Features and Detectability of German Texts Generated with Large Language Models

Verena Irrgang[*1], **Veronika Solopova**[*1], **Steffen Zeiler**[1], **Robert Nickel**[2], **Dorothea Kolossa**[1],
[1]Technische Universität Berlin, [2]Bucknell University,
Correspondence: veronika.solopova@tu-berlin.de

Abstract

The proliferation of generative language models poses significant challenges in distinguishing between human- and AI-generated texts. This study focuses on detecting German texts produced by various Large Language Models (LLMs). We investigated the impact of the training data composition on the model's ability to generalize across unknown genres and generators and still perform well on its test set. Our study confirms that models trained on data from a single generator excel at detecting that very generator, but struggle to detect others. We expanded our analysis by considering correlations between linguistic features and results from explainable AI. The findings underscore that generator-specific approaches are likely necessary to enhance the accuracy and reliability of text generation detection systems in practical scenarios. Our code can be found in the Github repository[1].

1 Introduction

The newest generation of generative models, such as GPT-4 (OpenAI, 2023) and Gemini (Anil et al., 2024) achieve unprecedented levels of text quality. While humans are less likely to believe AI-generated headlines (Longoni et al., 2022), they are not reliable annotators when it comes to determining whether a text was AI-generated or human-generated (cf. (Clark et al., 2021; Kreps et al., 2020; Brown et al., 2020)). Up to 75% of articles generated by GPT-2 were found to be credible, often even more credible than the original source text for the article (Solaiman et al., 2019, p. 10). This uncovers the dark potential of generative models to be used to create credible fake news, amplifying the already existing challenges for democracies posed by human-written fake news. For instance, Zellers et al., 2019 showed that their model outperformed human-written fake news in terms of credibility. Although e.g. ChatGPT is "censored" via Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022), it can still be used for malicious purposes, and open-source models have even bigger potential in that regard (Newhouse et al., 2019). For instance, Llama 2 outputs have been shown to have increased values for "Toxicity" and "Bias" compared to human-written texts (Touvron et al., 2023). In addition, large language models also tend to invent factual statements and notoriously hallucinate (Das et al., 2022).

Recent developments have shown that AI techniques also have the potential to *detect* AI-generated content. Chakraborty et al., 2023 provided a meta-level proof that as the number of samples increases, so does the possibility of detecting AI-generated text in an automated fashion, even when individual samples might be near indistinguishable. Lyu et al., 2022 identified such features of AI-generated text as topic drift, prolix sentences, abnormal paragraphs, poor text length control, overused phrases, and sparsity of uncommon characters. At the same time, generators improve and may display less easy-to-detect characteristics over time, making older detectors obsolete. Many new generative models appear while detectors are hypothesised to have poor generalization to new models and unknown domains (Tulchinskii et al., 2023). The general Anglo-centrism inherent to NLP research applies to this task. Meanwhile, there is more potential to create harmful content in other languages than in English. For instance, ChatGPT testers reported that it refused to generate recruitment propaganda for terrorist groups when prompted in English but it did so in Farsi (Murgia, 2023).

In light of this, our goal was to create a detector tailored to German language texts, specifically from genres in which fake news often appears. Our study

[*] These authors contributed equally to this work.

[1] https://github.com/vernsy/generated_text_detector

Proceedings of the 20th Conference on Natural Language Processing (KONVENS 2024), pages 264–280
September 10-13, 2024 ©2024 Association for Computational Linguistics

encompassed multiple generative models and their fine-tuned versions. Our best detector achieved an F1 score of 0.95. Moreover, we investigated how our models behave on texts from unseen genres and source generators. We assessed calibration, and performed model-agnostic and gradient-based analysis of model predictions and compared them with a statistical analysis of linguistic features displayed by each of the generative sources in our training data. Finally, we also determined token probabilities in an autoregressive fashion to measure differences between human- and AI-generated texts.

2 Related Work

Considerable research has already been conducted on the automatic detection of generated texts.

2.1 Architectures

Different detectors' architectures were explored. A transformer architecture was used by Alamleh et al., 2023 to detect English texts generated by ChatGPT; Lyu et al., 2022 and Guo et al., 2023a worked with Mandarin and Cantonese, while Gritsay et al., 2022 used a similar method on a Russian dataset. Schaaff et al., 2023 looked at multi-lingual solutions working with French, German, Spanish, and English, using statistical methods with XGBoost, a random forest (RF), and a multi-layer perceptron (MLP) network, as well as linguistic features. The models had varying performance levels on different languages and were much more accurate in detecting AI-generated texts than AI-rephrased ones.

Bakhtin et al., 2019 remark that the difference in the architecture between models that generate text and the detectors is one of the reasons for decreased robustness. Indeed, Small Language Models (SLMs) were shown to perform this task better than e.g. logistic regressors (Solaiman et al., 2019). In this study, it was also shown that a bidirectional model, RoBERTa (Liu et al., 2019a), outperformed a unidirectional model, i.e. the initial GPT. This was confirmed by Gehrmann et al., 2019 and Guo et al., 2023a. In contrast, Zellers et al., 2019 showed the advantages of using a detection model with the same architecture as the generating model. However, with the abundance of high-quality open-source generative models, this approach lacks cost-effectiveness.

2.2 Existing detectors

As for publicly available detectors, *GPTZero* (Edward et al., 2024) was the first generated text detector (Tian and Cui, 2023; Renbarger, 2023). While the current version deploys a transformer model and reports 98% accuracy, its initial version considered textual features, such as *perplexity* and *burstiness* which measure how "unexpected" a sequence of tokens is (Tian, 2023). Other popular detectors include *Originality-AI* and *Copyleaks*, which still rely on statistical methods (Orginality.ai, 2024; Copyleaks, 2024), as well as deep-leaning-based methods *Content at Scale* (at Scale, 2023), *Writer* (Writer, 2023) and *ZeroGPT* (ZeroGPT, 2024) all reporting 90% to 98% accuracy.

2.3 XAI studies

Much less has been done in terms of explainability for generated text detectors. Guo et al., 2023b tried to explain the choices of the model by extracting features with layer-wise relevance propagation. Alamleh et al., 2023 utilized the explainable AI framework SHAP (Lundberg and Lee, 2017) to find patterns such as politeness, lack of detail, fancy vocabulary, or reduced expressiveness in the texts. Mitchell et al., 2023; Gehrmann et al., 2019 and Ippolito et al., 2020 base their detectors partly on token probabilities to get explainable decisions.

To the best of our knowledge, our work is the first research analysing and explaining LLM-generated text detectors tailored for German and trained on data from a wide range of generators.

3 Methods

In this Section, we describe our dataset curation, the choice of the pre-trained model, and the experimental setup for training and explainability.

3.1 Data

Human-written. As we intended to use weak labels, when collecting German human texts we made sure to only scrape texts strictly from before 2016, the time before the first generative models went online, to make sure the data is indeed authentic (Foote, 2023). We selected genres stylistically similar to fake news (Grieve and Woodfield, 2023; Tsai, 2023). These are newspaper articles and social media posts, as well as Wikipedia articles and scientific publication's abstracts because fake news were shown to often use an explanatory tone (Khan et al., 2021). Namely, we collected scientific texts (*Springer Nature* Gruppe, 2023), journalistic texts (*die Tageszeitung*, taz Verlags u. Vertriebs GmbH, 2023), literary texts (*Wikimedia* zur Förderung Freien Wissens e.V., 2024a), encyclopedic texts (*Wikipedia* zur Förderung Freien Wissens e.V., 2024b) and everyday language from blog discussion threads (*Zeit Online* GmbH, 2023). To ensure that Wikipedia and Wikimedia texts were not updated since 2016, we extracted archived versions from The Internet Archive (Archive, 2023) and Wikipedia dump (zur Förderung Freien Wissens e.V., 2008). Wikimedia contains poems and novels, which, due to their linguistic sophistication, ensure our dataset can compete stylistically with elaborately formulated fake news.

AI-generated. The generated texts were produced with:

- Llama 2 (Touvron et al., 2023) with a temperature of 0.8 and a `top_p` value for nucleus sampling of 0.9,
- GPT3.5 via the OpenAI API (OpenAI, 2022) with `temperature` 0.6[2] and `top_p` 0.4,
- Snoozy (Anand et al., 2023) and Wizard (Xu et al., 2023), which are both GPT4All Llama 2 fine-tuned systems,
- Mistral (Hermeo)[3], a German-English model merged from DPOpenHermes-7B-v2 and leo-mistral-hessianai-7b-chat using mergekit, both fine-tuned versions of chat Mistral-7B-v0.1 (Mistral-AI, 2023).

While normally, a Q&A prompting format produces higher quality output than simple story completion (Guo et al., 2023a), fake news would usually not look like an answer to the question. We, therefore, prompted generators with two sentences from our human dataset of fixed length and asked them to complete the story with a similar length as the original text. We also included texts of various lengths, as it was shown to be beneficial for better generalisation of the classifier (Ippolito et al., 2020; Gritsay et al., 2022). Outputs with obvious repetitions were removed. The resulting data proportions are shown in the Appendix, Table 3.

3.2 Models

We performed a baseline experiment to select a pre-trained model for our main experiments out of three candidates: RoBERTa (Liu et al., 2019b), BERT (Devlin et al., 2019) and DistilBert (Sanh et al., 2019)[4]. We used LLaMA 2 (Touvron et al., 2023) generated text and human text for our baseline experiments. Over 8 training repetitions, 3 epochs each, DistilBert performed slightly better with an F1 score of 0.94, while RoBERTa and BERT achieved an F1 score of 0.93.

3.3 Training experiments

Our main set of experiments included fine-tuning DistilBert on 3 different datasets. While the human-written proportion remained unchanged, we experimented with different proportions of AI-generated training data.

In **Experiment 1** we trained a model (a) with Llama texts only (Model (a) train set). Then, in **Experiment 2**, for model (b) we reduced the Llama data portion and added a mixed-generated dataset, without Mistral and GPT Wikimedia data (model (b) train set).

For experiments (a) and (b) we used 4 test sets:

1. Model (a) test set (Llama 2 data versus human);
2. Model (b) test set (mixed data and reduced Llama 2 subset versus human);
3. Mistral test (unseen model);
4. GPT-Wikimedia (seen model, unseen genre).

Finally, for **Experiment 3**, we trained a model (c) on *all* of the data with a 0.2 train/test ratio (model (c) train/test). The exact train and test sample numbers are shown in the Appendix, Table 4.

We verified how well models were calibrated with the expected calibration error (ECE) (Guo et al., 2017) for each subset. An error analysis revealed which data subsets were more challenging for the models.

3.4 Explainability

Statistical linguistic analysis. We pre-selected relevant linguistic features from (Solopova et al., 2023a) based on various studies on fake news and propaganda detection and describe various morpho-syntactic and shallow semantic language parameters. We also used a sentiment analysis model from (Guhr et al., 2020) to annotate data with the probability of neutral, negative, and positive sentiment. A Gibbs cycle detection model from (Solopova et al., 2023b) was employed, where the detected elements correspond to different stages of the cognitive reflective cycle (description, evaluation, analysis, etc.), which we hypothesised to be more present

[2] Different temperature parameters for different generators were chosen empirically based on output quality.

[3] https://huggingface.co/malteos/hermeo-7b

[4] All three models are base and uncased versions of the respective pre-trained models.

in human texts. Additionally, we developed features intended to capture residual errors in generated texts. Namely, we measured the functional to lexical word ratio, the number of repeating lemmas, and mean, maximum, and medium sentence similarity. For the latter, we used the multilingual MiniLM-L12 v2 sentence transformer (Reimers and Gurevych, 2019) pairwise comparing all sentences of the text against each other. We also added the adjective-to-noun ratio and adverb-to-verb ratio, as AI-generated texts were said to possess more sophisticated language. A comprehensive list of all features can be seen in the Appendix, Table 5. The resulting 75 features were normalized by the number of tokens (in case of the counts) and separated into 6 groups depending on their source: human, Llama, GPT, Mistral, Wizard, and Snoozy. First, to understand the structure of our data, we performed principal component analysis (PCA). Then, after analyzing the distributions of our features, and concluding that all features, except for noun frequencies are not normally distributed, we performed the non-parametric Kruskal-Wallis Test, followed by the Mann-Whitney U-Test. We then selected features with the lowest p-values which were significant ≥ 2 comparison pairs.

Model-agnostic and attribution method. We decided to use several explainability methods to see how comparable their conclusions would be. We chose LIME (Ribeiro et al., 2016) and a gradient-based method (Janizek et al., 2021). While LIME approximates the local decision boundary of the model by generating a new dataset consisting of perturbed samples around a given input and observing how the model's predictions change with these perturbations, gradients analysis computes derivatives of the model's outputs for its inputs, tracking gradients with respect to input embeddings during the backward pass to see how changes in each input embedding dimension could affect the model's prediction.

We applied both methods on a merged test set of all test subsets. We averaged LIME coefficients and attribution scores and collected the top 95% of the highest class-bias coefficients and attribution scores with their corresponding tokens. In the case of gradient analysis, we followed the advice of Wang et al., 2020, and filtered out all stop-word tokens, tokens smaller than 3 characters, and those including "#".

Llama token probability. The token-succession probabilities that a Llama 2 model produces when it is not run in a generative mode, but rather in a text-analysis mode, can also be used to uncover specific differences between human- and AI-generated texts. Taking the first two sentences of a sample as a prompt, we ran the Llama 2 model to extract the tensor of probabilities of all possible next tokens to be chosen. We identified the probability of the respective next token of the given text sample and repeated the process autoregressively by appending the token from the last iteration to the input prompt. In this way, we produced a sequence of token probabilities for each token of a given input text. In the subsequent analysis, we divided samples into four groups of either correctly or falsely classified AI samples and correctly or falsely classified human-samples.

4 Results

This section presents the results for the training and explainability experiments.

4.1 Detectability

We trained each model over three epochs, 20 times per experiment with 20 different train/validation splits with a 0.8 to 0.2 split ratio. The training results for the native data set, with extended metrics, are illustrated in Table 1, while the performance in terms of the F1 score can be seen in Table 2.

We can see that by all metrics, model (a), only trained on Llama 2 and human data, performs the best on a test set drawn from its own distribution. The Matthews correlation coefficient (MCC) value across all experiments points to the fact that the model's effectiveness varies across classes. Complete model (c) visibly performs the worst with much lower MCC and recall overall, but only slightly dropping in area under ROC (AUROC) and precision compared to other settings.

However, when we look at the set-wise performance shown in Table 2, model (a) drops the most among the experimental settings on the mixed dataset (down to an F1 score of $\approx$0.6). Rather good F1 scores are, furthermore, misleading in the case of the Mistral and GPT test sets, as the model has a very low recall. This means that it mostly classifies GPT and Mistral samples as human.

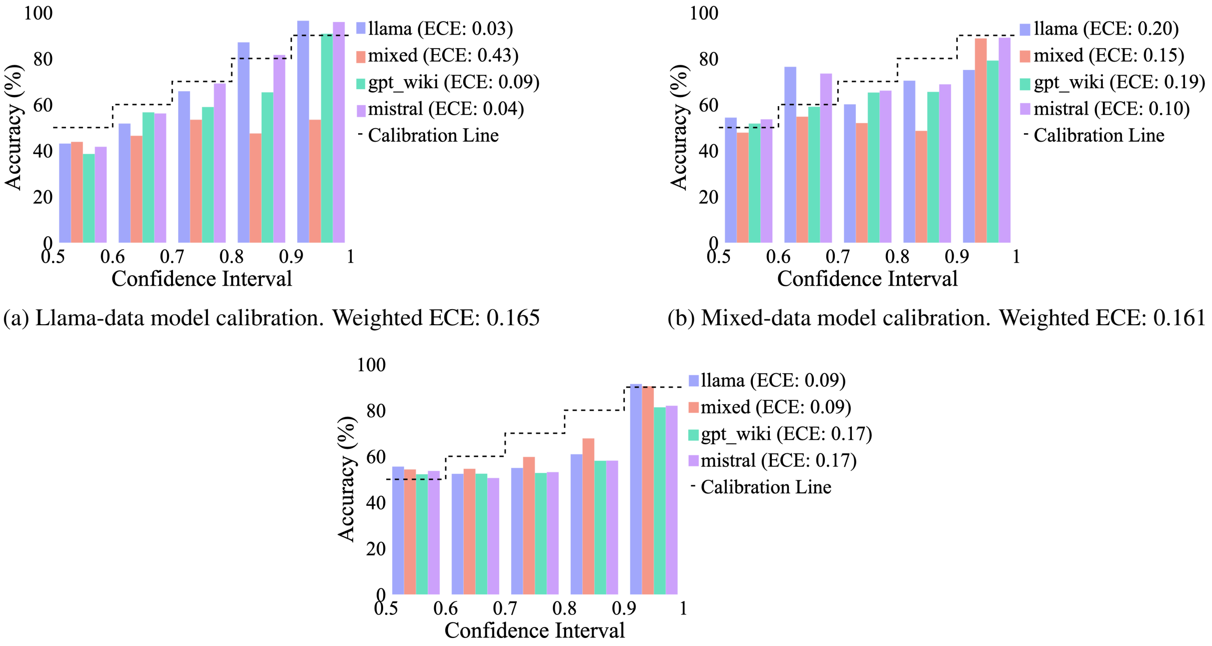
The reduction of the number of Llama 2 samples and the addition of other sources in the Mixed-data model (b) significantly decreases the model's performance on the Llama test set. Interestingly, model (b) achieves better results than model (a) on both the GPT test of the unknown genre and the

Metric	Baseline	(a) Llama	(b) Mixed	(c) Complete
F1 score	0.93	0.95	0.93	0.90
MCC	0.87	0.90	0.86	0.79
AUROC	0.98	0.99	0.98	0.97
Precision	0.93	0.95	0.93	0.92
Recall	0.93	0.95	0.93	0.86

Table 1: (a) Llama-data model, trained only on human and Llama data; (b) Mixed-data model, trained on data from all sources, except for Mistral data, GPT data generated from Wikimedia prompts, and a reduced Llama portion; (c) Complete model trained on all sources.

Testset	(a) Llama	(b) Mixed	(c) Complete*
LLaMA testset	0.95	0.78	0.86
Mixed testset	0.59	0.93	0.85
GPT testset	0.76 (R: 0.27)	0.74 (R: 0.59)	0.74 (R: 0.79)
Mistral testset	0.80 (R: 0.17)	0.80 (R: 0.60)	0.75 (R: 0.86)

Table 2: F1 scores for each of the models described in Table 1 on various test datasets. *The sizes of the test sets for the complete model are reduced (see Section 4.1). R: recall is provided for test datasets with unequal proportions.



(a) Llama-data model calibration. Weighted ECE: 0.165

(b) Mixed-data model calibration. Weighted ECE: 0.161

(c) Complete model calibration. Weighted ECE: 0.116

Figure 1: Calibration plots for the 3 models. The first bar always corresponds to the model's performance on the Llama test set, the second on mixed, the third on GPT Wikimedia and the last on Mistral.

unseen model Mistral in terms of recall. The values are still low for both models (a) and (b) pointing to poor generalization capabilities for both unknown genres and sources.

The GPT-Wikimedia and Mistral test sets for model (c) are smaller than for models (a) and (b), which prevents perfect comparison. Nonetheless, having been trained on the same number of Llama samples, model (c) performs better than mixed model (b) on the Llama portion of the test set, but worse than model (a), which was trained purely on this source. Similarly, model (c) is around 25%-points more accurate on mixed test data than model (a), but also around 7%-points less accurate than model (b). It has a much higher recall for both GPT and Mistral test sets. Overall, all three models seem to default to the human-written class when uncertain (see confusion matrices in Appendix, Figures

8 to 10.)

4.2 Model calibration

Figure 1 shows the calibration measures for our three models (a), (b), and (c) on our test sets 'llama', 'mixed', 'gpt_wiki', and 'mistral'. Models (a) and (b) have a similar overall ECE value of about 0.16. Model (a) itself is fairly well calibrated on its own (matching) test set. The model performs slightly worse for Mistral, and GPT-wikimedia, while its calibration on the mixed dataset is the worst. This variation in performance across different datasets indicates a limited ability of model (a) to generalize beyond its training context and may reflect the model's tendency to overfit when trained on one source. Hence, when deployed in more diverse or dynamic settings, it may not prove to be reliable.

Model (b) is underconfident in intervals from 0.5 to 0.7 on the Llama and Mistral dataset, and overconfident when assigning higher probabilities. Therefore, the model exhibits variations in its reliability, which are more pronounced outside the central probability range.

According to its overall ECE score of 0.11, the complete model (c) is the best-calibrated one of the three. It performs almost equally well on the Llama and mixed subsets. However, it does not perform better on the GPT-wikimedia and Mistral subsets, despite their presence in the training data. Thus, adding a small subset of unseen data does not necessarily translate into better performance, and DistilBert needs a substantial amount of samples to learn distributions of a new source.

4.3 Statistical linguistic analysis

Considering simple averages between human and generated sources overall, mostly only task-specific features showed commensurate variability (Figure 4). Human texts have marginally fewer repeating lemmas, adverbs in proportion to verbs, and a lower ratio of functional words (like prepositions, and conjunctions) to lexical words (like nouns, and verbs) compared to AI-generated data. Maximum sentence similarity has an especially strong negative value (-4). In contrast, positive values are seen in description sentence probability, as well as mean and median sentence similarity, indicating a higher count of these indicators in human texts.

PCA. According to our PCA analysis shown in Figure 3, PC1 explains almost 10% of our linguistic features while PC2 explains 6%. Mistral and Llama data seem to be outliers, while Wizard appears to produce data that is the most similar to human text.

Many features are especially discriminative in the case of Llama data: the number of foreign words, simple sentences, positive sentiment probability and repeating lemmas; all point to lower quality of data produced. Amount of nouns and neutral sentiment probability are the only two separating the Mistral and GPT data from the rest, while the number of 2nd person pronouns, pronouns in general, adverbs, subordinating conjunctions, complex sentences and verbs, differentiate Wizard and human text from all others. Interestingly, the probability of text being descriptive also points in the opposite direction from the human texts.

Kruskal-Wallis and U-Test. Due to the high number of data points, features, and comparison pairs (6 subsets produce 15 pairwise comparisons), the Kruskal-Wallis test resulted in most of the features being significant. Even after the U-test (with a 0.03 threshold) and FDR (Benjamini-Hochberg) multiple test correction, multiple features were significant for at least Llama-versus-the-rest comparisons. Thus, we decided to only consider the features further that were significant for at least 3 pairs, resulting in 6 final features. These are illustrated in Figure 2.

The p-values and effect sizes are stored in our OSF repository[5]. Repeating Lemmas (RL) and Maximum Sentence similarity (MSS) scores appeared to be significant for 13 pairs, especially recurring for human, GPT, and Llama comparisons between each other and with other subsets. Foreign words are relevant for 9 comparisons, especially Mistral, Snoozy, and Llama. Positive sentiment probability showed relevance for 6 comparisons, Snoozy and Mistral in particular, while discourse markers were significant for 4 pairs, mostly GPT and Llama. Finally, analysis and comparatives were present in 3 pairs, always involving the human subset.

4.4 Model-agnostic analysis

A detailed graphical representation of the LIME results is shown in the Appendix, Figure 11. Looking at overall PoS tag significance, the only part of speech clearly biased towards human texts was interjections, while most frequently high-significance terms were nouns and proper nouns. Many high-score terms are recurrent and have similar biases in all 3 models. Foreign words tend to have an

[5] Models and linguistic features analysis can be found here: https://osf.io/uhd4a

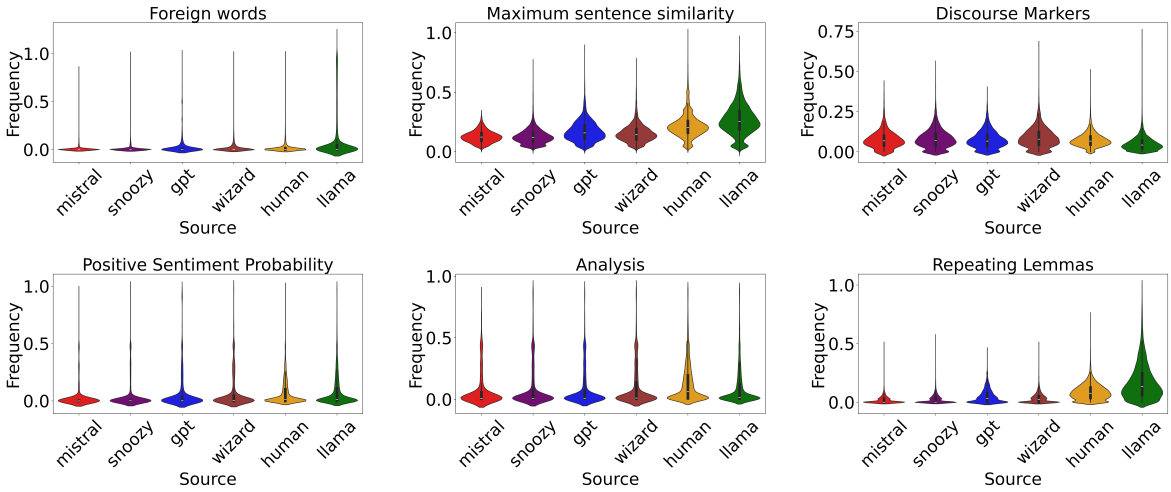


Figure 2: Distributions of the most significant linguistic features after a KW+U-test analysis.

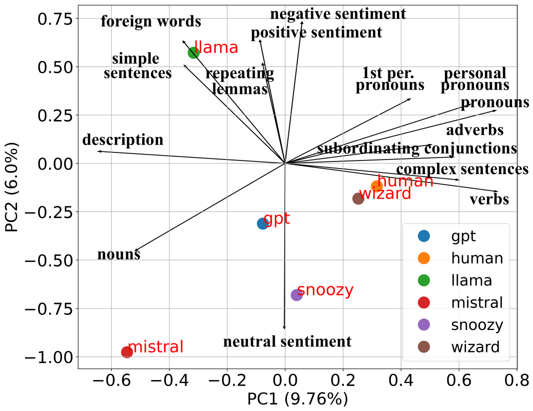


Figure 3: Top 15 loadings of linguistic features according to a PCA analysis for the overall data.

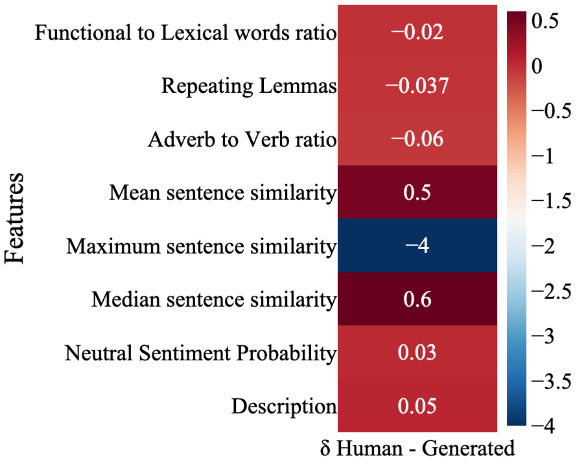


Figure 4: δ between human and aggregated generated feature averages. The values are z-score normalised.

AI-generated class bias (“bibliografia”, “explores”, “festival”, “including”, “explores”). Politically loaded terms are more often biased towards the human-written class (e.g. “berufsdemonstranten”, “kinderarmut”, “diktator”, “religionsfreiheit”, “asylbewerber”). Creative usages (“notgeil”, “idealfall”, “antispeziesismus”), words describing human subjective experiences (e.g. “träume”, “geschluckt”, “psychisch”, “gewichtszuname”) and proper nouns, possibly denoting usernames (“starfish1”, “mieep”, “deftone”, “theodosius”) have a strong human bias. The words with human bias also tend to be less frequent or fully absent from our training set (see Appendix, Figure 6).

4.5 Gradient analysis

As seen in the Appendix, Figure 12, almost all terms with the highest and lowest gradient magnitude, which indicate a bias towards one or the other class, are foreign words from French and English. Examples are “lyrics”, “interval”, “track” for model (a), “block”, “murders”, “windows”, “comment” for model (b) and “pendant”, “sound”, ”angle” for model (c). Another frequent category is proper nouns, mostly human names, especially not originally German ones (e.g. “Marcelo”, “Marlene”, “Scott”, “Nicole”, “Castro”), which seem to have a slight bias towards the AI-generated class, at least for models (a) and (c). Interestingly, more stereotypical German names and surnames like “Richter”, “Dietrich”, “Einstein”, “Brahms”, “Johannes”, “Philipp” and “Wolfgang” have strong human bias among all 3 models. This can be explained by the historic licence-free literature that is present in the human Wikipedia and Wikimedia genre.

Cities and countries are another frequent category. They are especially relevant for model (c), with 6 terms out of the top 30 terms being cities and countries. The cities might have a slight bias towards the AI-generated class, as *Vancouver*, *Freiburg*, *Manchester* and *Peking* are generated-biased, while only *Cadiz* is human-biased. The only city toponym for model (b) terms is *Mainz*, also biased towards the AI-generated class, while for model (a) we may, again, have the foreign/local division, as *Milan* is generated-biased and *Basel* and *Stuttgart* are human-biased. Organizations are also often reappearing (e.g. "Yahoo", "Telegram", "Windows, "Reuters", "Microsoft"), where only the first one appears to have a more human-directed bias. Abstract foreign nouns are more often associated with the AI-generated class: "terrorism", "irrational", "integration", and "proportional".

Using a Chi-Square model, we also verified if there was a correlation between bias of the term and how frequent it was in the training set. With a marginal p-value of about 0.05 for all 3 models, we rejected this hypothesis (see Appendix, Figure 7).

4.6 Token probability evaluation

Our results in Figure 5 show a clear distinction between token probability densities per group (see Section 3.4). The two groups of correctly classified examples constitute both outer ends of the density scale. The groups of misclassified samples show characteristics of a probability distribution typical for the respective opposite class. A lower density expresses, in the logic of this experiment, a wider range of next tokens, with, accordingly, lower probability values. Higher density means a narrow choice of next tokens. It follows that human texts have a more narrow selection of tokens to choose from, whereas predicting the next word in a generated text progression requires consideration of more possibilities.

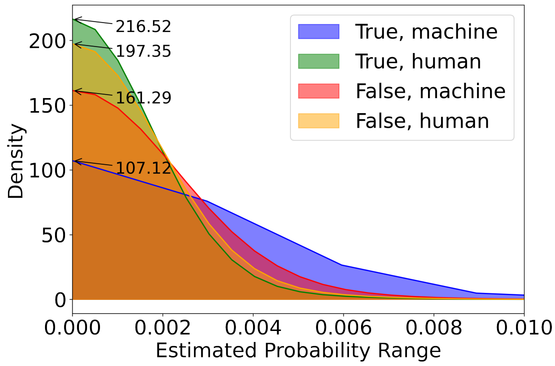


Figure 5: KDE plot showing the probability ranges of tokens grouped by correctly and wrongly classified human- and AI-generated text samples. The arrows point to the peak values of each distribution.

5 Discussion and Conclusion

Based on an in-depth analysis of detectability and features of AI-generated texts compared to human ones, different generative models appear to have strongly differing idiosyncrasies. According to our PCA analysis, fine-tuned versions of Llama 2, Wizard and Snoozy are extremely different from each other and their base model, while Wizard also seems to be imitating a human writing style the best. Generally, Llama 2 and Mistral are the greatest outliers. A classifier needs a large number of samples generated by Llama 2 to recognize this source well. The reduction of Llama samples from 22.500 to 3728 in the mixed-data model combined with the addition of other sources strongly degraded the model's performance on the Llama test set. The addition of Mistral data into the complete-data model does not seem to improve results significantly. While it possesses a better calibration on the overall set, Llama-data model (a) which was trained on a large amount of single-source samples seems to be the best calibrated for its test.

Hence, our results suggest that training a separate detector model for each generative source would lead to the most accurate detection, but this is neither cost-effective nor practical due to the rapid proliferation of new models. Grouping sources with similar distributions or significantly increasing the sample size for each subset may be viable solutions.

Foreign words are a recurrent feature throughout our analysis: they function as a statistically significant linguistic identifier, distinguishing Mistral, GPT and Llama texts. Many terms with high attribution and gradient scores are in general foreign words and names. Overall, many of our linguistic features capture lower-quality generations: repeating lemmas and highly similar sentences in one text, as well as overused adverbs and discourse markers. However, we can also see particular features of human texts. Low maximum sentence similarity suggests that human texts vary in sentence structure and content to avoid high levels of repetition. However, human writers also typically strive for a cohesive narrative or argument, reiterating certain points for emphasis or clarity, leading to a higher average sentence similarity. Human texts also more often contain positive sentiment, they are less descriptive and more analytical. However, as it can be seen from token probability analysis, generated texts are overall more random. This interesting characteristic may be explained by the fact that LLMs are trained on a large part of the internet data, while individual vocabulary is based on a single-life experience.

Although we can see some patterns in the high score terms from the interpretability analysis, overall, except for the foreign words and proper nouns, there is no major overlap between the explainability and linguistic analysis results. This suggests that many more linguistic categories could have been learnt. However, it might be an inherent limitation of semantic embeddings, which are powerful when applied to tasks involving explicit semantic differences between the classes, as in case of Sentiment Analysis or Topic Modeling. In contrast, the differentiation between generated and human content requires capturing more implicit indicators, that the models seemingly fail to consider. This is evident as the classifier defaults to the human class when uncertain, indicating limited learning of human text features. Future efforts should aim to enhance the transformer-based classifiers' capabilities in this regard.

Ethical Considerations

Several ethical considerations need to be considered in the context of the detection of AI-generated text, especially when we are dealing with texts produced by large language models. Firstly, the collection of texts for the training of the detection models needs to be fair and not in violation of privacy rights, especially if the texts contain personal or sensitive information. In our work, however, we made every attempt to minimize the impact of privacy issues by only using texts that were published and/or made publicly available by the respective authors and/or the entities that produced the texts. Secondly, there is certainly a risk of bias in our trained models since it was not possible to fairly consider all possible genres or styles of writing due to limitations in data collection. These biases have to be considered in the interpretation of the results. Thirdly, transparency is an important issue, in that, a precise account of how models were trained is provided. We strove to accomplish this through our detailed description in Section 3. Fourth, the ability to detect AI-generated text can be misused to suppress certain types of speech, or in contexts where anonymity is crucial, such as in political dissent, for example. It is unfortunately not possible for us to control how our proposed methods will be used, but it is an issue that we are aware of. Furthermore, if the detection of AI-generated texts is excessively promoted and overemphasized in the media, then this could potentially further erode the trust of society at large in digital communications.

Limitations

The detection of AI-generated text, especially when machine learning mechanisms are involved, is generally subject to certain limitations. As was confirmed by our work, detectors struggle to generalize across different types of generative models as well as data types. Since generative models tend to be constantly refined and re-tuned, detectors that were trained for particular LLMs will likely have to be updated and re-tuned as well. We anticipate that, with the ever-increasing quality of the text produced by LLMs, it will become harder and harder to distinguish between texts from LLMs and high-quality human-written texts. Furthermore, none of the proposed schemes is perfect. There is a non-negligible probability for false positives and false negatives, as reported via the precision and recall values in Tables 1 and 2.

The selection of the human data, sampled uniquely from before 2016, might induce a time domain shift that can be exploited by models, so while there is almost no other way to ensure that the data with weak labels is indeed human-written, it may have a negative effect, namely on models' capacity to generalise to current data.

Lastly, explainability is always a challenge in machine learning scenarios, including ours, even in light of the explainability results presented in Section 4.

Acknowledgements

The work on this paper was performed as part of the “noFake” project funded by the German Federal

Ministry of Research and Education (BMBF) under Award Identifier F16KIS1519.

References

Hosam Alamleh, Ali Abdullah S. AlQahtani, and AbdElRahman ElSaid. 2023. Distinguishing human-written and ChatGPT-generated text using machine learning. In *2023 Systems and Information Engineering Design Symposium (SIEDS)*. IEEE.

Yuvanesh Anand, Zach Nussbaum, Brandon Duderstadt, Benjamin Schmidt, and Andriy Mulyar. 2023. Gpt4all: An ecosystem of open source compressed language models. https://gpt4all.io/reports/GPT4All_Technical_Report_3.pdf.

Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, and Gemini team. 2024. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Internet Archive. 2023. Internet wayback machine. https://archive.org/. Accessed: October 23rd 2023.

Content at Scale. 2023. https://contentatscale.ai/ai-content-detector/. Accessed: September 13th, 2023.

Anton Bakhtin, Sam Gross, Myle Ott, Yuntian Deng, Marc'Aurelio Ranzato, and Arthur Szlam. 2019. Real or fake? learning to discriminate machine from human generated text. *Preprint*, arXiv:1906.03351.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Souradip Chakraborty, Amrit Singh Bedi, Sicheng Zhu, Bang An, Dinesh Manocha, and Furong Huang. 2023. On the possibilities of ai-generated text detection. *Preprint*, arXiv:2304.04736.

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. All that's 'human' is not gold: Evaluating human evaluation of generated text. *Preprint*, arXiv:2107.00061.

Copyleaks. 2024. https://copyleaks.com/. Accessed: January 5th, 2024.

Souvik Das, Sougata Saha, and Rohini Srihari. 2022. Diving deep into modes of fact hallucinations in dialogue systems. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 684–699, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

Edward, Alex (CEO, and CTO). 2024. Gpt-zero. https://gptzero.me/. Accessed: January 16th, 2024.

Keith D. Foote. 2023. A brief history of large language models. https://www.dataversity.net/a-brief-history-of-large-language-models/. Accessed: February 12th, 2024.

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

ZEIT ONLINE GmbH. 2023. Zeit online. https://www.zeit.de. Accessed: February 12th, 2024.

Jack Grieve and Helena Woodfield. 2023. *The Language of Fake News*. Elements in Forensic Linguistics. Cambridge University Press.

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

Springer Nature Gruppe. 2023. Springer nature. wissenschaftliche bücher. https://www.springernature.com/de. Accessed: February 12th, 2024.

Oliver Guhr, Anne-Kathrin Schumann, Frank Bahrmann, and Hans Joachim Böhme. 2020. Training a broad-coverage german sentiment classification model for dialog systems. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 1620–1625, Marseille, France. European Language Resources Association.

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. *Preprint*, arXiv:2301.07597.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. *Preprint*, arXiv:1706.04599.

Mengjie Guo, Limin Liu, Meicheng Guo, Siyuan Liu, and Zhiwei Xu. 2023b. Accurate generated text detection based on deep layer-wise relevance propagation. In *2023 IEEE 8th International Conference on Big Data Analytics (ICBDA)*. IEEE.

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.

Joseph D. Janizek, Pascal Sturmfels, and Su-In Lee. 2021. Explaining explanations: axiomatic feature interactions for deep networks. *J. Mach. Learn. Res.*, 22(1).

Ali Khan, Kathryn Brohman, and Shamel Addas. 2021. The anatomy of 'fake news': Studying false messages as digital objects. *Journal of Information Technology*, 37(2):122–143.

Sarah Kreps, R. Miles McCain, and Miles Brundage. 2020. All the news that's fit to fabricate: AI-generated text as a tool of media misinformation. *Journal of Experimental Political Science*, 9(1):104–117.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019a. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Chiara Longoni, Andrey Fradkin, Luca Cian, and Gordon Pennycook. 2022. News from generative artificial intelligence is believed less. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc.

Mingyang Lyu, Chenlong Bao, Jintao Tang, Ting Wang, and Peilei Liu. 2022. Automatic detection for machine-generated texts is easy. In *2022 IEEE Smartworld, Ubiquitous Intelligence & Computing, Scalable Computing & Communications, Digital Twin, Privacy Computing, Metaverse, Autonomous & Trusted Vehicles (SmartWorld/UIC/ScalCom/DigitalTwin/PriComp/Meta)*. IEEE.

Mistral-AI. 2023. Mistral transformer. `https://github.com/mistralai/mistral-src`. Accessed: January 18th, 2024.

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

Madhumita Murgia. 2023. Openai's red team: the experts hired to 'break' chatgpt. `https://www.ft.com/content/0876687a-f8b7-4b39-b513-5fee942831e8`. Accessed: April 14, 2023.

Alex Newhouse, Jason Blazakis, and Kris McGuffie. 2019. The industrialization of terrorist propaganda. the industrialization of terrorist propaganda neural language models and the threat of fake content generation. Last accessed on January 15th, 2024.

OpenAI. 2022. Introducing chatgpt. `https://openai.com/blog/chatgpt`. Accessed: January 16th, 2024.

OpenAI. 2023. Gpt-4 technical report. `https://cdn.openai.com/papers/gpt-4.pdf`. Accessed: January 16th, 2024.

Orginality.ai. 2024. `https://originality.ai/`. Accessed: January 5th, 2024.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Madeline Renbarger. 2023. How a 23-year-old college student built one of the leading ai detection tools. `https://www.businessinsider.com/`. Accessed: February 12th, 2024.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA. Association for Computing Machinery.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Kristina Schaaff, Tim Schlippe, and Lorenz Mindner. 2023. Classification of human- and AI-generated texts for English, French, German, and Spanish. In *Proceedings of the 6th International Conference on Natural Language and Speech Processing (ICNLSP 2023)*, pages 1–10, Online. Association for Computational Linguistics.

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

Veronika Solopova, Oana-Iuliana Popescu, Christoph Benzmüller, and Tim Landgraf. 2023a. Automated multilingual detection of pro-kremlin propaganda in newspapers and telegram posts. *Datenbank Spektrum*, 23(1):5–14.

Veronika Solopova, Eiad Rostom, Fritz Cremer, Adrian Gruszczynski, Sascha Witte, Chengming Zhang, Fernando Ramos López, Lea Plößl, Florian Hofmann, Ralf Romeike, Michaela Gläser-Zikuda, Christoph Benzmüller, and Tim Landgraf. 2023b. Papagai: Automated feedback for reflective essays. In *KI 2023: Advances in Artificial Intelligence*, pages 198–206, Cham. Springer Nature Switzerland.

taz Verlags u. Vertriebs GmbH. 2023. taz. die tageszeitung. https://taz.de/. Accessed: February 12th, 2024.

Edward Tian. 2023. https://gptzero.me/news/perplexity-and-burstiness-what-is-it. Accessed: January 5th, 2024.

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

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, and Amjad Almahairi et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Chih-Ming Tsai. 2023. Stylometric fake news detection based on natural language processing using named entity recognition: In-domain and cross-domain analysis. *Electronics*, 12(17).

Eduard Tulchinskii, Kristian Kuznetsov, Laida Kushnareva, Daniil Cherniavskii, Serguei Barannikov, Irina Piontkovskaya, Sergey Nikolenko, and Evgeny Burnaev. 2023. Intrinsic dimension estimation for robust detection of ai-generated texts. *Preprint*, arXiv:2306.04723.

Junlin Wang, Jens Tuyls, Eric Wallace, and Sameer Singh. 2020. Gradient-based analysis of NLP models is manipulable. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 247–258, Online. Association for Computational Linguistics.

Writer. 2023. https://writer.com/ai-content-detector/. Accessed: September 13th, 2023.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023. Wizardlm: Empowering large language models to follow complex instructions. *Preprint*, arXiv:2304.12244.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 9054–9065. Curran Associates, Inc., Red Hook, NY, USA.

ZeroGPT. 2024. Trusted gpt-4, chatgpt and ai detector tool by zerogpt. https://www.zerogpt.com/. Accessed: Febuary 19th, 2024.

Wikimedia Deutschland – Gesellschaft zur Förderung Freien Wissens e.V. 2008. Wikipedia dump. https://dumps.wikimedia.org/other/static_html_dumps/current/de/. Accessed: October 23rd 2023.

Wikimedia Deutschland – Gesellschaft zur Förderung Freien Wissens e.V. 2024a. Wikimedia. the free media repository. https://commons.wikimedia.org/wiki/Main_Page. Accessed: February 12th, 2024.

Wikimedia Deutschland – Gesellschaft zur Förderung Freien Wissens e.V. 2024b. Wikipedia. die freie enzyklopädie. https://www.wikipedia.de/. Accessed: February 12th, 2024.

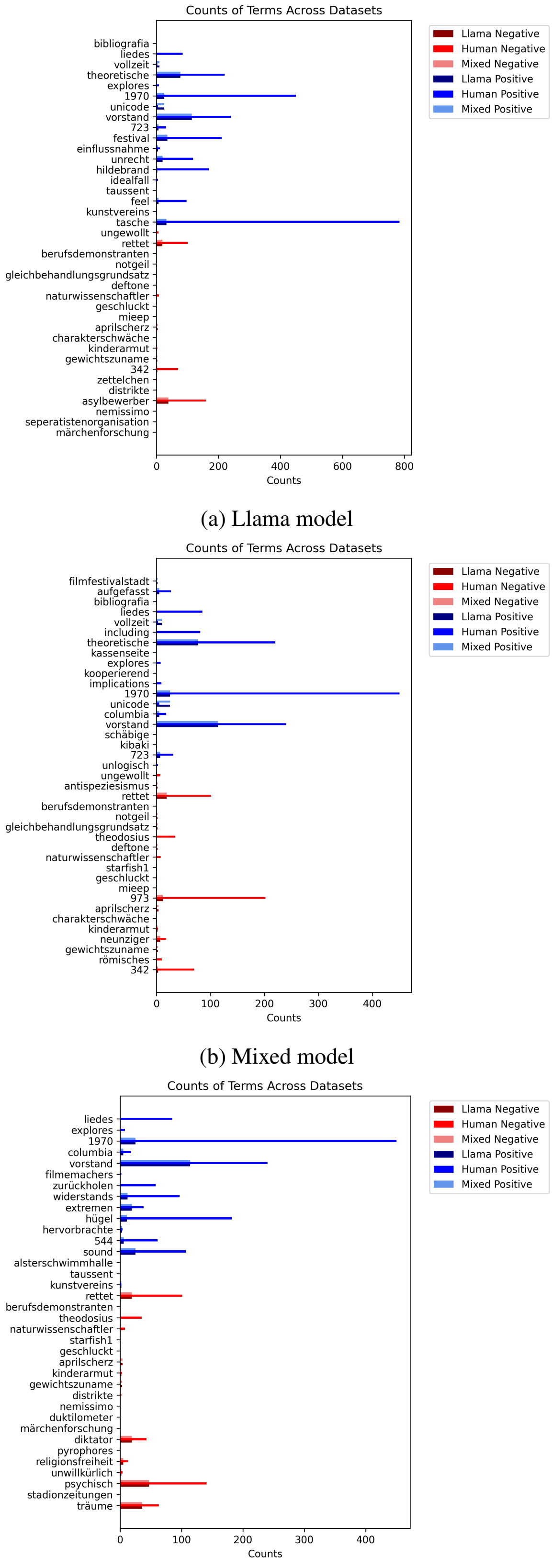
A Appendix

Name	TAZ	Springernature	Wikipedia	Wikimedia	Zeit Online
Llama	5000	5000	5000	5000	5000
Wizard	1806	1026	1893	-	2181
Snoozy	4162	985	734	-	1266
GPT3.5	800	800	800	800	800
Human	5000	5000	5000	5000	5000
Mistral	471	-	-	-	-

Table 3: Overall number of samples per genre and source. While Llama 2 and human samples are present for all of the genres, the rest of the sources are only collected based on the experimental set-up described in Section 3.3

Set/subset	Llama	Mixed	Human	Mistral	Gpt-wiki
Model (a) train	22.500	-	22.500	-	-
Model (a) test	2500	-	2500	-	-
Model (b) train	3728	18772	22.500	-	-
Model (b) test	373	1877	2500	-	-
Model (c) train	22.500	22.500	22.500	249	655
Model (c) test	2500	2500	2500	222	146
Mistral test	-	-	2500	471	-
GPT Wikimedia	-	-	2500	-	801

Table 4: Number of samples per source in train and test sets for each model.

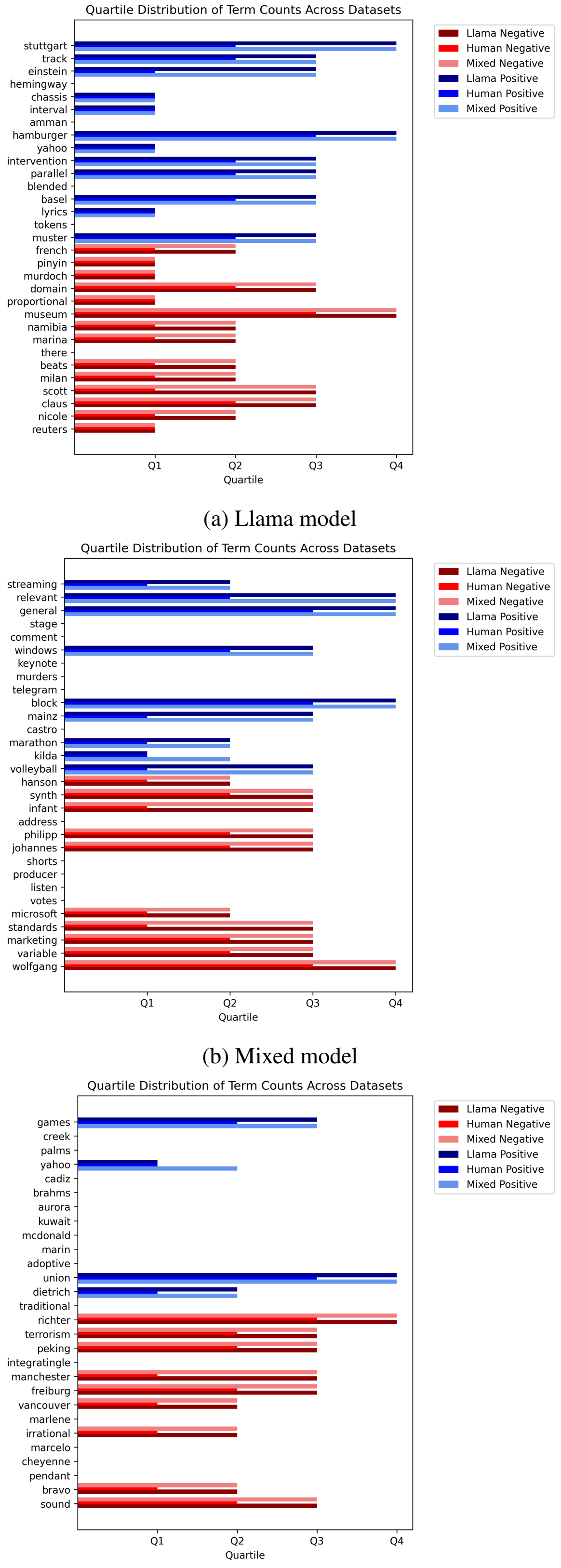


(a) Llama model

(b) Mixed model

(c) Complete model

Figure 6: Counts of important terms for LIME experiment.



(a) Llama model

(b) Mixed model

(c) Complete model

Figure 7: Counts of important terms from gradients experiment.

	Predicted	
Actual	HW	MG
HW	2307	193
MG	59	2441

(a) Llama dataset + HW dataset

	Predicted	
Actual	HW	MG
HW	2307	193
MG	1830	670

(b) Mixed dataset + HW dataset

	Predicted	
Actual	HW	MG
HW	2307	193
MG	582	219

(c) GPT Wikimedia + HW dataset

	Predicted	
Actual	HW	MG
HW	2307	193
MG	389	82

(d) Mistral Wikipedia + HW dataset

Figure 8: HW means human-written data, and MG means AI-generated data. Llama model (a) confusion matrices.

	Predicted	
Actual	HW	MG
HW	2227	273
MG	828	1672

(a) Llama dataset + HW dataset

	Predicted	
Actual	HW	MG
HW	2227	273
MG	103	2397

(b) Mixed dataset + HW dataset

	Predicted	
Actual	HW	MG
HW	2227	273
MG	571	230

(c) GPT Wikimedia + HW dataset

	Predicted	
Actual	HW	MG
HW	2227	273
MG	323	148

(d) Mistral Wikipedia + HW dataset

Figure 9: Mixed model (b) confusion matrices.

	Predicted	
Actual	HW	MG
HW	3679	1301
MG	24	4611

(a) Llama dataset + HW dataset

	Predicted	
Actual	HW	MG
HW	3679	1301
MG	38	3985

(b) Mixed dataset + HW dataset

	Predicted	
Actual	HW	MG
HW	3679	1301
MG	23	123

(c) GPT Wikimedia + HW dataset

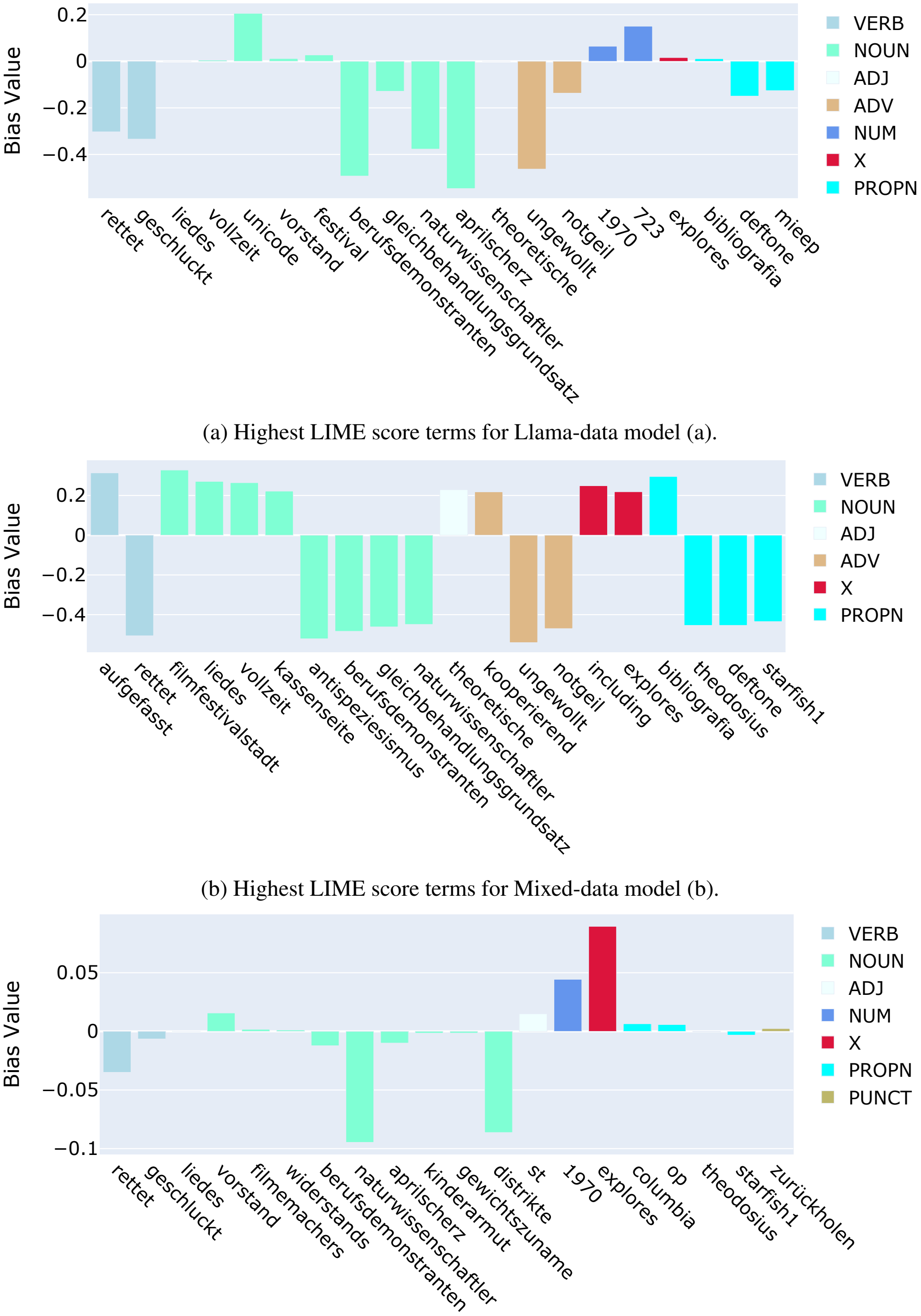
	Predicted	
Actual	HW	MG
HW	3679	1301
MG	2	220

(d) Mistral Wikipedia + HW dataset

Figure 10: Complete model (c) confusion matrices.

Table 5: Full list of Linguistic Features.
The results of the statistical testing of the features can be found in the abovementioned OSF repository.

Morphology	**Syntax**	**Semantics**
Adjectives	Clause of Purpose	Voc: Remembering
Nouns	Clause of Reason	Voc: Understanding
Verbs (Finite, Infinitives, Stative)	Clause of Condition	Voc: Application
Comparatives	Consecutive clause	Voc: Evaluation
Superlatives	Complex sentences	Voc: Analysis
Adverbs	Simple sentences	Voc: Creation
Adjective to Noun ratio	Relative clause	Voc: Assertion
Adverb to Verb ratio	Modal clause	Voc: Cognition
Abstract nouns	Concessive clause	Description
Passive voice	Adversative clause	Evaluation
Pronouns (All types)	1st person + finite verb	Analysis
Modal verbs	Subordinating conjunctions	Conclusion
Negations	Coordinating conjunctions	Positive Sentiments
Subjunctive mood	Questions	Negative Sentiments
Foreign words		Neutral Sentiments
Present, Past, Future		High modality words
1st person pronouns		Feelings
2nd person pronouns		Supports
Indefinite pronouns		Claims
		Future Actions
Generation Errors		
Mean sentence similarity		
Maximum sentence similarity		
Median sentence similarity		
Repetitive Lemmas		
Sentence Length Variation		
Functional to Lexical words ratio		

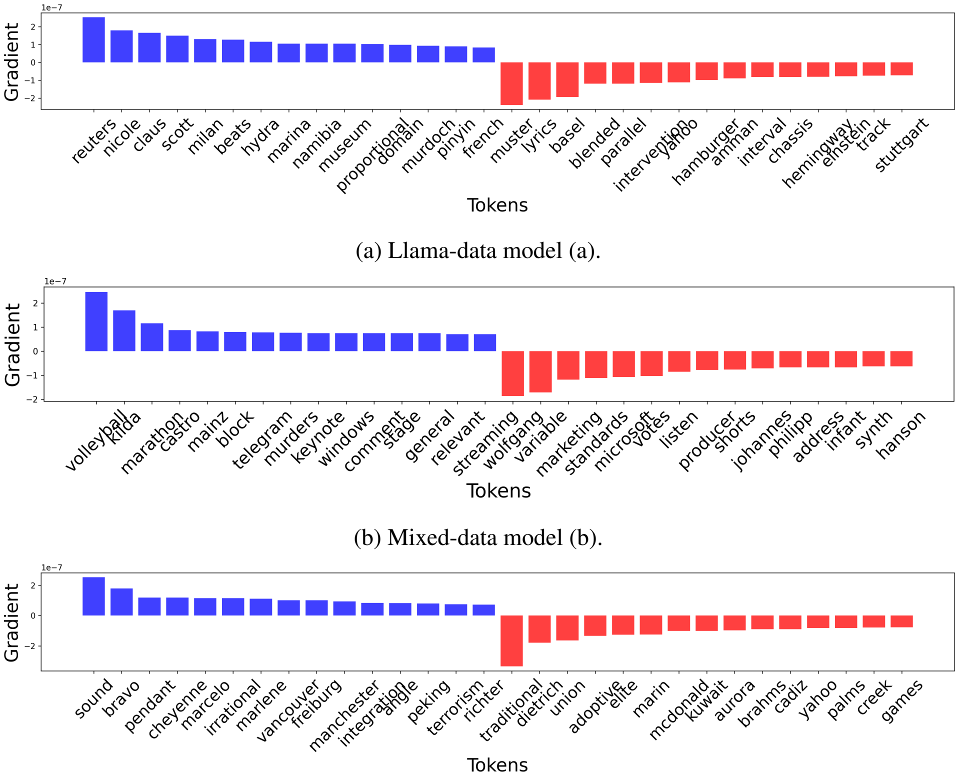


(a) Highest LIME score terms for Llama-data model (a).

(b) Highest LIME score terms for Mixed-data model (b).

(c) Highest LIME score terms for Complete-data model (c).

Figure 11: LIME analysis results.



(a) Llama-data model (a).

(b) Mixed-data model (b).

(c) Complete-data model (c).

Figure 12: Highest gradient magnitude terms.