advances in Neural Information Processing Systems 32*.

Haolan Zhan, Xuanli He, Qiongkai Xu, Yuxiang Wu, and Pontus Stenetorp. 2023. G3Detector: General GPT-generated text detector. *arXiv preprint arXiv:2305.12680*.