Overview of the 2024 ALTA Shared Task: Detect Automatic AI-Generated Sentences for Human-AI Hybrid Articles

Diego Mollá and **Qiongkai Xu**
Macquarie University
Sydney, Australia
diego.molla-aliod@mq.edu.au
qiongkai.xu@mq.edu.au

Zijie Zeng and **Zhuang Li**
Monash University
Melbourne, Australia
zhuang.li1@monash.edu
zijie.zeng@monash.edu

Abstract

The ALTA shared tasks have been running annually since 2010. In 2024, the purpose of the task is to detect machine-generated text in a hybrid setting where the text may contain portions of human text and portions machine-generated. In this paper, we present the task, the evaluation criteria, and the results of the systems participating in the shared task.

1 Introduction

The advent of large language models (LLMs) has revolutionized artificial intelligence (AI), leading to a significant surge in AI-generated text and the rise of human-AI collaborative writing. While this collaboration offers exciting opportunities, it also introduces challenges — particularly in distinguishing between human-authored and AI-generated content within a single document. Although AI refers to various technologies, our focus in this shared task is specifically on the text generated by LLMs. Detecting such content has become essential not only as a deterrent against misuse but also as a safeguard, particularly in news reporting, journalism, and academic writing.

Previous efforts, such as the 2023 ALTA shared task (Molla et al., 2023), focused on corpus-level detection of AI-generated text, assuming that entire documents are either human-written or AI-generated. However, with the rise of human-AI collaborative writing, it is increasingly common for a single document to contain a mix of sentences authored by human and AI. Our proposed task addresses this realistic scenario by automatically identifying AI-generated sentences within hybrid articles.

Detecting AI-generated content at the sentence level is crucial for analyzing hybrid texts, which are becoming more prevalent in fields like news reporting, content marketing, and academic writing (Ma et al., 2023). Identifying AI-generated content at a finer granularity introduces a more nuanced challenge than distinguishing entirely AI-generated documents from those solely by human writers.

To tackle this challenge, our study leverages a newly available public dataset from Zeng et al. (2024b) and a private test set we collected for this shared task, both of which contain diverse and realistic hybrid articles. These datasets offer ideal benchmarks for exploring AI-generated text detection, as they include a mixture of human-written and AI-generated sentences across a range of topics within two key domains: academic writing and news reporting.

By examining the accuracy of identifying AI-generated sentences within texts that combine human and AI-authored content, we aim to develop more sophisticated and effective detection methods for collaborative writing scenarios. This work complements existing corpus-level detection efforts by offering a more comprehensive approach to understanding and identifying AI-generated content at different scales and contexts. The insights gained from this shared task will be valuable not only for preserving integrity in written communication but also for promoting transparency and responsibility in AI-assisted content creation.

The website of the 2024 ALTA shared task is https://www.alta.asn.au/events/sharedtask2024/.

2 Related Work

Recent advances in LLMs have created unprecedented challenges for content authenticity. Following the comprehensive related work presented by Zeng et al. (2024a), we examine how the ability of AI to generate human-like text raises significant concerns across multiple scenarios — from education and journalism to scientific research (Ma et al., 2023) and social media. While these technologies

Proceedings of The 22nd Annual Workshop of the Australasian Language Technology Association, pages 197–202
December 2-4, 2024 ©2024 Association for Computational Linguistics

offer tremendous benefits, they also present risks of academic dishonesty (Mitchell et al., 2023) and the potential spread of misinformation. Current detection approaches predominantly employ binary classification at the document level (Koike et al., 2024; Hu et al., 2024; He et al., 2023; Mitchell et al., 2023; Pagnoni et al., 2022; Rosati, 2022; Li et al., 2024). These methods assume the content is either entirely AI-generated or entirely human-written, an assumption that fails to reflect real-world usage patterns. As noted in emerging research (Dugan et al., 2023), modern content creation often involves human-AI collaboration, requiring more fine-grained detection approaches. A promising direction in hybrid text analysis has emerged, focusing on the identification of mixed authorship within documents. This approach draws inspiration from classical text segmentation techniques while addressing the unique challenges of AI text detection (Ghinassi et al., 2023; Xia and Wang, 2023). Recent work has explored both boundary detection methods (Zeng et al., 2024b; Lukasik et al., 2020; Yu et al., 2023; Xing et al., 2020; Li et al., 2022; Somasundaran et al., 2020; Koshorek et al., 2018) and more sophisticated approaches that integrate boundary identification with content classification (Bai et al., 2023; Lo et al., 2021; Gong et al., 2022; Tepper et al., 2012; Zeng et al., 2024a; Wang et al., 2023).

3 Data Description

For this shared task, we constructed a dataset comprising hybrid articles with mixed human-written and GPT-3.5-turbo-generated[1] content to facilitate the evaluation of AI-generated sentence detection methods.

Data Production. The training data was primarily sourced from the publicly available dataset curated by Zeng et al. (2024b), created via systematically replacing selected sentences in human-written articles with GPT-3.5-turbo-generated alternatives. For each sentence replacement, GPT-3.5-turbo was prompted to generate a contextually appropriate substitute that preserved the coherence and style of the original article.

Additionally, we expanded the dataset by generating hybrid articles from human-written news content sourced from the CC-News dataset (Hamborg et al., 2017). We randomly selected 3,000 articles with token lengths between 100 and 300 and tokenized them using the NLTK tokenizer[2]. Following the methodology outlined by Zeng et al. (2024b), we processed these articles by replacing selected sentences with GPT-3.5-turbo-generated content. For more details on the prompt format used, please refer to Zeng et al. (2024b).

Content Structure. Each hybrid news article includes a mix of human-written and GPT-3.5-turbo-generated sentences, with sentence-level authorship labels. We employed four distinct construction patterns to organize the human and machine-generated sentences, aligning with the methods in Zeng et al. (2024b):

- `h-m`: Human-written sentences followed by machine-generated sentences.
- `m-h`: Machine-generated sentences followed by human-written sentences.
- `h-m-h`: Human-written sentences, followed by machine-generated sentences, and then human-written sentences.
- `m-h-m`: Machine-generated sentences, followed by human-written sentences, and then machine-generated sentences.

Domain Focus. While the training data includes both academic and news domains, the evaluation exclusively targets sentence-level predictions in the news domain.

Table 1 presents the statistics of the training and test datasets.

4 Baselines

To establish baseline performance metrics for the task, we have implemented three approaches for AI-generated sentence detection:

- **Context-Aware BERT Classifier**: A fine-tuned BERT (Devlin et al., 2019) model that incorporates contextual information by processing three-sentence windows (the target sentence and one sentence before and after). These contextual embeddings are passed through a feed-forward neural network with a binary classification head for authorship prediction.

[1] https://platform.openai.com/docs/models/gpt-3-5-turbo

[2] https://www.nltk.org/api/nltk.tokenize.html

Dataset	Domain	Documents	Sentences	
			Human	Machine
Train	Academic	14,576	67,647	132,002
Train	News	1,500	4,574	8,571
Phase 1 Test	News	500	1,624	2,640
Phase 2 Test	News	1,000	3,310	5,342

Table 1: Statistics of the shared task datasets

- **TF-IDF Logistic Regression Classifier**: A logistic regression model trained on TF-IDF vectors computed from individual sentences. The model processes each sentence independently, using these statistical features to learn discriminative patterns between human-written and AI-generated text. This baseline has been made available to the shared task participants.[3]
- **Random Guess Classifier**: A naive approach that assigns authorship labels randomly, providing a lower bound for performance evaluation.

5 Evaluation Framework

5.1 Evaluation Setup

The evaluation was hosted as a CodaLab competition[4] with three phases.

- In phase 1 (“Development”), labelled training data was made available, together with a labelled test set to test the participant systems. The CodaLab page allowed each participant to submit up to 100 system runs based on the test set of phase 1. The evaluation results of this phase appeared in a leaderboard but were not used for the final ranking.
- In phase 2 (“Test”), a new unlabelled test set was made available. Each team could make up to 3 submissions, the evaluation results of which were used for the final ranking.
- Phase 3 (“Unofficial submissions”) was open after the end of phase 2, where participating systems can make up to 999 submissions of the output of the test set of phase 2 for final analysis. The evaluation results of phase 3 were not used for the final ranking. Phase 3 is open indefinitely, and new teams are encouraged to participate and compare their systems against the published results.

The labels of the test set used in phases 2 and 3 are not publicly available.

5.2 Evaluation Metrics

Participants are tasked with identifying the authorship of each sentence in a hybrid article A consisting of n sentences $\{s_1, s_2, \ldots, s_n\}$. Each sentence is either human-written or AI-generated. Formally, we define a function f that maps the hybrid article A to a sequence of predicted labels $\hat{L}$:

$$f(A) \rightarrow \hat{L}, \quad \text{where} \quad \hat{L} = \{\hat{l}_1, \hat{l}_2, \ldots, \hat{l}_n\} \tag{1}$$

Each label $\hat{l}_i$ indicates the predicted authorship of the corresponding sentence s_i, being either human-written (H) or AI-generated (A).

The performance is primarily evaluated using Cohen’s Kappa score, with accuracy serving as a supplementary metric.

Cohen’s Kappa Score. This robust statistic, which determines the final system rankings, measures inter-rater agreement while accounting for chance agreement:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{2}$$

where p_o is the observed agreement (accuracy), and p_e is the expected agreement by chance. The Kappa score effectively handles imbalanced datasets where one class may dominate, making it particularly suitable for evaluating detection performance across varying distributions of human-written and AI-generated content.

Accuracy. As a supplementary metric, we also report the proportion of correctly classified sentences across all test articles.

[3] https://github.com/altasharedtasks/ALTA_2024_demo

[4] https://codalab.lisn.upsaclay.fr/competitions/19633

The evaluation metrics have been implemented using scikit-learn functions `cohen_kappa_score` and `accuracy_score`.

6 Participating Systems and Results

As in previous years, there were two categories of participating teams:

- **Student**: All team members must be university students. No participating members can be full-time employees or have completed a PhD in a relevant field. The only exception is student supervisors.
- **Open**: Any other teams fall into the open category.

A total of 4 teams made submissions in the test phase, and the results are shown in Table 2. The Kappa score was used for the final ranking, while the Accuracy score is provided to facilitate comparisons with previous and future work. As shown in Table 3, all participating teams outperformed the logistic regression and random baselines, while two teams achieved better results than the BERT baseline.

The difference between the top team and second best is statistically different[5], so the winning team is “null-error”.

A brief description of the participating systems who provided their information follows.

Team Dima (Galat, 2024) used a 4-bit quantized LlaMA 3.1-8B-Instruct fine-tuned on domain-specific data. They also tested their system’s ability to handle automatic rewrites.

Team ADSN (Thomas et al., 2024) used an ensemble of lightweight classification methods inspired on traditional authorship attribution approaches.

7 Conclusions

This paper described a shared task for sentence-level detection of GPT-3.5-turbo-generated content within hybrid texts. By moving beyond traditional corpus-level detection to sentence-level analysis, this task addresses the practical challenges of identifying AI-generated sentences in collaborative writing scenarios. The multi-domain training approach, combined with a focused evaluation of news articles, provides a rigorous framework for developing and evaluating fine-grained detection methods. Through this shared task, we aim to establish benchmarks for sentence-level AI content detection and advance our understanding of the distinctive characteristics of human-AI collaborative writing.

[5] Tests of statistical significance were based on NcNemar test on the system outputs, using the tool provided by Dror et al. (2018).

References

Haitao Bai, Pinghui Wang, Ruofei Zhang, and Zhou Su. 2023. Segformer: a topic segmentation model with controllable range of attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 12545–12552.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker’s guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392, Melbourne, Australia. Association for Computational Linguistics.

Liam Dugan, Daphne Ippolito, Arun Kirubarajan, Sherry Shi, and Chris Callison-Burch. 2023. Real or fake text?: Investigating human ability to detect boundaries between human-written and machine-generated text. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 12763–12771.

Dima Galat. 2024. Advancing LLM detection in the ALTA 2024 shared task: Techniques and analysis. In *Proceedings of the 22nd Annual Workshop of the Australasian Language Technology Association*.

Iacopo Ghinassi, Lin Wang, Chris Newell, and Matthew Purver. 2023. Lessons learnt from linear text segmentation: a fair comparison of architectural and sentence encoding strategies for successful segmentation. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 408–418, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Zheng Gong, Shiwei Tong, Han Wu, Qi Liu, Hanqing Tao, Wei Huang, and Runlong Yu. 2022. Tipster: A topic-guided language model for topic-aware text segmentation. In *International Conference on Database Systems for Advanced Applications*, pages 213–221. Springer.

Team	Category	Kappa	Accuracy
Dima	Student	0.9416	0.9724
SamNLP	Student	0.9245	0.9642
Adventure Seeker	Student	0.8183	0.9163
ADSN	Open	0.6955	0.8548

Table 2: Results of participating systems on the phase 2 evaluation set.

Method	Kappa	Accuracy
Context-Aware BERT	0.8461	0.9294
Logistic Regression	0.5674	0.7973
Random Guess	0.0012	0.4973

Table 3: Results of baseline systems on the phase 2 evaluation set.

Felix Hamborg, Norman Meuschke, Corinna Breitinger, and Bela Gipp. 2017. news-please: A generic news crawler and extractor. In *Proceedings of the 15th International Symposium of Information Science*, pages 218–223.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTav3: Improving deBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

Beizhe Hu, Qiang Sheng, Juan Cao, Yuhui Shi, Yang Li, Danding Wang, and Peng Qi. 2024. Bad actor, good advisor: Exploring the role of large language models in fake news detection. In *Proceedings of the 38th AAAI Conference on Artificial Intelligence*, Vancouver, Canada.

Ryuto Koike, Masahiro Kaneko, and Naoaki Okazaki. 2024. Outfox: Llm-generated essay detection through in-context learning with adversarially generated examples. In *Proceedings of the 38th AAAI Conference on Artificial Intelligence*, Vancouver, Canada.

Omri Koshorek, Adir Cohen, Noam Mor, Michael Rotman, and Jonathan Berant. 2018. Text segmentation as a supervised learning task. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 469–473.

Raymond Li, Wen Xiao, Linzi Xing, Lanjun Wang, Gabriel Murray, and Giuseppe Carenini. 2022. Human guided exploitation of interpretable attention patterns in summarization and topic segmentation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10189–10204.

Zhuang Li, Yuncheng Hua, Thuy-Trang Vu, Haolan Zhan, Lizhen Qu, and Gholamreza Haffari. 2024. Scar: Efficient instruction-tuning for large language models via style consistency-aware response ranking. *arXiv preprint arXiv:2406.10882*.

Kelvin Lo, Yuan Jin, Weicong Tan, Ming Liu, Lan Du, and Wray Buntine. 2021. Transformer over pre-trained transformer for neural text segmentation with enhanced topic coherence. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3334–3340.

Michal Lukasik, Boris Dadachev, Kishore Papineni, and Gonçalo Simões. 2020. Text segmentation by cross segment attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4707–4716.

Yongqiang Ma, Jiawei Liu, and Fan Yi. 2023. Is this abstract generated by ai? a research for the gap between ai-generated scientific text and human-written scientific text.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature. In *Proceedings of the 40th International Conference on Machine Learning*, ICML.

Diego Molla, Haolan Zhan, Xuanli He, and Qiongkai Xu. 2023. Overview of the 2023 ALTA shared task: Discriminate between human-written and machine-generated text. In *Proceedings of the 21st Annual Workshop of the Australasian Language Technology Association*, pages 148–152, Melbourne, Australia. Association for Computational Linguistics.

Artidoro Pagnoni, Martin Graciarena, and Yulia Tsvetkov. 2022. Threat scenarios and best practices to detect neural fake news. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1233–1249.

Domenic Rosati. 2022. SynSciPass: detecting appropriate uses of scientific text generation. In *Proceedings of the Third Workshop on Scholarly Document Processing*, pages 214–222, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Swapna Somasundaran et al. 2020. Two-level transformer and auxiliary coherence modeling for improved text segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7797–7804.

Michael Tepper, Daniel Capurro, Fei Xia, Lucy Vanderwende, and Meliha Yetisgen-Yildiz. 2012. Statistical section segmentation in free-text clinical records. In *Lrec*, pages 2001–2008.

Joel Thomas, Gia Bao Hoang, and Lewis Mitchell. 2024. Simple models are all you need: Ensembling stylometric, part-of-speech, and information-theoretic models for the ALTA 2024 shared task. In *Proceedings of the 22nd Annual Workshop of the Australasian Language Technology Association*.

Pengyu Wang, Linyang Li, Ke Ren, Botian Jiang, Dong Zhang, and Xipeng Qiu. 2023. SeqXGPT: Sentence-level AI-generated text detection. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1144–1156, Singapore. Association for Computational Linguistics.

Jinxiong Xia and Houfeng Wang. 2023. A sequence-to-sequence approach with mixed pointers to topic segmentation and segment labeling. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2683–2693.

Linzi Xing, Brad Hackinen, Giuseppe Carenini, and Francesco Trebbi. 2020. Improving context modeling in neural topic segmentation. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 626–636.

Hai Yu, Chong Deng, Qinglin Zhang, Jiaqing Liu, Qian Chen, and Wen Wang. 2023. Improving long document topic segmentation models with enhanced coherence modeling. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5592–5605.

Zijie Zeng, Shiqi Liu, Lele Sha, Zhuang Li, Kaixun Yang, Sannyuya Liu, Dragan Gašević, and Guanliang Chen. 2024a. Detecting ai-generated sentences in realistic human-ai collaborative hybrid texts: Challenges, strategies, and insights. *Proceedings of the 33rd International Joint Conference on Artificial Intelligence*.

Zijie Zeng, Lele Sha, Yuheng Li, Kaixun Yang, Dragan Gašević, and Guangliang Chen. 2024b. Towards automatic boundary detection for human-ai collaborative hybrid essay in education. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 22502–22510.