GenAI Content Detection Task 1: English and Multilingual Machine-Generated Text Detection: AI vs. Human

Yuxia Wang[1], Artem Shelmanov[1], Jonibek Mansurov[1], Akim Tsvigun[2,3], Vladislav Mikhailov[4], Rui Xing[1], Zhuohan Xie[1], Jiahui Geng[1], Giovanni Puccetti[5], Ekaterina Artemova[6], Jinyan Su[1, 10], Minh Ngoc Ta[9], Mervat Abassy[13], Kareem Elozeiri[11], Saad El Dine Ahmed[13], Maiya Goloburda[1], Tarek Mahmoud[1], Raj Vardhan Tomar[14], Alexander Aziz[15], Nurkhan Laiyk[1], Osama Mohammed Afzal[1], Ryuto Koike[7], Masahiro Kaneko[1,7], Alham Fikri Aji[1], Nizar Habash[1,8], Iryna Gurevych[1,12], Preslav Nakov[1]

[1]MBZUAI [2]Nebius AI [3]KU Leuven [4]University of Oslo [5]ISTI-CNR [6]Toloka AI
[7]Institute of Science Tokyo [8]New York University Abu Dhabi
[9]BKAI Research Center, Hanoi University of Science and Technology [10]Cornell University
[11]Zewail City of Science and Technology [12]TU Darmstadt [13]Alexandria University
[14]Cluster Innovation Center, University of Delhi [15]University of Florida
{yuxia.wang, artem.shelmanov, preslav.nakov}@mbzuai.ac.ae

Abstract

We present the GenAI Content Detection Task 1 – a shared task on binary machine generated text detection, conducted as a part of the GenAI workshop at COLING 2025. The task consists of two subtasks: Monolingual (English) and Multilingual. The shared task attracted many participants: 36 teams made official submissions to the Monolingual subtask during the test phase and 27 teams – to the Multilingual. We provide a comprehensive overview of the data, a summary of the results – including system rankings and performance scores – detailed descriptions of the participating systems, and an in-depth analysis of submissions.[1]

1 Introduction

The success and popularity of Large Language Models (LLMs) have led to the proliferation of generative artificial intelligence (GenAI) content, which is now widely applied across numerous aspects of daily life. However, this widespread adoption has brought several concerns to light, including challenges to the integrity of student assignments and the potential for fabricated content to mislead individuals (Wang et al., 2024d). As generative LLMs continue to advance rapidly, it is becoming increasingly difficult for humans to distinguish machine-generated content from authentic human-authored text. Consequently, developing effective methods to address these challenges is crucial. To this end, we propose a GenAI content detection task, with Task 1 focusing specifically on the detection of machine-generated text in both English and multilingual contexts. This task is the continuation of SemEval Shared Task 8 (Wang et al., 2024b). The new task introduces a broader range of languages and domains while incorporating updated generators that leverage the latest LLMs.

The task consists of two subtasks: **Monolingual (English) subtask A** and **Multilingual subtask B**. The data for the shared task covers various domains and LLM generators. The data for English subtask covers diverse domains, including peer reviews, student essays, scientific papers, news articles, social media, emails, speech content and so on, similar for multilingual subtask data, with the test set involving more than 8 domains. To construct the data for the shared task, we produced machine-generated texts (MGTs), using state-of-the-art LLMs, including GPT-4/4o, Mistral (Jiang et al., 2023), Llama-3.1 (Dubey et al., 2024), Vikhr-Nemo (Nikolich et al., 2024), Qwen-2 (Yang et al., 2024), etc. Multilingual subtask data encompasses 21 unique languages.

The task attracted 36 participants who made official submissions during the test phase for the monolingual subtask A and 27 participants who made official submissions to the multilingual subtask B.

2 Related Work

This section discusses prior work about machine-generated text detection methods, datasets and shared tasks.

2.1 Detection Methods

There are mainly two commonly used approaches for detecting machine-generated text, training-free and training-based. Training-free detection

[1] https://github.com/mbzuai-nlp/COLING-2025-Workshop-on-MGT-Detection-Task1

Proceedings of the 1st Workshop on GenAI Content Detection (GenAIDetect), pages 244–261

methods leverage statistical characteristics of texts for identifying MGTs (Solaiman et al., 2019; Gehrmann et al., 2019). Various features have been explored, such as perplexity (Vasilatos et al., 2023), perplexity curvature (Mitchell et al., 2023), log rank (Su et al., 2023), intrinsic dimensionality (Tulchinskii et al., 2024) and N-gram analysis (Yang et al., 2023). Revise-Detect hypothesizes that machine-generated texts would be edited less by LLMs than human-written texts (Zhu et al., 2023). Binoculars (Hans et al., 2024) employs two LLMs to calculate the ratio of perplexity to cross-perplexity, assessing how one LLM responds to the next token predictions of another. Training based detectors typically fine-tune a pre-trained model for binary classification (Yu et al., 2023; Zhan et al., 2023), utilizing techniques such as adversarial training (Hu et al., 2023) and abstention (Tian et al., 2023). Verma et al. (2023) fine-tune a linear classifier on top of the learned representations.

2.2 Datasets

There are many efforts in detecting machine-generated text benchmarks. HC3 (Guo et al., 2023a) contains both Chinese and English text from ChatGPT. Other datasets such as MGTBench (He et al., 2023), ArguGPT (Liu et al., 2023) and DeepfakeTextDetect (Li et al., 2023) consider texts generated by various LLMs. M4 and M4GT-Bench(Wang et al., 2024d,a) are two comprehensive datasets covering multiple domains, languages and generators. MULTITuDE (Macko et al., 2023) includes texts in 11 languages, while the MAiDE-up dataset (Ignat et al., 2024) focuses on hotel reviews generated in 10 languages by GPT-4. MultiSocial (Macko et al., 2024) benchmarks MGT detection in the social media domain for 22 languages and 5 social media platforms.

2.3 Shared Tasks

Several shared tasks have been organized to address the problem of detecting machine-generated texts. *2023 ALTA shared task* (Molla et al., 2023) focuses specifically on identifying GPT-generated texts. *DAGPap22 shared task* (Chamezopoulos et al., 2024) targets the detection of machine-generated scientific papers. *SemEval 2024 shared task 8* (Wang et al., 2024b) introduced four subtasks: monolingual and multilingual binary classification (whether the text is generated by machine or written by human), multi-way classification distinguishing different generators, and human-machine text boundary detection, attracting participation from hundreds of teams.

There has been growing interest in detecting machine-generated text in non-English languages, such as Russian in *RuATD Shared task 2022* (Shamardina et al., 2022, 2024), Spanish in *IberLEF 2023* (Sarvazyan et al., 2023), and Dutch in *CLIN33* (Fivez et al., 2024). The multilingual detection task on SemEval-2024 Task 8 (Wang et al., 2024b) covers 9 languages, utilizing the M4GT-Bench dataset (Wang et al., 2024c).

3 Shared Task Description

3.1 Overview

The shared task was conducted in two phases: the development phase August 27, 2024 – October 29, 2024 and the test phase October 30 – November 4, 2024. During the training phase, the participants were given access to the texts and labels of the training and validation subsets, as well as to the texts of the dev-test subset. The dev-test set was made available to participants to evaluate the generalization capabilities of their detectors on distinct data during the development phase.

After the start of the test phase, we opened the labels of the dev-test and provided access to the texts of the test subset with a limited number of submission attempts to prevent leakage. After the finish of the test phase, we have released the labels of the test set, so the participants could perform some ablation studies.

As per the rules of the Task, participants were required to use only the data provided by the organizers to develop their models and were prohibited from utilizing any additional training data.

3.2 Datasets

The data for the Task is split into four subsets: training, development, dev-test, and test. Texts and labels for all subsets are publicly available at Github repository. Tables 1 and 2 present the descriptive statistics of the data.

3.2.1 Training and Development Sets

The training data for both English and multilingual subtasks was constructed using three large-scale multilingual machine-generated text datasets — HC3 (Guo et al., 2023b), M4GT-Bench (Wang et al., 2024c), and MAGE (Li et al., 2024). We merged all collected data, removed repeated texts,

Split	Source	Data License	#Generators	#Domains	Human	MGT	H+M	Total
Train	HC3	CC BY-SA-4.0	1	5	39,140	18,671	57,811	610,767
	M4GT	CC BY-SA-4.0	14	6	86,782	181,081	267,863	
	MAGE	Apache-2.0	27	14	103,000	182,093	285,093	
Dev	HC3	CC BY-SA-4.0	1	5	16,855	7,917	24,772	261,758
	M4GT	CC BY-SA-4.0	14	6	37,220	77,267	114,487	
	MAGE	Apache-2.0	27	14	44,253	78,246	122,499	
Dev-test	RAID	MIT	0	–	13,371	0	13,371	32,557
	LLM-DetectAIve	CC BY-SA-4.0	5	–	0	19,186	19,186	
Test	CUDRT	CC BY-SA-4.0	6	6	12,287	10,691	22,978	73,941
	IELTS	Apache-2.0	2	1	11,382	13,318	24,700	
	NLPeer	Apache-2.0	1	1	5,326	5,376	10,702	
	PeerSum	Apache-2.0	2	1	5,080	6,995	12,075	
	MixSet	CC BY-SA-4.0	7	9	600	2,886	3,486	
Total					375,296	603,727	979,023	**979,023**

Table 1: **English subtask** statistical information of training, development, dev-test, and test sets.

Split	Source	Data License	Lang	#Generators	#Domains	Human	MGT	H+M	Total
Train	HC3	CC BY-SA-4.0	zh, en	1	9	54,655	30,670	85,325	674,083
	M4GT	CC BY-SA-4.0	9	16	13	100,359	203,525	303,884	
	MAGE	Apache-2.0	en	27	14	102,954	181,920	284,874	
Dev	HC3	CC BY-SA-4.0	zh, en	1	9	22,981	12,718	35,699	288,894
	M4GT	CC BY-SA-4.0	9	16	13	42,886	87,591	130,477	
	MAGE	Apache-2.0	en	27	14	44,299	78,419	122,718	
Dev-test	MULTITuDE	GPL-3.0	11	8	–	7,992	66,089	74,081	74,081
Test	29 sources	–	15	19	–	73,634	77,791	151,425	151,425
Total						449,760	738,723	1,188,483	**1,188,483**

Table 2: **Multilingual subtask** statistics of training, development, dev-test, and test sets. M4GT includes 9 languages: en, de, id, it, zh, bg, ar, ur, ru. MULTITuDE includes 11 languages: de, en, uk, es, nl, ca, ru, pt, ar, zh, cs.

and randomly split into train and development sets by the ratio of 7:3. See detailed distribution over different languages, domains and generators in Appendix A.1.

3.2.2 Dev-Test Set

English Subtask A: we utilized 13,371 human-written texts from RAID (Dugan et al., 2024) and sampled 19,186 MGTs from LLM-DetectAIve (Abassy et al., 2024). The latter contains MGTs of three types: (i) fully MGTs, (ii) human-written and then machine-polished texts, and (iii) machine-generated and then machine-humanized texts.

Multilingual Subtask B: we sampled data from MULTITuDE (Macko et al., 2023) as the multilingual dev-test set.

3.2.3 Test Set

For the test set, in addition to leveraging MixSet (Zhang et al., 2024) and CUDRT (Tao et al., 2024), the majority of test sets is collected by our team, particularly multilingual subtask. Note that the dataset of CUDRT has not been released to the public before we used it as a subset of the test set.

English Subtask A uses Mixset and a subset of CUDRT. Based on the IELTS essays, we collected generations from *Llama3.1-70B-versatile* and *GPT-4o-mini*. We further generated academic paper peer reviews based on NLPeer and PeerSum, using *GPT-4o* and *GPT-4o-mini*.

Multilingual Subtask B: in addition to two datasets — we used Urdu fake news detection datasets generated by Ali et al. (2024), and sampled data from the CUDRT Chinese subset, the rest of multilingual test set was all newly collected, involving 27 different corpus and spanning 15 languages, with six of them are not seen in training, dev and dev-test sets. It covers Arabic, Chinese, Dutch, German, **Hebrew**, **Hindi**, Indonesian, Italian, **Japanese**, **Kazakh**, **Norwegian**, Russian, Spanish, Urdu, and **Vietnamese** (languages highlighted with the bold font were not seen in the training data).[2] See detailed distribution over sources,

[2]We included 15 languages in the training, dev and dev-test

Task	Set	Accuracy	F1
Subtask A	Dev	96.2	95.9 / 96.2
	Dev-Test	83.1	81.6 / 82.6
	Test	74.9	73.4 / 73.8
Subtask B	Dev	95.2	94.8 / 95.2
	Dev-Test	84.7	65.5 / 85.7
	Test	74.7	74.2 / 74.3

Table 3: Baseline performance on the Dev, Dev-Test, and Test sets for according to accuracy and macro F1.

domains, and models in Appendix A.2.

3.3 Baselines

Detector We fine-tuned pre-trained transformer-based models on the training sets as baselines. For *subtask A*, we fine-tuned RoBERTa, and XLM-R for *subtask B* to handle with multilingual data.

Fine-tuning was performed using the Hugging Face `Trainer` API with the following configuration: learning rate of 2×10^{-5}, batch size of 16 for both training and evaluation, weight decay of 0.1, and a total of 3 training epochs. Models were evaluated at the end of each epoch, and we keep the best model determined by development set performance, for the subsequent testing.

Results on Dev, Dev-test, and Test Sets Baseline results on the dev, dev-test, and test sets for both subtask A and B are demonstrated in Table 3. The baseline models showed strong performance on the development (dev) sets, particularly for subtask A, achieving high accuracy and F1-scores. However, performance declined on the dev-test and test sets, indicating potential overfitting or challenges in adapting to unseen data distributions.

For subtask B, the multilingual setting introduced additional complexity, as reflected in the relatively lower macro-average metrics, which emphasizes the difficulty of generalizing across multiple languages. These baseline results provide a reference point for participants and highlight the challenges of detecting machine-generated text, especially in multilingual contexts.

4 Participants' Submissions

In this section, we first describe ranking, macro-F1 and accuracy of participants, followed by a brief description of all submitted systems. We classify methods into (1) above vs. below the baseline, (2) black-box vs. white-box, (3) zero-shot vs. fine-tuning, (4) fine-tuning based on small models vs. large models, and (5) ensemble or not.

sets — Arabic, **Bulgarian**, **Catalan**, Chinese, Czech, Dutch, **English**, **German**, Indonesian, Italian, **Portuguese**, Russian, Spanish, **Ukrainian**, and Urdu.

Rank	Team	Macro-F1	Accuracy
1	Advacheck	83.07	83.11
2	Unibuc-NLP	83.01	83.33
3	Fraunhofer SIT	82.80	82.89
4	Grape	81.88	82.23
5	TechExperts(IPN)	81.53	81.81
6	TurQUaz	80.68	80.74
7	SzegedAI	79.10	79.29
8	AAIG	78.74	79.34
9	DCBU	77.13	78.01
10	Alfa	75.37	76.42
11	L3i++	74.63	75.54
12	LuxVeri	74.58	75.68
13	azlearning	74.14	75.17
14	honghanhh	73.94	75.14
–	**Baseline**	73.42	74.89
15	VX1291	72.93	74.83
–	cuettransform	72.32	73.16
16	rockstart	72.24	73.89
17	batirsdu	71.01	71.42
18	IPN-CIC	70.68	72.42
19	Ai-Monitors	70.57	72.65
20	semanticcuet	70.05	71.96
21	hmcgovern	68.48	69.51
22	abhirak0603	68.02	70.50
23	cnlpnitspp	65.02	68.76
24	mail6djj	64.66	68.46
25	bennben	63.32	67.48
26	saehyunma	62.80	67.25
27	yuwert777	62.14	66.69
28	seven	59.09	63.20
29	fangsifan	58.48	62.68
30	yaoxy	57.28	64.20
31	jojoc	54.16	60.37
32	dominikmacko	49.94	50.78
33	tropaleum	49.57	50.60
34	starlight1	47.57	56.65
35	nitstejasrikar	44.89	57.24

Table 4: **English subtask** leaderboard results. The main performance metric is macro-F1. Accuracy is used as an auxiliary performance metric.

To describe systems participating in the English and Multilingual subtasks separately, in the text we add the subscript **English:rank** to participants in the English subtask and the subscript **Multi:rank** to participants in the multilingual subtask. For example the team **Fraunhofer SIT** is ranked 3rd in the English subtask and referred to as **Fraunhofer SIT**$_{\textbf{English:3}}$ while it is ranked 10th in the Multilingual subtask and thus referred to as **Fraunhofer SIT**$_{\textbf{Multi:10}}$.

Team Name	Ranking	Small PLM	LLM	Feature Combination	Ensemble
Advacheck	1	✓			
Unibuc-NLP	2		✓		
Fraunhofer SIT	3	✓			
Grape	4	✓	✓		✓
TechExperts(IPN)	5		✓		
TurQUaz	6	✓	✓	✓	✓
SzegedAI	7	✓			✓
AAIG	8	✓			
DCBU	9	✓	✓	✓	✓
L3i++	11		✓		
LuxVeri	12	✓		✓	
IPN-CIC	18	✓			
Ai-Monitors	19	✓			

Table 5: **English subtask** participants overview.

4.1 English Subtask

4.1.1 Results and Rank

The English subtask attracted 36 submissions in total. Table 4 presents the complete rankings. The competition saw a remarkably tight race among the top performers with only 0.27 macro-F1 points separating the top three teams: **Advacheck$_{\mathbf{English:1}}$** (83.07), **Unibuc-NLP$_{\mathbf{English:2}}$** (83.01), and **Fraunhofer SIT$_{\mathbf{English:3}}$** (82.8). Interestingly, while the team **Advacheck$_{\mathbf{English:1}}$** secured the first place by the main metric, **Unibuc-NLP$_{\mathbf{English:2}}$** achieved a slightly higher accuracy (83.33 vs. 83.11), highlighting the razor-thin margins between top performers.

Fourteen teams outperformed the baseline (73.42 macro-F1) according to the main metric with scores varying from 83.07 to 44.89. The inability of most submissions to surpass the baseline underscores the complexity of the task.

4.1.2 System Description

Table 5 presents an overview of the English subtask participants' systems.[3]

Team Advacheck$_{\mathbf{English:1}}$ (Gritsai et al., 2025) develops a multi-task system with a shared Transformer Encoder (DeBERTa-v3-base) between several classification heads. This system includes a primary binary classification head and additional multiclass heads for text domain classification. The ablation studies show that multi-task learning outperforms single-task modes, with simultaneous tasks forming cluster structures in the embeddings space.

Team Unibuc-NLP$_{\mathbf{English:2}}$ (Teodor-George Marchitan, 2025) utilized both masked (XLM-RoBERTa-base) and causal language models (Qwen2.5-0.5B; Yang et al. (2024)),[4] with the Qwen-based classifier performing on par with Gritsai et al.. The authors report that LORA fine-tuning XLM-RoBERTa promotes a strong performance.

Team Fraunhofer SIT$_{\mathbf{English:3}}$ (Schäfer and Steinebach, 2025) combined an MGT detection adapter with a multi-genre natural language inference adapter over RoBERTa-base.

Team Grape$_{\mathbf{English:4}}$ (Doan and Inui, 2025), first, finetuned Llama-3.2-1B (Dubey et al., 2024) and gemma-2-2b (Team et al., 2024) for the MGT detection task. Second, they combined linguistic features with the model predictions by leveraging ensemble learning for more robust classification.

Team TechExperts(IPN)$_{\mathbf{English:5}}$ similar to Doan and Inui fine-tuned gemma-2b for the MGT detection task, which confirms the effectiveness of the small model in identifying the generated content.

Other teams ranked in top-20 developed the MGT detectors by (i) fine-tuning a model (**Team TurQUaz$_{\mathbf{English:6}}$**; Keleş and Kutlu, 2025; **Team AAIG$_{\mathbf{English:8}}$**; Bhandarkar et al., 2025; **Team IPN-CIC$_{\mathbf{English:18}}$**; Abiola et al., 2025; **Team Ai-Monitors$_{\mathbf{English:19}}$**; Singh et al., 2025); (ii) ensembling models and features (**Team SzegedAI$_{\mathbf{English:7}}$**; Kiss and Berend, 2025; **Team DCBU$_{\mathbf{English:9}}$**; Zhang et al., 2025; **Team Lux Veri$_{\mathbf{English:12}}$**; Mobin and Islam, 2025); and (ii) utilizing label supervision (**Team L3i++$_{\mathbf{English:11}}$**; Tran and Nguyen, 2025).

4.2 Multilingual Subtask

4.2.1 Results and Ranks

The multilingual subtask received 27 submissions with complete rankings demonstrated in Table 6.

The most notable feature of this subtask was the exceptional performance of the team "Grape", achieving macro-F1 score of 79.16, significantly outperforming other competitors. A substantial gap of 3.59 macro-F1 points between the winner and the second place "rockstart" (75.57) underscores the effectiveness of the "Grape" team approach to multilingual MGT detection.

In this subtask, only seven teams managed to surpass the baseline score of 74.16 with scores

[3]Top ranking teams that lack a system description do so because the authors did not submit a manuscripts and did not provide a short description of their system.

[4]https://qwenlm.github.io/blog/qwen2.5/

Rank	Team	Macro-F1	Accuray
1	Grape	79.16	79.62
–	jykim*	75.96	76.56
2	rockstart	75.57	75.64
3	Nota AI	75.32	75.91
4	LuxVeri	75.13	75.27
5	TechExperts(IPN)	74.63	74.74
6	azlearning	74.36	74.49
7	nampfiev1995	74.27	74.40
–	**Baseline**	74.16	74.74
8	starlight1	73.78	73.92
9	abit7431	72.65	73.48
10	Fraunhofer SIT	72.58	73.61
11	mail6djj	72.24	73.34
12	saehyunma	72.20	73.52
13	seven	71.40	72.00
14	jojoc	70.72	70.99
15	OSINT	70.67	71.87
16	yaoxy	69.54	71.51
17	VX1291	69.47	70.50
18	bennben	69.13	69.63
19	fangsifan	68.60	69.57
20	yuwert777	68.45	70.65
21	honghanhh	67.61	67.91
22	tmarchitan	66.29	67.11
–	keles	64.24	64.41
23	batirsdu	62.59	63.05
24	sohailwaleed2	52.53	52.59
25	dominikmacko	51.03	51.05

Table 6: **Multilingual subtask** leaderboard results. Submissions marked with "*" use additional training data and, therefore, are not incorporated in the ranking.

Team Name	Ranking	Small PLM	LLM	Feature Combination	Ensemble
Grape	1	✓	✓		✓
Nota AI	3	✓	✓	✓	✓
LuxVeri	4	✓			✓
TechExperts(IPN)	5		✓		
Fraunhofer SIT	10	✓			
OSINT	15	✓			

Table 7: **Multilingual subtask** participants overview.

ranging from 79.16 to 51.03. This indicates the increased difficulty of detecting MGT text among multiple languages simultaneously.

The overall lower scores in this subtask compared to the English subtask (top score 79.16 vs. 83.07) highlight the additional complexity introduced by multilingual detection and room for improvement.

4.2.2 System Description

Table 7 presents an overview of the multilingual subtask participants' systems.

Team Grape$_{\textbf{Multi:1}}$ (Doan and Inui, 2025), ranked 1 in the multilingual leaderboard, adopted two approaches in the task. They first separately fine-tuned small language models tailored to the specific subtask and then trained an ensemble model on top of them. Through evaluating and comparing these approaches, the team identified the most effective techniques for detecting machine-generated content across languages.

Team NotaAI$_{\textbf{Multi:3}}$ (Park et al., 2025) secured the third place in the task. They developed the system that addresses the challenge of detecting MGT in languages not observed during training, where model accuracy tends to decline significantly. The proposed multilingual MGT detection system employs a two-step approach: first, a language identification tool determines the language of the input text. If the language has been observed during training, the text is processed using a model fine-tuned on a multilingual PLM. For languages not seen during training, the system utilizes a model that combines token-level predictive distributions extracted from various LLMs with a meaning representation derived from a multilingual PLM.

Team LuxVeri$_{\textbf{Multi:4}}$ (Mobin and Islam, 2025) earned the 4th place. They utilized an ensemble of models, where weights are assigned based on each model's inverse perplexity to improve classification accuracy. The system combined RemBERT, XLM-RoBERTa-base, and BERT-base-multilingual-cased using the same weighted ensemble strategy. The results highlight the effectiveness of inverse perplexity-based weighting for robust detection of machine-generated text in both monolingual and multilingual settings.

Team TecExperts(IPN)$_{\textbf{Multi:5}}$ (Mehak et al., 2025) leveraged the gemma-2b model, fine-tuned specifically for the Shared Task 1 datasets to achieve strong performance.

Team L3i++$_{\textbf{Multi:7}}$ (Tran and Nguyen, 2025) studied a label-supervised adaptation configuration for LLaMA-as-a-judge for the task. In detail, they explore the feasibility of fine-tuning LLaMA with label supervision in masked and unmasked, unidirectional and bidirectional settings, to discriminate the texts generated by machines and humans in monolingual and multilingual corpora.

Other Systems The other systems explored various approaches, including exploring the integration of additional features such as perplexity and Tf-IDF (**Team TurQUaz$_{\textbf{Multi:22}}$**; Keleş and

Rank	All	MixSet	CUDRT	IELTS	PeerReview
1	83.1	48.0	67.1	89.9	97.2
2	83.3	66.7	75.9	82.6	94.1
3	82.9	58.9	71.0	88.8	92.1
4	82.2	64.7	73.2	79.1	97.4
5	81.8	59.2	72.7	80.8	95.5
6	80.7	47.2	72.6	78.1	96.9
7	75.7	54.9	71.0	63.1	97.2
8	79.3	62.3	75.4	69.0	97.2
9	78.0	60.0	74.6	66.3	96.9
10	76.4	59.8	75.5	64.2	93.2
11	75.5	60.9	70.3	66.9	92.5
12	75.7	56.6	74.0	61.9	95.2
13	75.2	62.8	70.8	65.3	92.2
14	75.1	66.6	72.8	62.7	92.2
BL	74.9	62.0	72.1	63.4	92.2
15	74.8	73.2	71.9	63.0	90.8
-	73.2	53.5	71.3	62.8	89.3
16	73.9	64.3	71.2	62.6	90.3
17	71.4	53.9	69.6	70.8	76.6
18	72.4	65.4	70.6	62.2	86.5
19	72.7	72.6	70.4	63.6	84.8
20	72.0	69.8	70.4	66.5	79.8
21	69.5	50.7	64.0	65.7	82.0
22	70.5	70.6	66.7	65.3	80.0
23	68.8	73.7	66.9	61.7	77.6
24	68.5	65.7	67.3	57.4	82.0
25	67.5	67.6	67.7	58.0	77.5
26	67.2	68.2	67.2	57.3	78.0
27	66.7	67.4	67.1	57.1	76.5
28	63.2	68.3	67.8	57.1	64.4
29	63.5	67.7	68.6	57.6	64.0
30	64.2	77.7	64.5	58.6	67.9
31	60.4	77.7	64.6	58.3	55.6
32	50.8	56.0	49.7	51.1	50.7
33	50.6	56.7	49.1	50.7	51.0
34	56.6	80.8	60.6	54.9	50.9
35	57.2	82.3	56.4	54.0	57.8

Table 8: **English subtask detection accuracy** across four domains.

Kutlu, 2025), finetuning models such as XLM-RoBERTa on the training set for the final evaluation, as incorporating adapter fusion led to worse results (**Team Fraunhofer SIT$_{\text{Multi:10}}$**; Schäfer and Steinebach, 2025), XML-R and mBERT models (**Team IPN$_{\text{Multi:9}}$**; Abiola et al., 2025 and QWen and RoBERTa models (**Team Unibuc-NLP$_{\text{Multi:22}}$**; Teodor-George Marchitan, 2025); and combining language-specific embeddings with fusion techniques to create a unified, language-agnostic feature representation (**Team OSINT$_{\text{Multi:15}}$**; Agrahari and Sanasam, 2025).

5 Analysis

Based on the test set, we analyze submitted systems by comparing the detection accuracy on (1) in-domain vs. out-of-domain, (2) seen vs. unseen languages, and (3) generations produced using normal prompts vs. prompts attempting to fill the gap between human and machine based on observations in manual annotations.

5.1 English In-domain vs. Out-of-domain

Results in Table 8 show the accuracy of 36 submitted systems across four component datasets in the English test set. Significant variance across domains reveals different generalization and robustness across detection systems.

Performance for in-domain datasets, such as IELTS and PeerReview, is generally higher than out-of-domain datasets MixSet and CUDRT. Top systems ranking 1-5 achieve scores around 80% on in-domain datasets. For example, top1 Team "Advacheck" scored 83.1% on IELTS essays and 89.9% on PeerReview. Moreover, accuracies are ≥90% for all teams above the baseline on PeerReview including the baseline itself. The consistently-high performance suggests that peer reviews (PeerRead) in the M4GT-Bench training set have effectively facilitated detectors in capturing domain-specific patterns during training, and thus generalizing well to similar-content PeerReview in the test set. For IELTS essays, the performance trend differs slightly from PeerReview. Despite student essays presented in the training set M4GT-Bench, only the first five teams managed to achieve scores ≥80%. This lies in the fact that essays sampled from OUTFOX in M4GT-Bench were written by English native speakers, while English is the second language for authors who attended the IELTS test. Subtle differences between essays in the training and test result in accuracy declines on the test set, which to some extent reveals the vulnerability of detectors against tiny distribution perturbations.

Out-of-domain dataset MixSet is the most challenging subset due to its varied and unseen content genres including game reviews, email, blog, and speech content. Top-ranked teams (ranks 1–5) experienced a substantial performance drop on MixSet — accuracy in the range of 48–66.7%. This may also attribute to the humanization and adaption of machine-generated text in MixSet. The former refers to modifying MGT to more closely mimic the natural noise that human writing always brings, introducing typo, grammatical mistakes, links, and tags. The latter refers to modifying MGT to ensure its alignment to fluency and naturalness to human linguistic habits without introducing any error expression. Detection systems struggle with highly heterogeneous and less structured data, which is exacerbated by the humanization and adaption operations of MGT in MixSet.

A surprising observation on MixSet is that all

Rank Size	All 151,425	News 57,590	Wiki 11,687	Essay 2,201	QA 24,854	Summary 13,600	Tweet 1,325	GovR 19,736	Other 4,214
1	79.6	65.1	80.2	99.3	98.9	70.0	94.5	87.0	84.2
2	75.6	64.0	87.1	81.0	91.9	79.1	100.0	69.1	48.2
3	75.9	60.7	81.0	97.7	96.2	65.2	72.0	81.7	91.1
4	75.3	60.7	87.9	91.0	93.2	71.7	98.9	75.2	58.6
BL	74.8	61.6	85.2	97.7	94.1	58.6	94.4	76.2	83.2
5	74.7	60.2	74.7	97.7	98.9	59.7	65.3	75.0	96.2
6	74.5	59.8	79.6	90.9	95.1	82.8	95.5	62.6	82.7
7	74.4	59.8	79.7	90.7	95.2	82.1	93.8	62.9	79.4
8	73.9	58.1	81.2	98.5	92.9	73.5	29.1	81.2	70.7
9	73.5	61.1	85.0	94.7	94.5	64.8	87.8	78.7	60.3
10	73.6	60.8	77.3	94.2	95.4	61.3	91.9	80.5	86.8
11	73.3	60.2	83.9	96.7	94.9	60.0	56.0	82.4	61.8
12	73.5	62.2	81.4	93.3	95.9	64.8	41.0	83.5	68.2
13	72.0	56.3	42.3	99.2	99.2	70.9	33.7	89.0	67.3
14	71.0	56.0	55.2	97.0	92.4	76.3	0.1	81.1	85.6
15	50.3	51.0	42.4	60.0	51.2	49.7	33.9	61.9	62.1
16	71.5	59.6	44.0	97.0	99.2	59.5	57.7	89.3	58.1
17	50.2	50.8	43.2	57.7	50.7	49.9	36.6	59.8	60.8
18	69.6	55.0	45.8	97.7	92.2	71.5	2.3	82.7	85.2
19	70.5	54.5	33.5	99.1	99.1	73.1	6.4	88.7	77.6
20	70.7	60.9	41.7	93.5	99.1	63.5	45.3	86.8	61.3
21	67.9	61.7	69.9	63.6	78.1	78.0	49.4	71.8	60.7
22	67.1	57.4	51.8	83.4	94.7	61.5	100.0	80.7	20.9
23	49.7	49.1	57.0	45.5	49.1	50.3	64.5	40.1	39.4
24	52.6	45.3	35.0	83.0	72.4	67.3	99.3	46.6	17.8
25	51.0	50.4	53.0	51.0	51.8	52.0	56.1	48.4	48.9

Table 9: **Multilingual subtask detection accuracy** across eight domains (Wiki: Wikipedia, GovR: GovReport).

teams above the baseline struggled to improve ≤5% compared to the baseline 62%, while 15 teams below the baseline achieved improvements ≥5%, with remarkable scores achieved by the last two teams — 80.8% and 82.3%, showing a stark contrast to their performance on other datasets.

Domains involved in CUDRT partially overlapped with the training data domains (e.g., news), while thesis is out of the training data though similar to academic papers, leading to the accuracy between Mixset and PeerReview. Most teams including the baseline scored between 65–75%, demonstrating moderate adaptability to this dataset.

5.2 Multilingual Subtask

We analyze submissions from three perspectives.

5.2.1 In-domain vs. Out-of-domain

We divided 29 sources across 15 languages into 8 domains: News, Wikipedia, Essay, question answering (QA), Summary, Tweet, government reports (GovReport), and others (e.g., poetry).

Table 9 presents the multilingual Subtask accuracy across 8 domains. In-domain datasets (News, Wikipedia, QA and Summary) consistently achieve higher accuracies due to their structured and training-aligned nature. Baseline accuracies for these domains are relatively strong, with significant improvements by the top-performing teams. Notably, the top-ranked team achieved peak performance of over 98% in QA, while the second-ranked team attained over 87% in Wikipedia. Though the genre of summary presented in the training data, they are English text. Summaries in the test set are Russian and Arabic, so summary domain posed notable challenges for detector, performing poorly across both baselines and team submissions. This underscores the difficulty of distinguishing machine-generated summaries from human-written ones in this domain.

Conversely, out-of-domain datasets (Essay, Tweet, GovReport, and Other) presented greater challenges, reflecting the systems' struggles to generalize to unseen styles or informal text. While structured datasets like essays and GovReport performed moderately well, with top-team accuracies exceeding 85%, informal and noisy domains such as tweets exhibited the lowest performance, with accuracies peaking at just 69.99%. This stark contrast highlights the need for more effective generalization strategies. Interestingly, we observed an anomaly in the tweet domain: two teams (ranked second and 22nd) achieved perfect accuracy (100%). This suggests that specialized

Rank Size	All 151,425	Fill-gap 32,487	Original 17,017	Others 101,921
1	79.6	91.1	94.2	73.5
2	75.6	75.9	84.0	74.1
3	75.9	89.7	92.2	68.8
4	75.3	81.5	86.9	71.4
BL	74.8	87.6	89.0	68.3
5	74.7	84.6	96.6	67.9
6	74.5	75.6	90.1	71.5
7	74.4	75.4	90.3	71.4
8	73.9	88.5	87.1	67.0
9	73.5	86.7	93.1	66.0
10	73.6	92.9	93.0	64.2
11	73.3	88.3	91.6	65.5
12	73.5	91.6	94.3	64.3
13	72.0	93.7	95.7	61.1
14	71.0	90.4	86.3	62.3
15	50.3	66.7	64.8	42.7
16	71.5	93.2	96.4	60.4
17	50.2	64.7	62.9	43.5
18	69.6	91.6	86.5	59.8
19	70.5	94.9	95.1	58.6
20	70.7	93.8	96.1	59.0
21	67.9	79.9	71.5	63.5
22	67.1	84.6	94.4	57.0
23	49.7	36.1	37.4	56.1
24	52.6	66.4	60.3	46.9
25	51.0	48.2	48.5	52.4

Table 10: **Multilingual subtask detection accuracy** between generations using original prompts vs. prompts aiming to fill the gap between human and machine, corresponding to columns of *Original* vs. *Fill-gap*. All is the whole multilingual test set.

approaches tailored to this domain can yield exceptional results, though these may involve overfitting to specific dataset patterns.

Overall, the results reveal a persistent gap between in-domain and out-of-domain performance, emphasizing the importance of domain adaptation and robust methods for handling unstructured or unseen data. At the same time, the findings demonstrate the potential for domain-specific optimizations in challenging contexts.

5.2.2 General Prompts vs. Improved Prompts

We compare system's accuracy results on text generated by ordinary prompts and the well-designed prompts that are used to fill the human and machine generations gap. MGTs using the improved prompts appear to make detection tasks more challenging. Our improved prompts aim to make machine-generated text more similar to human-written text by instructing LLMs how to generate human-like text and to avoid presenting distinguishable signals in formats, where these features were summarized from our observations in manual annotations in distinguishing human and machine text.

As shown in Table 10, in scenarios where detectors are tasked with identifying machine-generated text created using our improved prompts (Fill-gap in the Table 10), there is a noticeable decrease in accuracy compared to detecting machine-generated text created with the original prompts. This decline is particularly evident in higher ranks, with team 2 experiencing an 8% drop, team 5 a 12% drop, and teams 6 and 7 around a 15% drop. This decrease in performance suggests that the improved prompts, which were designed to narrow the gap between machine-generated and human-generated texts, may have inadvertently made the machine output too similar to human-like text, complicating the detector's ability to distinguish between the two. However, there are exceptions to this trend. Notably, team 8 (rank 8) and team 14 (rank 14) show improved results when using Fill-gap prompts, with accuracy increasing from 87.08% to 88.55% for team 8 and from 86.30% to 90.39% for team 14. This improvement may be due to a misalignment of features between their detector design and our improved machine-generated prompt design.

This suggests that we can learn from machine-generated examples to design better prompts that make the machine-generated text more natural and less detectable. However, it also exposes the vulnerability of detectors — they can be easily fooled when we adjust the prompts.

5.2.3 Seen Languages vs. Unseen Languages

Table 11 presents the detection accuracy on the multilingual subtask across 15 languages, including seen and unseen languages during the training process. The top-performing languages in terms of detection accuracy are generally those seen during training, with the highest accuracy observed on Chinese (94.2), followed by Russian (89.6) and Spanish (89.5). For Arabic (AR), Italian (IT), and Dutch (NL), the performance is slightly lower but still competitive, demonstrating the model's steady generalization to seen languages.

For unseen languages, such as Hindi (HI) and Hebrew (HE), there is a noticeable drop in performance compared to seen languages. For example, the top-performing team achieved only 51.8 on Hindi. It is challenging for models to generalize to unseen languages, due to the limited exposure to linguistic patterns, structures, and features during training. It is worth noting that some unseen

Rank Size	**All** 151,425	**ZH** 63,009	**UR** 30,505	**RU** 27,158	**AR** 10,670	**IT** 5,296	**KK** 2,471	**VI** 2,326	**DE** 1,865	**NO** 1,544	**ID** 1,200	**NL** 1,200	**ES** 1,200	**HI** 1,199	**HE** 1,182	**JA** 600
1	79.6	94.2	68.7	67.1	71.2	52.9	55.5	90.5	88.3	80.3	89.6	82.2	89.5	51.8	86.7	77.0
2	75.6	84.7	64.6	74.2	57.9	52.9	83.8	83.5	96.4	76.0	51.7	90.6	91.2	69.6	96.8	95.3
3	75.9	90.2	67.2	58.9	66.8	52.9	92.5	74.7	88.8	72.2	87.4	68.9	47.1	70.6	96.4	72.2
4	75.3	87.6	64.6	63.9	61.3	52.9	75.8	83.4	94.9	88.5	53.5	92.2	90.4	73.0	97.3	92.2
BL	74.8	87.3	68.4	55.3	68.4	52.9	82.8	85.3	85.2	69.8	68.2	92.5	90.5	71.3	89.3	90.0
5	74.7	90.1	64.1	56.0	69.1	52.9	62.9	87.6	59.6	69.8	93.8	81.0	90.4	69.1	96.5	95.0
6	74.5	84.2	65.0	67.9	66.8	52.9	47.5	81.8	93.5	83.2	83.9	85.9	88.9	69.1	89.8	78.2
7	74.4	84.4	64.9	67.7	65.4	52.9	47.5	82.0	92.2	85.8	83.4	85.4	89.2	68.8	90.1	75.2
8	73.9	88.3	58.7	67.0	58.4	52.9	93.0	65.9	89.6	61.6	50.5	80.7	88.0	61.4	82.7	61.2
9	73.5	85.1	67.0	59.8	60.8	52.9	90.6	87.2	82.8	78.2	48.7	78.0	83.1	54.5	89.6	74.3
10	73.6	86.0	67.6	56.0	69.1	52.9	86.8	80.4	65.0	52.8	73.8	87.4	85.4	63.5	85.7	86.0
11	73.3	87.4	63.4	58.2	55.6	52.9	89.4	79.7	87.0	66.6	73.9	82.1	87.4	70.5	93.3	79.5
12	73.5	85.3	68.0	61.5	54.3	52.9	92.7	62.0	87.8	63.7	80.3	85.3	86.3	63.0	86.2	59.5
13	72.0	93.2	55.4	63.3	55.4	52.9	93.0	65.9	5.2	25.8	71.2	50.2	50.0	61.4	1.7	61.2
14	71.0	87.0	54.3	68.7	61.2	52.8	54.7	63.8	77.1	54.7	49.7	57.1	64.9	53.5	0.0	52.0
15	50.3	50.9	52.0	49.0	53.0	50.4	52.1	49.7	33.9	33.2	49.7	50.3	50.7	50.4	32.1	50.0
16	71.5	91.3	62.4	55.5	53.7	52.9	89.4	79.7	5.3	28.9	79.9	50.2	50.0	70.3	1.9	79.5
17	50.2	50.6	51.4	49.3	52.8	50.1	52.2	50.1	35.9	34.5	49.3	50.3	50.2	50.6	34.2	53.3
18	69.6	87.4	54.5	63.8	61.1	52.9	55.7	57.0	58.2	23.1	50.3	55.2	59.3	53.7	0.0	54.3
19	70.5	92.2	51.6	65.5	56.5	52.8	54.7	63.8	4.2	23.8	70.6	50.1	50.0	53.5	0.0	52.0
20	70.7	87.6	65.6	58.3	52.0	52.9	92.7	62.0	5.0	28.2	81.7	50.2	50.0	63.0	1.9	59.5
21	67.9	71.9	51.7	80.1	55.3	78.3	48.1	63.8	93.8	82.1	72.4	83.5	84.7	52.3	31.7	63.8
22	67.1	82.5	61.5	55.3	45.8	52.9	94.2	71.6	12.0	27.9	57.5	63.3	73.6	53.5	20.3	57.2
23	49.7	49.2	48.4	50.7	47.4	49.0	50.3	49.7	65.5	63.5	50.4	51.1	49.2	51.9	64.5	52.0
24	52.6	60.7	45.7	58.9	28.8	52.9	47.5	48.1	5.8	39.8	47.7	49.5	51.2	46.0	5.8	27.0
25	51.0	51.1	49.9	51.5	50.8	50.1	50.1	52.3	55.9	54.5	52.5	54.0	49.9	52.4	53.7	52.0

Table 11: **Multilingual subtask detection accuracy** across 15 languages. Underlined languages were not present in the training data.

languages perform relatively well, such as Kazakh (KK) and Vietnamese (VI), achieving relatively high scores. This may result from knowledge transfer from similar languages to the unseen, like Russian to Kazakh, and Chinese to Vietnamese.

Overall, the models perform well on seen languages, and scores decline significantly on unseen languages.The dataset size and the nature of a language (e.g., script, structure, and linguistic features) play an important role in the model's ability to generalize.

6 Conclusion

In this work, we presented the dataset, baseline, participating systems and a detailed analysis across various detection methods for GenAI shared task 1: binary machine generated text detection. We explored both English and multilingual settings with diverse domains, LLM generators, and languages. All submitted systems show good performance on domains and languages that are seen during training, while witness the significant declines on unseen domains and languages. Moreover, detectors show remarkable vulnerability when machine-generated text is adapted to mimic humans, either by introducing typo, link, and tags, or by using fill-human-machine gap prompts. We expect our task can attract more researchers to develop robust and generalized detection models, and our analysis insights can provide a direction for future work, advancing research in machine-generated content detection.

Limitations

Despite providing a comprehensive dataset that spans multiple generators and domains and testing both English and Multilingual settings our study encounters several limitations that pave the way for future research.

Firstly, all the text samples (human and machine generated) used in this work come from existing open-source datasets and resources. While the sources of the test set have not been released prior to the conclusion of the challenge there is a limited possibility of data leakage. Participants were not allowed to use any external data and we trust they did not, however, pre-trained models could have seen part of the test set during their training and it would be impossible to know it.

Secondly, we don't have a detailed analysis of the differences between the datasets we joined together so that it is hard to understand if they have replicated or near-replicated samples and more in general how similar or not they are. In the future we will try to measure the performance of MGT detectors trained on the train set of one of these datasets when tested on each of the others to measure how close are the distributions of each pair of datasets among those we used.

Finally, we only look at binary classification tasks (human vs. machine) while it would be relevant to understand the performance of detectors in a multiclass classification scenario (human vs. machine1 vs. machine2 vs. ...), this would have been difficult to arrange correctly using the different datasets we have collected since isolating the specific versions of each model becomes harder over time (specifically with closed source ones) and therefore we avoided doing it. Future work should account for this scenario too.

Ethics and Broader Impact

This section outlines potential ethical considerations related to our work.

Data Collection and Licenses A primary ethical consideration is the data license. We reused pre-existing dataset, such as HC3, M4GT-Bench, MAGE, RAID, OUTFOX and LLM-DetectAIve, which have been publicly released for research purposes under clear licensing agreements. We adhere to the intended usage of all these dataset licenses.

Security Implications The dataset underpinning our shared task aims to foster the development of robust MGT detection systems, which are vital in addressing security and ethical concerns. These systems play a crucial role in identifying and mitigating misuse cases, such as preventing the spread of automated misinformation campaigns, which can undermine public discourse, and protecting individuals and organizations from potential financial losses through deceptive machine-generated content. In sensitive domains like journalism, academia, and legal proceedings, where the authenticity and accuracy of information are incredibly important, MGT detection is vital to maintaining content integrity and public trust. Beyond these fields, robust detection mechanisms contribute to the broader goal of promoting digital literacy by raising public awareness of the strengths and limitations of LLMs. This fosters a healthy skepticism towards digital content, encouraging users to critically evaluate the information they encounter.

Moreover, in multilingual contexts, detecting MGT becomes significantly more challenging due to the diversity of linguistic and cultural nuances. Advanced detection systems should address these complexities to prevent vulnerabilities, such as exploitation of less-resourced languages for disinformation. By ensuring the reliability of multilingual machine-generated content, these systems enhance global trust in AI technologies and protect against the security risks that arise from their misuse.

References

Mervat Abassy, Kareem Elozeiri, Alexander Aziz, Minh Ngoc Ta, Raj Vardhan Tomar, Bimarsha Adhikari, Saad El Dine Ahmed, Yuxia Wang, Osama Mohammed Afzal, Zhuohan Xie, Jonibek Mansurov, Ekaterina Artemova, Vladislav Mikhailov, Rui Xing, Jiahui Geng, Hasan Iqbal, Zain Muhammad Mujahid, Tarek Mahmoud, Akim Tsvigun, Alham Fikri Aji, Artem Shelmanov, Nizar Habash, Iryna Gurevych, and Preslav Nakov. 2024. LLM-DetectAIve: a tool for fine-grained machine-generated text detection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 336–343, Miami, Florida, USA. Association for Computational Linguistics.

Tolulope O. Abiola, Tewodros A. Bizuneh, Fatima Uroosa, Nida Hafeez, Grigori Sidorov, Olga Kolesnikova, and Olumide E. Ojo. 2025. Advancing multilingual machine-generated text detection: Insights from the coling workshop task. In *Proceedings of the 1st Workshop on GenAI Content Detection (GenAIDetect)*, Abu Dhabi, UAE. International Conference on Computational Linguistics.

Shifali Agrahari and Ranbir Sanasam. 2025. Team osint at genai detection task 1: Multilingual mgt detection:leveraging cross-lingual adaptation for robust llms text identification. In *Proceedings of the 1st Workshop on GenAI Content Detection (GenAIDetect)*, Abu Dhabi, UAE. International Conference on Computational Linguistics.

Muhammad Zain Ali, Yuxia Wang, Bernhard Pfahringer, and Tony Smith. 2024. Detection of human and machine-authored fake news in urdu. *arXiv preprint arXiv:2410.19517*.

Avanti Bhandarkar, Ronald Wilson, and Damon Woodard. 2025. Aaig at genai content detection task 1: Exploring syntactically-aware, resource-efficient small autoregressive decoders for ai content detection. In *Proceedings of the 1st Workshop on GenAI Content Detection (GenAIDetect)*, Abu Dhabi, UAE. International Conference on Computational Linguistics.

Savvas Chamezopoulos, Drahomira Herrmannova, Anita De Waard, Domenic Rosati, and Yury Kashnitsky. 2024. Overview of the dagpap24 shared task on detecting automatically generated scientific paper. In *Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024)*, pages 7–11.

Nhi Doan and Kentaro Inui. 2025. Grape at genai content detection task 1: Llm agents in multilingual machine-generated text detection. 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, et al. 2024. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*.

Liam Dugan, Alyssa Hwang, Filip Trhlík, Andrew Zhu, Josh Magnus Ludan, Hainiu Xu, Daphne Ippolito, and Chris Callison-Burch. 2024. RAID: A shared benchmark for robust evaluation of machine-generated text detectors. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12463–12492, Bangkok, Thailand. Association for Computational Linguistics.

Pieter Fivez, Walter Daelemans, Tim Van de Cruys, Yury Kashnitsky, Savvas Chamezopoulos, Hadi Mohammadi, Anastasia Giachanou, Ayoub Bagheri, Wessel Poelman, Juraj Vladika, et al. 2024. The clin33 shared task on the detection of text generated by large language models. *Computational Linguistics in the Netherlands Journal*, 13:233–259.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander M Rush. 2019. Gltr: Statistical detection and visualization of generated text. *arXiv preprint arXiv:1906.04043*.

German Gritsai, Anastasia Voznyuk, Ildar Khabutdinov, and Andrey Grabovoy. 2025. Advacheck at genai detection task 1: Ai detection powered by domain-aware multi-tasking. In *Proceedings of the 1st Workshop on GenAI Content Detection (GenAIDetect)*, Abu Dhabi, UAE. International Conference on Computational Linguistics.

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

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

Abhimanyu Hans, Avi Schwarzschild, Valeriia Cherepanova, Hamid Kazemi, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2024. Spotting llms with binoculars: Zero-shot detection of machine-generated text. *arXiv preprint arXiv:2401.12070*.

Xinlei He, Xinyue Shen, Zeyuan Chen, Michael Backes, and Yang Zhang. 2023. Mgtbench: Benchmarking machine-generated text detection. *arXiv preprint arXiv:2303.14822*.

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.

Oana Ignat, Xiaomeng Xu, and Rada Mihalcea. 2024. Maide-up: Multilingual deception detection of gpt-generated hotel reviews. *arXiv preprint arXiv:2404.12938*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Kaan Efe Keleş and Mucahid Kutlu. 2025. Turquaz at genai detection task 1: Dr. perplexity or: How i learned to stop worrying and love the finetuning. In *Proceedings of the 1st Workshop on GenAI Content Detection (GenAIDetect)*, Abu Dhabi, UAE. International Conference on Computational Linguistics.

Mihály Kiss and Gábor Berend. 2025. Beyond binary: Soft-voting multi-class classification for binary machine-generated text detection across diverse language models. In *International Conference on Computational Linguistics*, Abu Dhabi, UAE. Association for Computational Linguistics.

Yafu Li, Qintong Li, Leyang Cui, Wei Bi, Longyue Wang, Linyi Yang, Shuming Shi, and Yue Zhang. 2023. Deepfake text detection in the wild. *arXiv preprint arXiv:2305.13242*.

Yafu Li, Qintong Li, Leyang Cui, Wei Bi, Zhilin Wang, Longyue Wang, Linyi Yang, Shuming Shi, and Yue Zhang. 2024. MAGE: Machine-generated text detection in the wild. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 36–53, Bangkok, Thailand. Association for Computational Linguistics.

Yikang Liu, Ziyin Zhang, Wanyang Zhang, Shisen Yue, Xiaojing Zhao, Xinyuan Cheng, Yiwen Zhang, and Hai Hu. 2023. Argugpt: evaluating, understanding and identifying argumentative essays generated by gpt models. *arXiv preprint arXiv:2304.07666*.

Dominik Macko, Jakub Kopal, Robert Moro, and Ivan Srba. 2024. Multisocial: Multilingual benchmark of machine-generated text detection of social-media texts. *arXiv preprint arXiv:2406.12549*.

Dominik Macko, Robert Moro, Adaku Uchendu, Jason Lucas, Michiharu Yamashita, Matúš Pikuliak, Ivan Srba, Thai Le, Dongwon Lee, Jakub Simko, and Maria Bielikova. 2023. MULTITuDE: Large-scale multilingual machine-generated text detection benchmark. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9960–9987, Singapore. Association for Computational Linguistics.

Gull Mehak, Amna Qasim, Abdul Gafar Manuel Meque, Nisar Hussain, Grigori Sidorov, and Alexander Gelbuk. 2025. Tecexperts(ipn) at genai detection task 1: Ensuring content authenticity: Detecting ai-generated text in english and multilingual contexts.

In *Proceedings of the 1st Workshop on GenAI Content Detection (GenAIDetect)*, Abu Dhabi, UAE. International Conference on Computational Linguistics.

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.

Md Kamrujjaman Mobin and Md Saiful Islam. 2025. Luxveri at genai detection task 1: Inverse perplexity weighted ensemble for robust detection of ai-generated text across english and multilingual contexts. In *Proceedings of the 1st Workshop on GenAI Content Detection (GenAIDetect)*, Abu Dhabi, UAE. International Conference on Computational Linguistics.

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.

Aleksandr Nikolich, Konstantin Korolev, Sergei Bratchikov, Igor Kiselev, and Artem Shelmanov. 2024. Vikhr: Constructing a state-of-the-art bilingual open-source instruction-following large language model for Russian. In *Proceedings of the Fourth Workshop on Multilingual Representation Learning (MRL 2024)*, pages 189–199, Miami, Florida, USA. Association for Computational Linguistics.

Hancheol Park, Jaeyeon Kim, Geonmin Kim, and Tae-Ho Kim. 2025. Nota ai at genai detection task 1: Unseen language-aware detection system for multilingual machine-generated text. In *Proceedings of the 1st Workshop on GenAI Content Detection (GenAIDetect)*, Abu Dhabi, UAE. International Conference on Computational Linguistics.

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. *arXiv preprint arXiv:2309.11285*.

Karla Schäfer and Martin Steinebach. 2025. Fraunhofer sit at genai content detection task 1: Adapter fusion for ai-generated text detection. In *Proceedings of the 1st Workshop on GenAI Content Detection (GenAIDetect)*, Abu Dhabi, UAE. International Conference on Computational Linguistics.

Tatiana Shamardina, Vladislav Mikhailov, Daniil Cherniankii, Alena Fenogenova, Marat Saidov, Anastasiya Valeeva, Tatiana Shavrina, Ivan Smurov, Elena Tutubalina, and Ekaterina Artemova. 2022. Findings of the the ruatd shared task 2022 on artificial text detection in russian. *arXiv preprint arXiv:2206.01583*.

Tatiana Shamardina, Marat Saidov, Alena Fenogenova, Aleksandr Tumanov, Alina Zemlyakova, Anna Lebedeva, Ekaterina Gryaznova, Tatiana Shavrina, Vladislav Mikhailov, and Ekaterina Artemova. 2024. Coat: Corpus of artificial texts. *Natural Language Processing*, pages 1–26.

Azad Singh, Vishnu Tripathi, Ravindra Kumar Pandey, Pragyanand Saho, Prakhar Joshi, Neel Mani, Richa Alagh, Pallaw Mishra, and Piyush Arora. 2025. Ai-monitors at genai detection task 1: Fast and scalable machine generated text detection. In *Proceedings of the 1st Workshop on GenAI Content Detection (GenAIDetect)*, Abu Dhabi, UAE. International Conference on Computational Linguistics.

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

Jinyan Su, Terry Zhuo, Di Wang, and Preslav Nakov. 2023. Detectllm: Leveraging log rank information for zero-shot detection of machine-generated text. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12395–12412.

Zhen Tao, Zhiyu Li, Dinghao Xi, and Wei Xu. 2024. CUDRT: benchmarking the detection of human vs. large language models generated texts. *CoRR*, abs/2406.09056.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving Open Language Models at a Practical Size. *arXiv preprint arXiv:2408.00118*.

Liviu P. Dinu Teodor-George Marchitan, Claudiu Creanga. 2025. Team unibuc - nlp at coling-2025 task 1: Qwen it detect machine-generated text. In *Proceedings of the 1st Workshop on GenAI Content Detection (GenAIDetect)*, Abu Dhabi, UAE. International Conference on Computational Linguistics.

Yuchuan Tian, Hanting Chen, Xutao Wang, Zheyuan Bai, Qinghua Zhang, Ruifeng Li, Chao Xu, and Yunhe Wang. 2023. Multiscale positive-unlabeled detection of ai-generated texts. *arXiv preprint arXiv:2305.18149*.

Hanh Thi Hong Tran and Nam Tien Nguyen. 2025. L3i++ at genai detection task 1: Can label-supervised llama detect machine-generated text? In *Proceedings of the 1st Workshop on GenAI Content Detection (GenAIDetect)*, Abu Dhabi, UAE. International Conference on Computational Linguistics.

Eduard Tulchinskii, Kristian Kuznetsov, Laida Kushnareva, Daniil Cherniavskii, Sergey Nikolenko, Evgeny Burnaev, Serguei Barannikov, and Irina Piontkovskaya. 2024. Intrinsic dimension estimation for robust detection of ai-generated texts. *Advances in Neural Information Processing Systems*, 36.

Christoforos Vasilatos, Manaar Alam, Talal Rahwan, Yasir Zaki, and Michail Maniatakos. 2023. Howkgpt: Investigating the detection of chatgpt-generated university student homework through context-aware perplexity analysis. *arXiv preprint arXiv:2305.18226*.

Vivek Verma, Eve Fleisig, Nicholas Tomlin, and Dan Klein. 2023. Ghostbuster: Detecting text ghostwritten by large language models. *arXiv preprint arXiv:2305.15047*.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Osama Mohanned Afzal, Tarek Mahmoud, Giovanni Puccetti, Thomas Arnold, Alham Fikri Aji, Nizar Habash, Iryna Gurevych, and Preslav Nakov. 2024a. M4gt-bench: Evaluation benchmark for black-box machine-generated text detection. *CoRR*, abs/2402.11175.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Osama Mohammed Afzal, Tarek Mahmoud, Giovanni Puccetti, and Thomas Arnold. 2024b. SemEval-2024 task 8: Multidomain, multimodel and multilingual machine-generated text detection. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 2057–2079, Mexico City, Mexico. Association for Computational Linguistics.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Osama Mohammed Afzal, Tarek Mahmoud, Giovanni Puccetti, Thomas Arnold, Alham Aji, Nizar Habash, Iryna Gurevych, and Preslav Nakov. 2024c. M4GT-bench: Evaluation benchmark for black-box machine-generated text detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3964–3992, Bangkok, Thailand. Association for Computational Linguistics.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Chenxi Whitehouse, Osama Mohammed Afzal, Tarek Mahmoud, Toru Sasaki, Thomas Arnold, Alham Aji, Nizar Habash, Iryna Gurevych, and Preslav Nakov. 2024d. M4: Multi-generator, multi-domain, and multi-lingual black-box machine-generated text detection. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1369–1407, St. Julian's, Malta. Association for Computational Linguistics.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Xianjun Yang, Wei Cheng, Yue Wu, 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*.

Xiao Yu, Yuang Qi, Kejiang Chen, Guoqiang Chen, Xi Yang, Pengyuan Zhu, Weiming Zhang, and Nenghai Yu. 2023. Gpt paternity test: Gpt generated text detection with gpt genetic inheritance. *CoRR*.

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

Qihui Zhang, Chujie Gao, Dongping Chen, Yue Huang, Yixin Huang, Zhenyang Sun, Shilin Zhang, Weiye Li, Zhengyan Fu, Yao Wan, and Lichao Sun. 2024. LLM-as-a-coauthor: Can mixed human-written and machine-generated text be detected? In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 409–436, Mexico City, Mexico. Association for Computational Linguistics.

Zhaowen Zhang, Songhao Chen, and Bingquan Liu. 2025. Dcbu at genai detection task 1: Enhancing machine-generated text detection with semantic and probabilistic features. In *Proceedings of the 1st Workshop on GenAI Content Detection (GenAIDetect)*, Abu Dhabi, UAE. International Conference on Computational Linguistics.

Biru Zhu, Lifan Yuan, Ganqu Cui, Yangyi Chen, Chong Fu, Bingxiang He, Yangdong Deng, Zhiyuan Liu, Maosong Sun, and Ming Gu. 2023. Beat llms at their own game: Zero-shot llm-generated text detection via querying chatgpt. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7470–7483.

Appendix

A Dataset Distributions

A.1 Training and Development Sets

Tables 12 and 13 respectively demonstrate the statistical information of training and development sets across different sources of English and multilingual subtasls, and Table 14 shows the distribution over generators for datasets of HC3, M4GT-Bench and MAGE — the three component datasets of training and development sets for both English and multilingual subtasks.

Source	Sub-source	Training Set			Development Set		
		Human	Machine	Total	Human	Machine	Total
HC3	finance	2579	3189	5768	1113	1301	2414
	medicine	886	883	1769	352	380	732
	open_g	823	2339	3162	364	1015	1379
	reddit_tl5	34329	11680	46009	14781	4959	19740
	wiki_sai	523	580	1103	245	262	507
M4GT-Bench	arxiv	22484	30684	53168	9487	13003	22490
	outfox	2162	40973	43135	995	17390	18385
	peerread	3300	16169	19469	1398	6749	8147
	reddit	20353	32609	52962	8663	14076	22739
	wikihow	19454	35305	54759	8532	15168	23700
	wikipedia	19029	25341	44370	8145	10881	19026
MAGE	cmv	6020	16592	22612	2618	7026	9644
	cnn	265	0	265	131	0	131
	dialogsum	210	0	210	98	0	98
	eli5	15347	21849	37196	6451	9340	15791
	hswag	6806	19169	25975	2903	8085	10988
	imdb	269	0	269	107	0	107
	pubmed	273	0	273	105	0	105
	roct	6916	20008	26924	2930	8439	11369
	sci_en	6613	14390	21003	2891	6145	9036
	squad	14519	14875	29394	6333	6330	12663
	tldr	5558	15808	21366	2329	6930	9259
	wp	7919	21215	29134	3393	9390	12783
	xsum	6992	22129	29121	2925	9621	12546
	yelp	25293	16058	41351	11039	6940	17979
Grand Total		228922	381845	610767	98328	163430	261758

Table 12: **Monolingual subtask**: statistical information of training and development sets across different sources.

Source	Sub-source	Lang	Training Set			Development Set		
			Human	Machine	Total	Human	Machine	Total
HC3	baike	zh	2996	3211	6207	1247	1378	2625
	finance	en	2638	3135	5773	1054	1355	2409
		zh	1103	1393	2496	438	560	998
	law	zh	494	353	847	196	145	341
	medicine	en	874	901	1775	364	362	726
		zh	741	739	1480	317	327	644
	nlpcc_dbqa	zh	1155	2718	3873	543	1094	1637
	open_qa	en	840	2394	3234	347	960	1307
		zh	5212	2683	7895	2148	1117	3265
	psychology	zh	3546	773	4319	1505	309	1814
	reddit_eli5	en	34510	11776	46286	14600	4863	19463
	wiki_csai	en	546	594	1140	222	248	470
M4GT-Bench	Baike/Web QA	zh	4068	4099	8167	1629	1819	3448
	CHANGE-it NEWS	it	0	4174	4174	0	1843	1843
	News/Wikipedia	ar	344	1770	2114	150	756	906
		de	231	4462	4693	102	1957	2059
	RuATD	ru	684	630	1314	316	284	600
	True & Fake News	bg	4205	3886	8091	1795	1694	3489
	Urdu-news	ur	2085	1676	3761	853	720	1573
	arxiv	en	22508	30649	53157	9463	13038	22501
	id_newspaper_2018	id	1895	2081	3976	886	917	1803
	outfox	en	2196	40878	43074	961	17485	18446
	peerread	en	3291	16174	19465	1407	6744	8151
	reddit	en	20385	32535	52920	8631	14150	22781
	wikihow	en	19492	35187	54679	8494	15286	23780
	wikipedia	en	18975	25324	44299	8199	10898	19097
MAGE	cmv	en	6009	16476	22485	2629	7142	9771
	cnn	en	275	0	275	121	0	121
	dialogsum	en	197	0	197	111	0	111
	eli5	en	15214	21714	36928	6584	9475	16059
	hswag	en	6780	19163	25943	2929	8091	11020
	imdb	en	260	0	260	116	0	116
	pubmed	en	262	0	262	116	0	116
	roct	en	6820	19875	26695	3026	8572	11598
	sci-gen	en	6682	14308	20990	2822	6227	9049
	squad	en	14495	14914	29409	6357	6291	12648
	tldr	en	5526	15858	21384	2361	6880	9241
	wp	en	7941	21406	29347	3371	9199	12570
	xsum	en	6991	22202	29193	2926	9548	12474
	yelp	en	25502	16004	41506	10830	6994	17824
Grand Total			257968	416115	674083	110166	178728	288894

Table 13: **Multilingual subtask**: statistical information of training and development sets across different sources and languages.

Source	Model	Training Set		Development Set	
		Human	Machine	Human	Machine
HC3	gpt-35	0	18671	0	7917
	human	39140	0	16855	0
	bloomz	0	21061	0	8991
	cohere	0	20808	0	8896
	davinci	0	19345	0	8210
	dolly	0	8932	0	3931
	dolly-v2-12b	0	1938	0	831
	gemma-7b-it	0	12162	0	5240
	gemma2-9b-it	0	8366	0	3629
M4GT-Bench	gpt-3.5-turbo	0	25856	0	11005
	gpt4	0	9956	0	4300
	gpt4o	0	10374	0	4247
	human	86782	0	37220	0
	llama3-70b	0	12333	0	5181
	llama3-8b	0	12057	0	5290
	mixtral-8x7b	0	15865	0	6623
	text-davinci-003	0	2028	0	893
MAGE	13B	0	5385	0	2367
	30B	0	5769	0	2380
	65B	0	5815	0	2404
	7B	0	5083	0	2166
	GLM130B	0	4398	0	1842
	bloom$_{7b}$	0	5151	0	2201
	flan$_{5,\text{base}}$	0	6566	0	2887
	flan$_{5,\text{large}}$	0	6500	0	2893
	flan$_{5,\text{small}}$	0	6570	0	2811
	flan$_{5,\text{xl}}$	0	6429	0	2739
	flan$_{5,\text{xxl}}$	0	6532	0	2777
	gpt-3.5-turbo	0	15991	0	6682
	gpt$_{j}$	0	3468	0	1480
	gpt$_{\text{neox}}$	0	4734	0	2021
	human	103000	0	44253	0
	opt$_{1.3b}$	0	5553	0	2351
	opt$_{125m}$	0	5735	0	2469
	opt$_{3b}$	0	4988	0	2296
	opt$_{2.7b}$	0	5736	0	2586
	opt$_{30b}$	0	5637	0	2376
	opt$_{350m}$	0	5128	0	2252
	opt$_{6.7b}$	0	5642	0	2378
	opt$_{\text{iml}30b}$	0	6008	0	2619
	opt$_{\text{iml,max}1.3b}$	0	6176	0	2660
	t0$_{1b}$	0	6309	0	2620
	t0$_{3b}$	0	6602	0	2849
	text-davinci-002	0	14884	0	6359
	text-davinci-003	0	15304	0	6781
Grand Total		228922	381845	98328	163430

Table 14: Generator distribution over three component of training and development sets.

A.2 Test Sets

Table 15 shows the statistical distribution of English test sets in different domains and generators. Tables 16 and 16 present the distribution of the multilingual test set over different languages, domains and generators (see details).

Source / Domain	License	# Human	# MGT	LLM Generator List
CUDRT-en subset	CC BY-SA 4.0	12939	10800	GPT-3.5-turbo, Llama2, Llama3, ChatGLM, Baichuan, Qwen (1800 samples each)
Mixset	CC BY-SA 4.0	600	3000	-
LLM-DetectAIve-IELTS	huggingface	1635	900	llama-3.1-70B-versatile (900 samples)
IELTSDuck	Apache-2.0	10932	12418	GPT-4o-mini-2024-07-18, (10932), llama-3.1-70B-versatile (1486)
NLPeer	Apache-2.0	5376	5376	GPT-4o-2024-05-13 (5376)
Peersum	Github	5157	6997	GPT-4o-2024-08-06 (3501), GPT-4o-mini-2024-07-18 (3496)
Total	-	36639	39491	-
After deduplication	-	35393	39363	-
After removing short text	-	34675	39266	-

Table 15: Statistics of the English test set

Source / Domain	Language	# Human	# MGT	LLM Generator List
Cudrt-Subset	Chinese	12565	1500	GPT-3.5 (300), Qwen (300), GPT-4 (300), ChatGLM (300), Baichuan (300)
High School Student Essay	Chinese	3502	1556	GLM-4-9b-chat (778), Claude-3.5-sonnet (778)
Zhihu-Qa	Chinese	12524	10269	GPT-4o-2024-08-06 (3423), GPT-4o-mini-2024-07-18 (6846)
Mnbvc-Qa-Zhihu	Chinese	3000	3000	GPT-4o-2024-05-13 (3000)
Govreport	Chinese	2975	17695	GPT-4o-2024-05-13 (5932), ChatGLM3-6B (5821)
Easc (Summary)	Arabic	153	306	GPT-4o-2024-08-06 (306)
Tweets	Arabic	1400	3400	GPT-4 (1700), GPT-4o-2024-08-06 (1400), Qwen-2.5 72B (300)
Kalimat Youm 7 News	Arabic	1000	2000	GPT-4o-2024-05-13 (1000), Ace-GPT (1000)
Sanad (News)	Arabic	3000	3000	GPT-4o-2024-05-13 (3000)
Summaries	Russian	6562	6582	GPT-4o-2024-08-06 (3300), Vikhrmodels/Vikhr-Nemo-12B-Instruct-R-21-09-24 (3282)
News	Russian	6494	6539	GPT-4o-2024-08-06 (3295), Vikhrmodels/Vikhr-Nemo-12B-Instruct-R-21-09-24 (3244)
Wikipedia	Russian	1025	3049	GPT-4-0613 (999), Vikhrmodels/it-5.4-fp16-orpo-v2 (1025), AnatoliiPotapov/T-lite-instruct-0.1 (1025)

Table 16: Statistics of the multilingual test sets, part 1

Source / Domain	Language	# Human	# MGT	LLM Generator List
Wikipedia	Hebrew	1182	2173	GPT-4-0613 (991), dicta-il/dictalm2.0-instruct (1182)
Wikipedia	German	1865	2529	GPT-4-0613 (957), LeoLM/leo-hessianai-13b-chat (1572)
Wikipedia	Norwegian	1544	2543	GPT-4-0613 (999), norallm/normistral-7b-warm-instruct (1544)
Wikipedia	Spanish	600	600	Llama 3.1 405B instruct (600)
Wikipedia	Dutch	600	600	Llama 3.1 405B instruct (600)
Wikipedia	kaz	1300	1300	GPT-4o-2024-08-06 (1300)
Dice (News)	Italian	2800	2800	Llama 3.1 405B instruct (2800)
News	Urdu	13497	17472	GPT-4o-2024-08-06 (17472)
News	Hindi	600	600	GPT-4o-2024-08-06 (600)
News	Japanese	300	300	GPT-4o-2024-08-06 (300)
News	Vietnamese	600	600	GPT-4o-2024-08-06 (600)
Wikipedia	Vietnamese	600	600	GPT-4o-2024-08-06 (600)
Poetry	Indonesian	600	600	GPT-4o-2024-08-06 (600)
Total	-	80288	91613	-
Non-duplicated	-	78424	79305	-
Remove Short Text	-	73634	77791	-

Table 17: Statistics of the multilingual test sets, part 2