Are AI Detectors Good Enough? A Survey on Quality of Datasets With Machine-Generated Texts

Anonymous ACL submission

Abstract

The rapid development of autoregressive Large Language Models (LLMs) has significantly improved the quality of generated texts, necessitating reliable machine-generated text detectors. A huge number of detectors and collections with AI fragments have emerged, and several detection methods even showed recognition quality up to 99.9% according to the target metrics in such collections. However, the quality of such detectors tends to drop dramatically in the wild, posing a question: Are detectors actually highly trustworthy or do their high benchmark scores come from the poor quality of evaluation datasets? In this paper, we emphasise the need for robust and qualitative methods for evaluating generated data to be secure against bias and low generalising ability of future model. We present a systematic review of datasets from competitions dedicated to AI-generated content detection and propose methods for evaluating the quality of datasets containing AI-generated fragments. In addition, we discuss the possibility of using high-quality generated data to achieve two goals: improving the training of detection models and improving the training datasets themselves. Our contribution aims to facilitate a better understanding of the dynamics between human and machine text, which will ultimately support the integrity of information in an increasingly automated world.

1 Introduction

The quality of the outputs of autoregressive Large Language Models (LLM) has grown tremendously in the last five years, making their output almost indistinguishable from human-written texts [Chang et al., 2024]. This expanded the application fields of such models, as many routine tasks can be entrusted to them nowadays. However, we may observe multiple cases of overuse of these models, when they are utilised for creating texts that are intended to be written and fact-checked by humans. Misuse could be revealed in the generation of fake news [Zellers et al., 2019, Zhou et al., 2023], which can mislead readers of such content. Another example is concern in academic society, where a lot of students complete assignments with LLMs [Koike et al., 2024, Ma et al., 2023], making professors grade AI-generated content and undervaluing the purpose of the educational process. The spread of machine-generated fragments and plagiarism in scholarly articles increases rapidly with the growth of chatbots and reaches several tens of percent [Liang et al., 2024, Gritsay et al., 2023a]. Field researches also showed that in the last year alone, more than 60,000 scientific papers contain evidence of the use of machine generation [Gray, 2024]. As generated texts are gaining rapid and widespread popularity, it is crucial to develop systems able to counter the uncontrolled proliferation of artificial data and signal to the reader that the content they read is generated.

Another concern is that the Web is overflowing with machine-generated content, often of poor quality. The fact that such texts contribute bias to publicly available texts on the Internet, through false facts, hallucinations and spelling errors, is worth considering. Given the current agenda of using trillions of tokens from the Internet to teach new language models, Villalobos et al. [2022] revealed that the data will run out by 2028. That means that the training sets for more advanced language models in the future will include a large amount of generated content, as it already has a place in the current picture of available texts from all over the Internet. It was shown in Alemohammad et al. [2023] that such *self-consuming* will result in substantial degradation of the model's abilities. Furthermore, the trend is evolving in such a way that human-written texts on almost any topic will be much harder to retrieve. While for texts dated 10 years ago we are confident that the usage of generation was extremely rare, we cannot state the same

for modern texts.

Detectors capable to distinguish human-written texts from AI-generated texts, and whose detection quality can be guaranteed, are required by numerous specialists in many fields. We consider that the key to success in building excellent detectors will be high-quality artificial text collections. In this paper we would like to estimate the quality of the available generated data from competitions and research papers. As we are still in a lack of qualitative detectors, even though we see the claimed metrics on various datasets up to 99.9%, meanwhile in the wild a noticeable decline in performance can be seen. Competitions are a rather good way to suggest attractive challenges for researchers, but the results, obtained in these competitions may look confusing: as the models become more and more advanced, seemingly making the detecting task more challenging, participants of competitions still reach almost perfect scores of metric, bringing up the problem about quality of generated data in the provided datasets, see Appendix C.

Our contributions are following:

1. We systemize information about datasets from articles and competitions, dedicated to the detection of AI-generated content task.
2. We suggest methods that may be helpful for evaluating the quality of the generated data and the datasets aimed to use for binary classification between human and machine texts.

2 Related Work

AI-generated content detection is usually a text classification task, meaning that the input is a text sequence and the output is a discrete, usually binary class prediction. When the task is binary, the common labels are "AI" or "human", however in some cases one may be interested in multi-class classification between several language models. The last task is usually called authorship attribution. Finally, there is a recent, more complex task, called hybrid writing detection, when text contains fragments of both authors and the task is to determine the borders when authors change.

Among the approaches to tackle the classification problem [Jawahar et al., 2020] is to calculate linguistic [Fröhling and Zubiaga, 2021], stylometric, and statistical features, as well as using classical machine learning methods such as Logistic Regression or Random Forest as classifiers on these features. Some other techniques include the calculation of internal metrics, such as token-wise log probability or rank [Gehrmann et al., 2019]. They can be calculated for each token and compared for consistency with the prior context. However, these methods are falling behind approaches that use BERT-based models and keep an eye on the context [Ippolito et al., 2020, Gritsay et al., 2022] such as RoBERTa [Liu et al., 2019] and DeBERTa [He et al., 2021a], that are showing state-of-the-art results on this task [Uchendu et al., 2021].

Regarding methods for evaluating the quality of the generated data itself, it has become more common to evaluate with the help of LLMs themselves [Xu et al., 2023] . The main advantage of such approach is that it does not require any human reference, unlike ROUGE [Lin, 2004] or MoverScore [Zhao et al., 2019]. However, the output of model-evaluator needs to be unified and is not always interpretable. Another approach is suggested by Zhu and Bhat [2020], where the text is evaluated based on several linguistic criteria, such as grammar or coherence.

There are many datasets, developed as benchmarks for AI-generated content detection and it would be infeasible to evaluate them all. We will describe the datasets we used in our analysis and experiments in Section 3. There are a number of works dedicated to summarization of information about those datasets, however, only few works focus on the quality of data in them. Participants of competitions often provide a brief overview of the data in their solution descriptions [2022b, 2022, 2023b, 2024, 2024, 2024, 2024, 2024]. In [Voznyuk and Konovalov, 2024] it was shown that in SemEval Task 8 [Wang et al., 2024b] there were a number of generations of poor quality, which made the task for detectors much easier. In this work, we analyse some other datasets, whether they also contain similar sort of generation errors, by diminishing the quality of detectors trained on them.

3 Data

3.1 Datasets From Shared Tasks

Shared tasks to detect artificially generated fragments appeared several years ago. Participants are suggested to solve some tasks connected with detection, either a binary classification, authorship attribution or hybrid human-AI collaboration. In this work, we focused on competitions in which a binary classification task was present. It should be

noted that the studies only present data that were made available for training stage, as not all competitions had gold labels provided for test samples after the end of the competition. All shared tasks contain text in English, unless stated otherwise:

RuATD 2022 [Shamardina et al., 2022a] contains human- and machine-written fragments in Russian with wide range of themes.

DAGPap 2022 [Kashnitsky et al., 2022] contains human- and machine-written scientific excerpts collected by Elsevier.

AuTexTification 2023 [Sarvazyan et al., 2023] provides texts for classification in English and Spanish. The organisers considered five different domains to solve with.

IberAuTexTiification 2024 [Sarvazyan et al., 2024] is a continuation of the previous competition. The novelty of this shared task is that it is multilingual (six Iberian languages), multi-domain and multi-model.

Voight-Kampff Generative AI Authorship Verification 2024 [Ayele et al., 2024] provides two texts, one authored by a human, one by a machine, and participants were suggested to pick the human one. Test data for this task was compiled from the submissions of another competition' participants and comprised multiple text genres. For shorter term in future mentions, we will refer to this competition as **PAN 2024**.

SemEval 2024 Task 8 [Wang et al., 2024b] focuses on shifts between domains, generators and languages of generated texts. Among available languages are Chinese, Urdu, Bulgarian, Indonesian, Russian, Arabic, German and Italian, however test set consisted only of texts in English, Italian, German and Arabic.

MGT Detection Task 1 [1] is the continuation and improvement of the SemEval Task 8. Organisers aim to refresh training and testing data with generations from novel LLMs and include new languages.

A basic overview of the analysed datasets is presented in Table 1, and a more detailed description of the data source and the topics presented can be seen in Appendix B.

3.2 Datasets from Papers

The number of collections with generated content has started to increase with an increasing number of available generators. Quite often, researchers together with a new approach for AI content detection publish a parallel dataset on which they have validated their method. In this paper, we picked collections with human- and machine-generated excerpts that are the most common and cited in other researchers' publications.

GPT2 Output Dataset [2] consists of text outputs generated by GPT-2 models across various prompts.

HC3 (Human Chatbot Conversations Corpus) [Su et al., 2024] features conversations between humans and chatbots, primarily used for research on chatbot responses and human-AI interaction analysis. This dataset is available for both English and Chinese, but we have focused only on the former.

GhostBuster [Verma et al., 2024] aimed at detecting AI-generated content by comparing it to human-written text, often used in the context of identifying machine-generated misinformation or spam.

MGTBench (Machine Generated Text Benchmark) [He et al., 2023] is a benchmark dataset designed to evaluate the quality of machine-generated text across various tasks, including fluency, coherence, and creativity.

MAGE (Model Augmented Generative Evaluation) [Li et al., 2024] evaluates the performance of generative models by comparing outputs with human annotations, aiding in the development of more accurate generative AI models.

M4 (Multilingual, Multimodal, Multitask, Massive Dataset) [Wang et al., 2023] is a large-scale dataset designed for training models that can handle multiple languages, tasks, and modalities, making it useful for developing versatile AI systems. It contains texts in Arabic, Bulgarian, Indonesian, Russian, Chinese and Urdu, but we focused only on texts in English.

4 Approach

We decided to evaluate all datasets on common setups to see how good standard approaches perform on them. We did not have the goal to obtain the highest score, but rather to compare the performance of the same method on different datasets.

4.1 Classifiers

DeBERTa Classifier. A standard approach for binary classification of documents is to fine-tune in supervised manner some classifier, usually

[1] COLING-25 MGT Detection Task 1

[2] gpt2-output-dataset

Dataset	Year	Language	Num. of Texts	Generated VS Human	Average Length	Median Length
RuATD	2022	ru	129k	64.5k / 64.5k	236.86 / 221.47	99.0 / 95.0
DAGPap22	2022	en	5.3k	3.6k / 1.6k	799.45 / 1180.07	680.0 / 1126.5
AuTex	2023	en, es	65.9k	33.1k / 32.8k	315.08 / 297.28	386.0 / 351.0
IberAuTex	2024	es, en, ca, gl, eu, pt	98k	52.5k / 45.4k	1036.92 / 1058.36	981.0 / 1018.0
PAN24	2024	en	15.2k	14.1k / 1.1k	2640.50 / 3007.04	2731.0 / 2868.0
SemEval24 Mono	2024	en	34.2k	18k / 16.2k	2465.12 / 2358.05	2570.0 / 2083.5
SemEval24 Multi	2024	en, ar, de, it	42.3k	22.1k / 20.2k	2217.87 / 2256.67	2270.0 / 2032.0
MGT Task 1 Mono	2025	en	610.7k	381.8k / 228.9k	1448.28 / 1541.18	1208.0 / 1080.0
MGT Task 1 Multi	2025	en, zh, it, ar, de, ru, bg, ur, id	674k	416.1k / 257.9k	1422.74 / 1445.33	1195.0 / 1032.0

Table 1: Statistics of the datasets from the shared tasks.

BERT-like model. In our case, we used mDeBERTa [He et al., 2021b], which is the current state-of-the-art model for machine-generated text detection [Macko et al., 2023].

DetectGPT. The DetectGPT framework [Mitchell et al., 2023] introduced novel perplexity-based scoring function, which involves perturbing a text passage, after what log-probabilities between original and modified texts are compared. This score is passed as an input to a classification model, with GPT-2 [Radford et al., 2019] as the base model and T5-Large [Raffel et al., 2019] as perturbations generator. However, as DetectGPT requires intensive computational costs, we utilized Fast-DetectGPT [Bao et al., 2024], that substitutes DetectGPT's perturbation step with a more efficient sampling step. Another advantage is it can be requested via API and approach is applicable without additional fine-tuning stages.

Binoculars [Hans et al., 2024] introduce modified score with perplexities and cross-perplexity of two close language models to separate human and machine texts and is proved to be better than approach with classic perplexity. Also, the approach is applicable without additional fine-tuning stages.

4.2 Topological Time Series calculation

It was shown in Tulchinskii et al. [2023] that if we take the inner dimensionality of the manifold on the set of embeddings, we could separate human-written texts from machine-generated ones. However, in that work authors estimated short texts, and therefore we followed the approach in Kushnareva et al. [2024] with sliding window. Topological Time Series calculate intrinsic dimensions (PHD) of the text within sliding window and they can be used as a feature for detector. Authors conclude that this metric can, in fact, help to differentiate the texts of different origin. To be able to compare datasets between each other, we came up with a score, utilizing KL-divergence. Let h_d, m_d be distributions of intrinsic dimensions for two types of texts from the same dataset, of human and machine origin, then our KL_{TTS} is following:

$$\text{KL}_{\text{TTS}}(h_d, m_d) = |D_{\text{KL}}(h_d||m_d) - D_{\text{KL}}(m_d||h_d)|$$

The lower this score, the closer h_d and m_d are, which means almost indistinguishable texts and vice versa.

4.3 Perturbations and Shuffling

Based on the results of text modification studies [Sadasivan et al., 2024, Mitchell et al., 2023],

Dataset	Year	Language	Num. of Texts	Generated VS Human	Average Length	Median Length
GPT2	2019	en	1250k	1000k / 250k	2941.28 / 2616.04	3245.0 / 2459.0
HC3	2023	en	85.4k	26.9k / 58.5k	1010.50 / 680.68	1012.0 / 422.0
GhostBuster	2023	en	21k	18k / 3k	3345.07 / 3391.26	3439.5 / 2911.5
MGTBench	2024	en	23.7k	20.7k / 3k	1595.94 / 3391.26	1226.0 / 2911.5
MAGE	2024	en	436k	152.3k / 284.2k	1138.75 / 1281.88	706.0 / 666.0
M4	2024	en	89.5k	44.7k / 44.7k	1587.62 / 3162.40	1454.0 / 1697.0

Table 2: Statistics of the datasets from the research papers.

which show how small perturbations affect machine reading comprehension systems, we decided to consider this way of possibly assessing the quality of a dataset. The key idea here is that AI models are sensitive to such adversarial changes, unlike humans. We considered two modification ideas: Adversarial Token Perturbation and Sentence Shuffling.

Adversarial Token Perturbation. In this approach we divide the text into tokens and randomly replace the token with a synonym from the WordNet [Miller, 1994] collection with a probability of 70%. We apply such a technique to each represented class. Using an encoder model, we obtain embeddings for each of the texts in the current dataset. Finally, we measure the average embedding shifts for the classes of human and generated texts. We obtain the embedding shifts using the cosine distance between the embeddings of the original texts and the modified ones. As a result, after modifications we obtain Δ_{shift} — the log difference of the average embedding shifts.

$$\Delta_{\text{shift}} = \log \frac{\frac{1}{n}\sum_{i=1}^{n} \cos_d(h^o_{h_i}, h^p_{h_i})}{\frac{1}{m}\sum_{j=1}^{m} \cos_d(h^o_{m_j}, h^p_{m_j})},$$

where n and m – number of samples in the human and generated parts of the dataset respectively, $h^o_{h_i}$ – embedding of i-th fragment of human part of data, $h^p_{h_i}$ – the same embedding after perturbation. The same goes for $h^o_{m_i}$ and $h^o_{m_i}$ for machine-generated part of data. Finally, $\cos_d$ is a function that measures the cosine distance between two vectors.

Sentence Shuffling. In this approach, we randomly swap sentences, thereby affecting the cohesion of the text. We try to find out the effect of artificial origin on the difference between the distributions after permutations. By dividing a fragment into sentences and randomly reversing the order of 70% of the selected sentences, we apply this technique to each represented class. Further, using the text encoding model, we obtain embeddings for each of the texts of the current dataset. Finally, we measure embedding shifts for the class of human and generated texts, and after that we convert the shifts into probability-like distributions. This allows us to obtain at the end $\text{KL}_{\text{shuffle}}(H, M)$ — the KL-divergence between the shifts of human and generated texts.

$$\text{KL}_{\text{shuffle}}(H, M) = \sum_i H(i) \log \frac{H(i)}{M(i)},$$

$$H(i) = \frac{\cos_d(h^o_{h_i}, h^p_{h_i}) + \epsilon}{\sum_j \left(\cos_d(h^o_{h_j}, h^p_{h_j}) + \epsilon\right)},$$

and $M(i)$ has the same structure, except that instead of human class texts the generated class texts are used, ϵ is a small constant added to avoid division by zero.

4.4 Attention Maps

We hypothesised that some patterns in the attention map might be useful in evaluating the quality of the texts, as in the case of perfect quality of generation, the performance of the model on human and

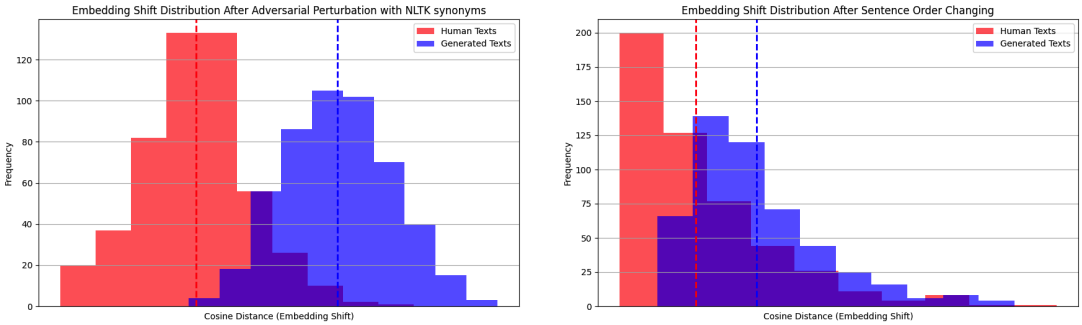

Figure 1: Comparisons of embedding shifts after two types of modifications for the HC3 dataset.

machine texts should be identical. As it is infeasible to evaluate attention maps visually for large amounts of texts, we came up with a statistic that may be useful in grasping some patterns. It was intended to grasp "attention columns" or massive activations [Sun et al., 2024] that appear on the attention map. For some texts there is no such pattern observed, for others there's only one distinct column, then there are cases when two or three distinct columns are seen. Our hypothesis was that attention maps for machine-generated texts should show more columns than for human ones, because there may be some tokens that model does not expect to see.

$$Attention_j(t) = \frac{1}{|t_{j:|t|}|} \sum_{i=j}^{|t|} Attention(q_i, k_j)$$

Therefore, until the end of the text, it still attends to such tokens, and we calculated the mean difference between the top-1 and the top-2 values across the dataset. Depending on the number of "attention columns", this difference will vary, but in general we assume that it should be comparable for texts of both origin.

5 Experiments

From each dataset, we sampled 1000 documents, balanced between two classes. As we dealt with multilingual datasets, we chose to finetune a multilingual model supporting all the presented languages, namely `mDeBERTa-v3-base`. Information about training can be found in the Appendix D.To evaluate the quality of baselines, Binoculars and Fast-DetectGPT, we launched `falcon-rw-1b` [Almazrouei et al., 2023] and `gpt-neo-2.7B` [Black et al., 2021] respectively. It is worth noting that with the last two methods we were only able to measure quality for samples in English. Our objective was to show that datasets of lower quality have shifts that will be easily recognised by the models "from the first step", hence we have not performed any hyperparameter tuning, only one iteration of fine-tuning and testing of the underlying models. In the experiment with topological features we used `RoBERTa-base`, just as the authors of original paper. In the experiment Perturbations and Shuffling, the `multilingual-e5-large` encoder was used to build embeddings of texts, which shows high metrics on encoding high-resource languages [Wang et al., 2024a]. Finally, in the experiments with attention maps we used `LLaMA2-7B` to obtain attention scores. We took 3 heads from Layer 15 and averaged them, as heads on this layer showed very distinct attention patterns.

6 Results

The results of comparison of the designed features on the test parts of the selected datasets are presented in Table 4. Regarding the TTS score, previous works have shown that texts of different origin have different PHD values, however this result was obtained for GPT-2 model, meanwhile current advanced models generate texts of much better quality, therefore PHD values become more similar. If texts of different origin have high KL_{TTS}, it means that it is easier for a detector to separate such texts. KL_{TTS} is also constrained for shorter texts, see Section 7.

The second column of Table 4 is based on attention map analyses. Here, on the left are the mean difference values of the human excerpts, on the right are the generated ones. When the attention maps of texts of different origin show different patterns, it also becomes easier for detectors to separate the texts. In the remaining columns we list the statistics observed on modified texts, and for both of these the lower the better, as this reflects the similar degree of resilience of the generated and

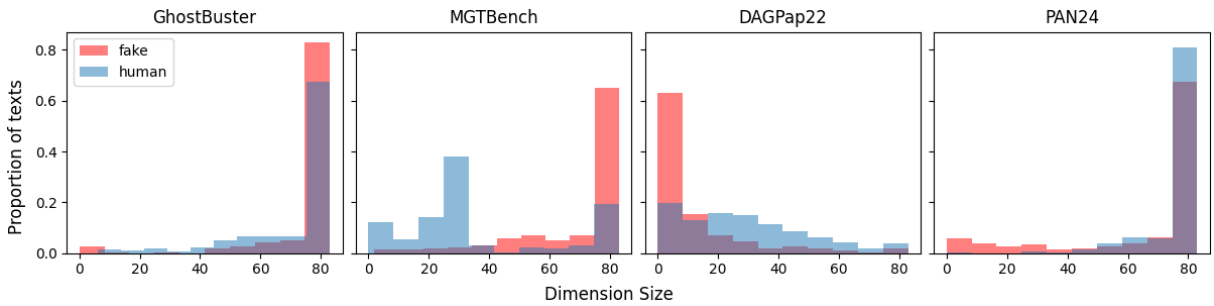

Figure 2: Topological Time Series for different datasets. These datasets obtained highest KL_{TTS}. The results for the remaining datasets selected in this paper can be found in Figure 4.

Dataset	DeBERTa	Binoculars	DetectGPT
GPT-2	0.972	0.495	0.412
HC3	0.998	0.931	0.972
GhostBuster	0.910	0.683	0.711
MGTBench	0.961	0.164	0.344
MAGE	0.835	0.632	0.654
M4	0.987	0.171	0.381
SemEval24 Mono	0.999	0.943	0.983
SemEval24 Multi	0.997	–	–
RuATD	0.765	–	–
DAGPap22	0.968	0.333	0.562
PAN24	0.826	0.411	0.890
AuTex23en	0.941	0.783	0.911
AuTex23es	0.933	–	–
IberAuTex	0.964	–	–
MGT-1 Mono	0.904	0.665	0.683
MGT-1 Multi	0.934	–	–

Table 3: Classification results with different detectors estimated using F_1-score. Binoculars and DetectGPT work only with English texts, thus we could not apply them to datasets with non-English texts.

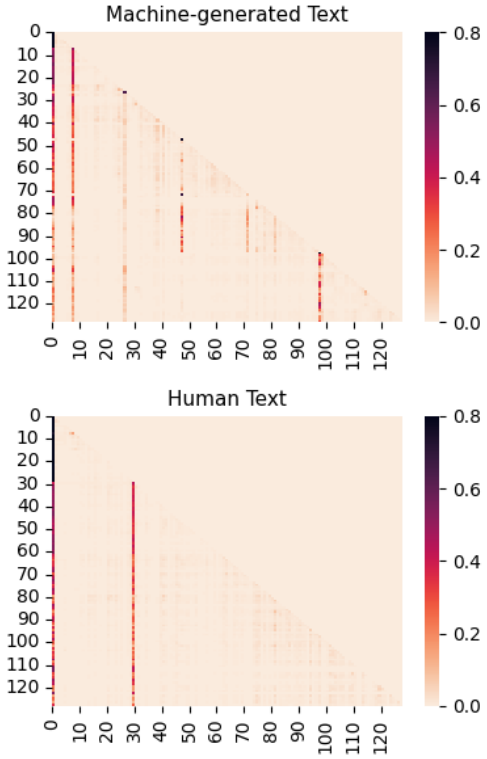

Figure 3: Attention maps on two excerpts from DAGPap22, Layer 15, Head 15.

human texts to adversarial attacks. Qualitatively generated data with no bias should take values close to human.

Finally, in Table 3 we show the results of applying modern classifiers to the chosen test parts. For instance, on the datasets that had low values in Table 4, a quality close to 1 can be achieved, which indicates the clear presence of detector bias towards them, or a structural feature that is too obvious for the model. It is not possible to judge the quality of the data only by achieving F_1 values close to 1, but by combining the values of the two tables we can estimate which set has better quality data and which has lower quality data.

7 Discussion

Regarding KL_{TTS}, on Fig. 2 we show 4 datasets with high value of it. While GhostBuster and PAN24 received such high score due to discrepancy on texts with higher dimensions, MGTBench and DAGPap22 did it due to the difference in distributions themselves. Note also that KL_{TTS} may not perform well with very short texts, since the internal method of computing PHD requires sufficiently long texts for stable computation. Therefore, we

Dataset	KL_{TTS} ↓	Attention Columns (h / m)	Δ_{shift} ↓	$KL_{shuffle}$ ↓
GPT-2	**0.014**	3.430 / 4.094	0.084	1.255
HC3	0.053	0.459 / 0.967	0.264	1.167
GhostBuster	0.053	**2.822 / 2.988**	**0.024**	**0.359**
MGTBench	0.043	1.961 / 2.639	**0.031**	**0.421**
MAGE	**0.011**	**2.289 / 2.166**	0.094	**0.310**
M4	0.036	3.842 / 2.256	0.107	**0.483**
SemEval24 Mono	**0.012**	1.540 / 0.766	0.191	2.576
SemEval24 Multi	**0.001**	2.123 / 0.830	0.059	2.046
RuATD	<u>0.007</u>	1.631 / 1.391	0.315	14.028
DAGPap22	0.083	**0.637 / 0.675**	**0.039**	**0.472**
PAN24	0.053	3.463 / 2.588	**0.050**	**0.331**
AuTex23-en	<u>0.021</u>	3.179 / 2.740	0.110	4.331
AuTex23-es	<u>0.001</u>	3.072 / 3.244	0.105	1.306
IberAuTex	**0.012**	**2.049 / 1.946**	0.223	5.516
MGT-1 Mono	**0.019**	2.070 / 1.783	**0.031**	0.587
MGT-1 Multi	**0.006**	3.313 / 3.117	**0.027**	0.522

Table 4: Calculated statistics on texts from chosen datasets. In "Attention Columns" we show the mean difference between the highest attention column and the second-placed. The first value is for human texts; the second value is for machine-generated. Some values for KL_{TTS} are underlined, because texts are too short, see Section 7.

discard KL_{TTS} on RuATD and AuTex23-es, as they do not fit the criteria, see Table 1. On top of that, it has already been shown that texts must be of sufficient length [Gritsay et al., 2022] to build reliable detectors.

Regarding attention maps, in Fig. 3, we depict an example of distinct patterns where human text has only one separate attention column besides attention on the first token, while machine-generated text has multiple attention columns.

Analysing the values in the table 4, we can trace the presence of sufficiently high quality data in the selected datasets. The developed attributes in aggregate are able to reflect the quality of the generated dataset from different perspectives and angles. We propose to utilise these attributes in combination with other statistical tools for evaluating data quality, e.g. Zipf's law [Powers, 1998].

Presented statistics can be utilised to estimate the quality of collections and to improve them. Also, datasets that collect machine-generated content may provide utility for the more general two other purposes as well. First, high-quality generated data can be utilised to evaluate the quality of the causal model during training, as one of the training objectives to improve model answers and make it more human-like. Secondly, good detectors can help to clean training sets, as large proportion of low-quality generated texts in those sets can result in emerging biases towards incorrect structure and rubbish fragments in the output of the model in the future.

8 Conclusion

In the current research, we discussed the problem of quality of datasets with AI-generated texts used for testing corresponding detectors. This problem is relevant, as the quality of test data directly influences the quality of widely used detectors. We conducted a review of datasets from competitions and scientific publications on datasets aimed at the detection of AI-generated content and proposed methods to evaluate the quality of datasets containing AI excerpts based on different structural features. We evaluated topological features, robustness to adversarial attacks, and similarity of attention patterns as estimators of data quality. We concluded that all analysed datasets fail in one another of our methods and do not allow to reliably estimate AI detectors. We encourage researchers to propose their own ways for quality assessment, which will allow to create a comprehensive system of evaluation of the detection datasets. Our work aims to contribute to a better understanding of the difference between human and machine text, which will ultimately contribute to preserving the integrity of information in the world.

9 Limitations

In our work we focused on the task of binary classification, thus suggested methods are not optimal for the task of detection of the hybrid AI-human content. Also, some methods do not work properly on short texts, however, so do the detectors.

References

Sina Alemohammad, Josue Casco-Rodriguez, Lorenzo Luzi, Ahmed Imtiaz Humayun, Hossein Babaei, Daniel LeJeune, Ali Siahkoohi, and Richard G. Baraniuk. 2023. Self-consuming generative models go mad. *Preprint*, arXiv:2307.01850.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. The falcon series of open language models. *Preprint*, arXiv:2311.16867.

Abinew Ali Ayele, Nikolay Babakov, Janek Bevendorff, Xavier Bonet Casals, Berta Chulvi, Daryna Dementieva, Ashaf Elnagar, Dayne Freitag, Maik Fröbe, Damir Korenčić, Maximilian Mayerl, Daniil Moskovskiy, Animesh Mukherjee, Alexander Panchenko, Martin Potthast, Francisco Rangel, Naquee Rizwan, Paolo Rosso, Florian Schneider, Alisa Smirnova, Efstathios Stamatatos, Elisei Stakovskii, Benno Stein, Mariona Taulé, Dmitry Ustalov, Xintong Wang, Matti Wiegmann, Seid Muhie Yimam, and Eva Zangerle. 2024. Overview of PAN 2024: Multi-Author Writing Style Analysis, Multilingual Text Detoxification, Oppositional Thinking Analysis, and Generative AI Authorship Verification. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction. 15th International Conference of the CLEF Association (CLEF 2024)*, Lecture Notes in Computer Science, Berlin Heidelberg New York. Springer.

Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. 2024. Fast-detectgpt: Efficient zero-shot detection of machine-generated text via conditional probability curvature. *Preprint*, arXiv:2310.05130.

Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. 2021. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow. If you use this software, please cite it using these metadata.

Galina Boeva, German Gritsai, and Andrey Grabovoy. 2024. Team ap-team at pan: Llm adapters for various datasets.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2024. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.*, 15(3).

Leon Fröhling and Arkaitz Zubiaga. 2021. Feature-based detection of automated language models: tackling gpt-2, gpt-3 and grover. *PeerJ Computer Science*, 7.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander Rush. 2019. GLTR: Statistical detection and visualization of generated text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 111–116, Florence, Italy. Association for Computational Linguistics.

Andrew Gray. 2024. Chatgpt" contamination": estimating the prevalence of llms in the scholarly literature. *arXiv preprint arXiv:2403.16887*.

German Gritsai, Ildar Khabutdinov, and Andrey Grabovoy. 2024. Multi-head span-based detector for AI-generated fragments in scientific papers. In *Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024)*, pages 220–225, Bangkok, Thailand. Association for Computational Linguistics.

German Gritsay, Andrey Grabovoy, and Yury Chekhovich. 2022. Automatic detection of machine generated texts: Need more tokens. In *2022 Ivannikov Memorial Workshop (IVMEM)*, pages 20–26.

German Gritsay, Andrey Grabovoy, Aleksandr Kildyakov, and Yury Chekhovich. 2023a. Artificially generated text fragments search in academic documents. In *Doklady Mathematics*, volume 108, pages S434–S442. Springer.

German Gritsay, Andrey Grabovoy, Aleksandr Kildyakov, and Yury Chekhovich. 2023b. Automated text identification: Multilingual transformer-based models approach. In *IberLEF@SEPLN*.

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. *Preprint*, arXiv:2401.12070.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021a. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *Preprint*, arXiv:2111.09543.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021b. Deberta: Decoding-enhanced bert with disentangled attention. *Preprint*, arXiv:2006.03654.

Xinlei He, Xinyue Shen, Zeyuan Chen, Michael Backes, and Yang Zhang. 2023. MGTBench: Benchmarking Machine-Generated Text Detection. *CoRR abs/2303.14822*.

Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. 2020. Automatic detection of generated text is easiest when humans are fooled. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1808–1822, Online. Association for Computational Linguistics.

Ganesh Jawahar, Muhammad Abdul-Mageed, and Laks Lakshmanan, V.S. 2020. Automatic detection of machine generated text: A critical survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2296–2309, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Yury Kashnitsky, Drahomira Herrmannova, Anita de Waard, George Tsatsaronis, Catriona Catriona Fennell, and Cyril Labbe. 2022. Overview of the DAGPap22 shared task on detecting automatically generated scientific papers. In *Proceedings of the Third Workshop on Scholarly Document Processing*, pages 210–213, Gyeongju, Republic of Korea. Association for Computational Linguistics.

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.

Laida Kushnareva, Tatiana Gaintseva, Dmitry Abulkhanov, Kristian Kuznetsov, German Magai, Eduard Tulchinskii, Serguei Barannikov, Sergey Nikolenko, and Irina Piontkovskaya. 2024. Boundary detection in mixed AI-human texts. In *First Conference on Language Modeling*.

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. *Preprint*, arXiv:2305.13242.

Weixin Liang, Yaohui Zhang, Zhengxuan Wu, Haley Lepp, Wenlong Ji, Xuandong Zhao, Hancheng Cao, Sheng Liu, Siyu He, Zhi Huang, Diyi Yang, Christopher Potts, Christopher D Manning, and James Y. Zou. 2024. Mapping the increasing use of LLMs in scientific papers. In *First Conference on Language Modeling*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Yongqiang Ma, Jiawei Liu, Fan Yi, Qikai Cheng, Yong Huang, Wei Lu, and Xiaozhong Liu. 2023. Ai vs. human – differentiation analysis of scientific content generation. *Preprint*, arXiv:2301.10416.

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.

George A. Miller. 1994. WordNet: A lexical database for English. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Pablo Miralles, Alejandro Martín, and David Camacho. 2024. Team aida at PAN: ensembling normalized log probabilities. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), Grenoble, France, 9-12 September, 2024*, volume 3740 of *CEUR Workshop Proceedings*, pages 2807–2813. CEUR-WS.org.

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'23.

David M. W. Powers. 1998. Applications and explanations of zipf's law. In *Proceedings of the Joint Conferences on New Methods in Language Processing and Computational Natural Language Learning*, NeMLaP3/CoNLL '98, page 151–160, USA. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Colin Raffel, Noam M. Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

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.

Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. 2024. Can AI-generated text be reliably detected?

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. *Preprint*, arXiv:2309.11285.

Areg Mikael Sarvazyan, José Ángel González, Francisco Rangel, Paolo Rosso, and Marc Franco-Salvador. 2024. Overview of iberautextification at iberlef 2024: Detection and attribution of machine-generated text on languages of the iberian peninsula. *Procesamiento del Lenguaje Natural*, 73(0):421–434.

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

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

Zhenpeng Su, Xing Wu, Wei Zhou, Guangyuan Ma, and Songlin Hu. 2024. Hc3 plus: A semantic-invariant human chatgpt comparison corpus. *Preprint*, arXiv:2309.02731.

Mingjie Sun, Xinlei Chen, J. Zico Kolter, and Zhuang Liu. 2024. Massive activations in large language models. *Preprint*, arXiv:2402.17762.

Eduard Tulchinskii, Kristian Kuznetsov, Laida Kushnareva, Daniil Cherniavskii, Sergey Nikolenko, Evgeny Burnaev, Serguei Barannikov, and Irina Piontkovskaya. 2023. Intrinsic dimension estimation for robust detection of ai-generated texts. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Adaku Uchendu, Zeyu Ma, Thai Le, Rui Zhang, and Dongwon Lee. 2021. TURINGBENCH: A benchmark environment for Turing test in the age of neural text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2001–2016, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Andric Valdez-Valenzuela, Ricardo Zavala-Reyes, Victor Morales Murillo, and Helena Goméz-Adorno. 2024. The iimasnlp team at iberautextification 2024: Integrating graph neural networks, multilingual llms, and stylometry for automatic text identification.

Vivek Verma, Eve Fleisig, Nicholas Tomlin, and Dan Klein. 2024. Ghostbuster: Detecting text ghostwritten by large language models. *Preprint*, arXiv:2305.15047.

Pablo Villalobos, Jaime Sevilla, Lennart Heim, Tamay Besiroglu, Marius Hobbhahn, and Anson Ho. 2022. Will we run out of data? an analysis of the limits of scaling datasets in machine learning. *arXiv preprint arXiv:2211.04325*.

Anastasia Voznyuk and Vasily Konovalov. 2024. DeepPavlov at SemEval-2024 task 8: Leveraging transfer learning for detecting boundaries of machine-generated texts. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1821–1829, Mexico City, Mexico. Association for Computational Linguistics.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024a. Multilingual e5 text embeddings: A technical report. *Preprint*, arXiv:2402.05672.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Chenxi Whitehouse, Osama Mohammed Afzal, Tarek Mahmoud, Alham Fikri Aji, and Preslav Nakov. 2023. M4: Multi-generator, multi-domain, and multi-lingual black-box machine-generated text detection.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Chenxi Whitehouse, Osama Mohammed Afzal, Tarek Mahmoud, Giovanni Puccetti, Thomas Arnold, Alham Fikri Aji, Nizar Habash, Iryna Gurevych, and Preslav Nakov. 2024b. Semeval-2024 task 8: Multigenerator, multidomain, and multilingual black-box machine-generated text detection. In *Proceedings of the 18th International Workshop on Semantic Evaluation*, SemEval 2024, Mexico, Mexico.

Wenda Xu, Danqing Wang, Liangming Pan, Zhenqiao Song, Markus Freitag, William Yang Wang, and Lei Li. 2023. Instructscore: Explainable text generation evaluation with finegrained feedback. *Preprint*, arXiv:2305.14282.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. *Defending against neural fake news*. Curran Associates Inc., Red Hook, NY, USA.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. *Preprint*, arXiv:1909.02622.

Yuan Zhao, Junruo Gao, Junlin Wang, Gang Luo, and Liang Tang. 2024. Utilizing an ensemble model with anomalous label smoothing to detect generated scientific papers. *Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024)*.

Jiawei Zhou, Yixuan Zhang, Qianni Luo, Andrea G Parker, and Munmun De Choudhury. 2023. Synthetic lies: Understanding ai-generated misinformation and evaluating algorithmic and human solutions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA. Association for Computing Machinery.

Wanzheng Zhu and Suma Bhat. 2020. GRUEN for evaluating linguistic quality of generated text. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 94–108, Online. Association for Computational Linguistics.

A Additional Results on Topological Time Series

In Figure 4, it can be seen the plotted Topological Time Series distributions for all the selected datasets for the currect research.

B Data Description

More detailed description with information on sources, topics and years of the datasets selected in this paper from competitions in Table 5 and research articles presented in Table 6.

C Evaluation results of competitions

Table 7 shows the winning scorers in the competitions reviewed in this paper. In the AuTex and IberAuTex competitions it was forbidden to use additional data to fine-tune the detection algorithms. In the other collections it was allowed, we can notice a high quality near perfect in them. We should note the low value of metrics on the RuATD dataset, which can be explained by the limited number of available high-quality language models in Russian during the competition.

Competition	Metric	Best result
RuATD	Accuracy	0.820
AuTex-en	Macro-F1	0.809
AuTex-es	Macro-F1	0.708
IberAuTex	Macro-F1	0.805
SemEval24 Mono	Accuracy	0.975
SemEval24 Multi	Accuracy	0.959
PAN24	Avg. of 5 metrics*	0.924
DAGPap22	Avg. F1-score	0.994

Table 7: Best results from each analysed competition. PAN24 used mean of 5 metrics, such as accuracy, F1 and other to evaluate *efficiency* of the system.

D Hyperparameters

Hyperparameters	Values
Epochs	5*
Learning rate (LR)	5e-5
Warmup steps	50
Weight decay	0.01

Table 8: Hyperparameters for fine-tuning mDeBERTa-base. We trained for 5 epochs with possibility of early exit.

The training was conducted on NVIDIA GeForce RTX 3090. See hyperparameters in Table 8.

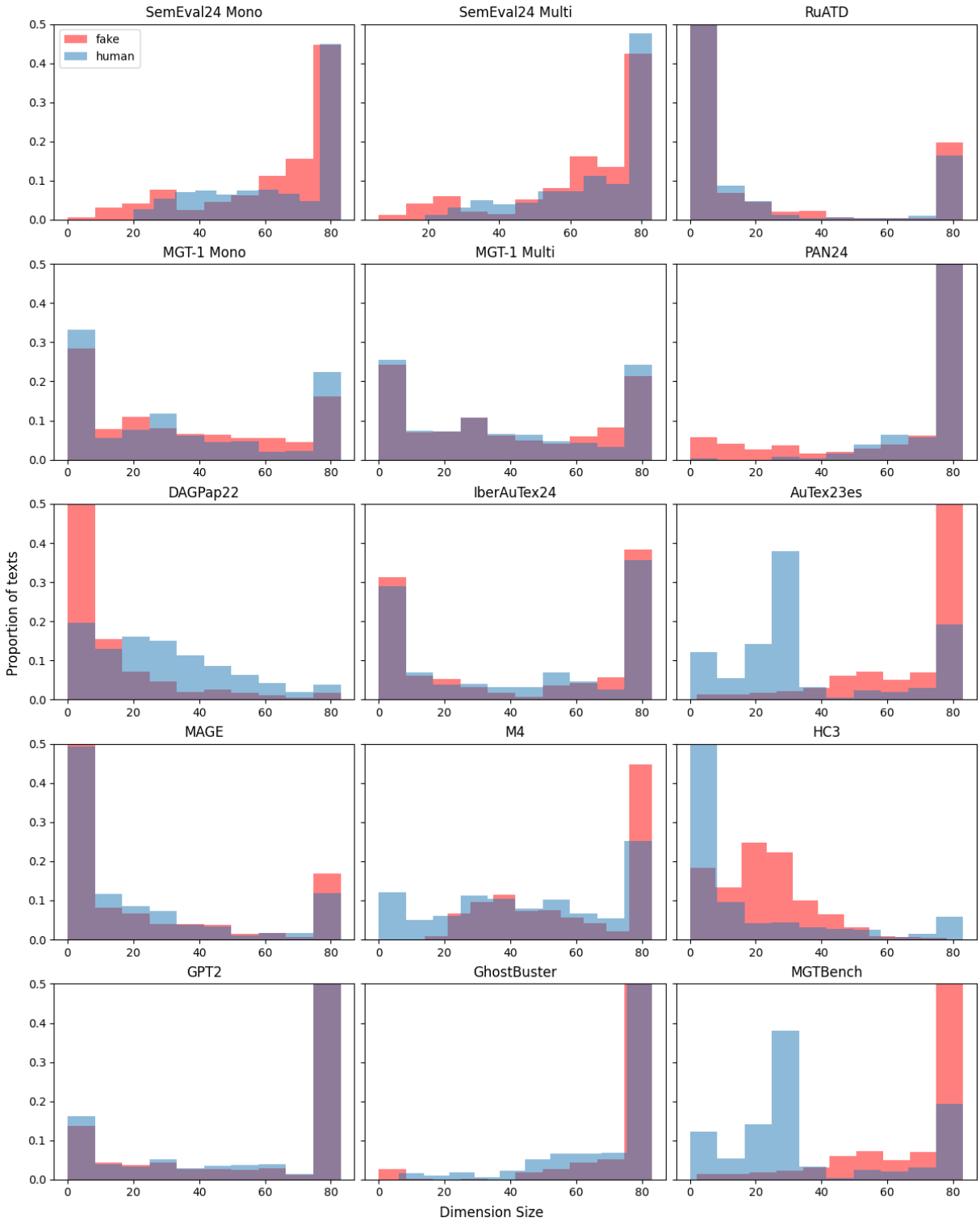

Figure 4: Topological Time Series for remaining datasets from current review. We omitted the results for AuTex23en, because virtually all texts there had the dimension of 0.

Dataset	Year	Themes	Sources
RuATD	2022	News, Social media, Wikipedia, Strategic Documents, Diaries	M-BART, M-BART50, M2M-100, OPUS-MT, mT5-Large, mT5-Small, ruGPT2-Large, ruGPT3-Large, ruGPT3-Medium, ruGPT3-Small, ruT5-Base, ruT5-Base-Multitask, ruT5-Large
DAGPap	2022	Scopus papers	Led-Large-Book-Summary, GPT-3, Spinbot, GPT-Neo-125M
AuTex	2023	Legal documents, Social media, How-to articles	BLOOM-1B7, BLOOM-3B, BLOOM-7B1, GPT-3 (Babbage, Curie, text-davinci-003)
IberAuTex	2024	News, Reviews, Emails, Essays, Dialogues, Wikipedia, Wikihow, Tweets	GPT, LLama, Mistral, Cohere, Anthropic, MPT, Falcon
PAN	2024	News	Alpaca-7B, BLOOM-7B1, Alpaca-13B, Gemini-Pro, ChatGPT (gpt-turbo-3.5, gpt-4-turbo), Llama-2-70B, Llama-2-7b, Mistral-7B, Mistral-8X7B, Qwen1.5-72B, GPT-2
SemEval Mono	2024	Wikipedia, WikiHow, Reddit, arXiv, PeerRead, Student Essays	ChatGPT (text-davinci-003, gpt-4), Cohere, Dolly-v2, BLOOMz
SemEval Multi	2024	Wikipedia, WikiHow, Reddit, arXiv, and PeerRead, Student Essays, News	ChatGPT (text-davinci-003, gpt-4), LLaMA2, Cohere, Dolly-v2, BLOOMz, Jais
MGT Detection Task 1 Mono	2025	CNN, DialogSum, Wikipedia, WikiHow, Eli5, Finance, Medicine, XSum, PubMed, SQuAD, IMDb, Reddit, arXiv, PeerRead	ChatGPT (text-davinci-002, text-davinci-003, gpt-turbo-3.5), OPT, LLama3, BLOOMz, FLAN-T5, Cohere, Dolly, Gemma, Mixtral
MGT Detection Task 1 Multi	2025	CNN, DialogSum, Baike, QA Wikipedia, WikiHow, Eli5, Finance, Medicine, Psychology, XSum, PubMed, SQuAD, IMDb, Reddit, arXiv, PeerRead	ChatGPT (text-davinci-002, text-davinci-003, gpt-turbo-3.5, gpt4o), GLM, GPT-J, GPT-Neo, OPT, Llama2, LLama3, BLOOMz, FLAN-T5, Cohere, Dolly, Gemma, Mixtral, Jais

Table 5: More detailed descriptive statistics about domains and generators of the chosen datasets from competitions.

Dataset	Year	Themes	Sources
GPT2	2019	WebText	GPT-2-117M, GPT-2-345M, GPT-2-762M, GPT-2-1542M
HC3	2023	ELI5, WikiQA, Wikipedia, Medicine, Finance	ChatGPT (gpt-turbo-3.5)
GhostBuster	2023	Student Essays, News Articles, Creative Writing	ChatGPT (gpt-3.5-turbo), Claude
MGTBench	2024	Student Essays, News Articles, Creative Writing	ChatGLM, Dolly, ChatGPT-turbo, GPT4All, StableLM, Claude
MAGE	2024	Opinions, Reviews, News, QA, Story Generation, Commonsense Reasoning, Knowledge Illustration, Scientific Writing	ChatGPT (text-davinci-002, text-davinci-003, gpt-turbo-3.5), LLaMA, GLM-130B, FLAN-T5, OPT, BigScience, EleutherAI
M4	2024	Wikipedia, Reddit ELI5, WikiHow, PeerRead, arXiv abstract	ChatGPT (text-davinci-003, gpt-turbo-3.5), Cohere, Dolly-v2, BLOOMz

Table 6: More detailed descriptive statistics about domains and generators of the chosen datasets from papers.