BUST: Benchmark for the evaluation of detectors of LLM-Generated Text

Joseph Cornelius[†*] **Oscar Lithgow**[†*] **Sandra Mitrović**[†] **Ljiljana Dolamić**[‡] **Fabio Rinaldi**[†]
[†] Dalle Molle Institute for Artificial Intelligence Research (IDSIA), Switzerland
[‡] armasuisse, Science & Technology, Switzerland
{joseph.cornelius,oscarwilliam.lithgow,sandra.mitrovic,fabio.rinaldi}@idsia.ch
ljiljana.dolamic@armasuisse.ch

Abstract

We introduce BUST, a comprehensive benchmark designed to evaluate detectors of texts generated by instruction-tuned large language models (LLMs). Unlike previous benchmarks, our focus lies on evaluating the performance of detector systems, acknowledging the inevitable influence of the underlying tasks and different LLM generators. Our benchmark dataset consists of 25K texts from humans and 7 LLMs responding to instructions across 10 tasks from 3 diverse sources. Using the benchmark, we evaluated 5 detectors and found substantial performance variance across tasks. A meta-analysis of the dataset characteristics was conducted to guide the examination of detector performance. The dataset was analyzed using diverse metrics assessing linguistic features like fluency and coherence, readability scores, and writer attitudes, such as emotions, convincingness, and persuasiveness. Features impacting detector performance were investigated with surrogate models, revealing emotional content in texts enhanced some detectors, yet the most effective detector demonstrated consistent performance, irrespective of writer's attitudes and text styles. Our approach focused on investigating relationships between the detectors' performance and two key factors: text characteristics and LLM generators. We believe BUST will provide valuable insights into selecting detectors tailored to specific text styles and tasks and facilitate a more practical and in-depth investigation of detection systems for LLM-generated text.

1 Introduction

In recent years, the development of large language models (LLMs) has surged, with notable releases such as GPT-4, Llama-2 and Falcon. As these models have advanced, distinguishing between texts written by humans and texts generated by LLMs has become challenging. Several studies show that most adults cannot distinguish between the two (Clark et al., 2021; Jakesch et al., 2023). This indistinguishability poses several challenges, ranging from the potential for fraudulent authoring to the automatic dissemination of targeted misinformation (Pan et al., 2023; Yu et al., 2023). There are two main approaches to identifying generated text - watermarking and detection systems - with the focus of this paper on the latter. Watermarking relies on the developer's watermark, whereas detection systems can be developed independently. However, the effectiveness and feasibility of current detectors in identifying machine-generated text (MGT) remains an active area of research. To effectively

CONTEXT: Belize, formerly known as British Honduras, is a Caribbean country located on the northeastern coast of Central America. Belize is bordered on the northwest by Mexico, on the east by the Caribbean Sea, and on the south and west by Guatemala. It has an area of 22,970 square kilometres (8,867 sq mi) and a population of 408,487 (2019). Its mainland is about 290 km (180 mi) long and 110 km (68 mi) wide. It has the lowest population and population density in Central America.

INSTRUCTION: Given this short paragraph about Belize, how many people live there and what is notable about its population density?

HUMAN: Belize has a population of 408,487 people and has the lowest population density in Central America.

FALCON 7B: The population of Belize is around 408,487 according to the Worldometer statistics. This figure is the estimated population of the country as of year 2019. Belize is one of the smallest countries in Central America with a population density of 38.93 people per square kilometers.

LLAMA-2 70B: According to the information provided, 408,487 people live in Belize. Belize has the lowest population and population density in Central America.

GPT-4: According to the paragraph, Belize has a population of 408,487 as of 2019. What is notable about its population density is that it has the lowest population and population density in Central America.

Figure 1: BUST dataset excerpt for the closed QA task: Presents a **HUMAN** response and machine-generated texts (e.g., **FALCON 7B**, **LLAMA-2 70B**, **GPT-4**) for a given **CONTEXT** and **INSTRUCTION**.

[*]Equal contribution, with Cornelius focusing on dataset creation and Lithgow on evaluation and data repurposing.

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 8026–8054

evaluate these detection systems, there is an urgent need for an adaptable benchmarking dataset that can cope with the influx of LLMs and covers various tasks (Tang et al., 2023).

In this paper, we introduce a comprehensive dataset[1] tailored for benchmarking detectors against instruction-tuned LLMs. This dataset consists of 3,180 instructions paired with corresponding responses from both humans and LLMs, resulting in a corpus of over 25,000 texts. A systematic evaluation of five distinct detectors was conducted, revealing notable variances in performance across the range of tasks. Additionally, we have conducted a meta-analysis, examining the dataset's characteristics and employing surrogate models to investigate the features that most significantly impact detector performance. Our findings indicate that while emotional content in texts enhanced the performance of some detectors, the most effective detector demonstrated consistent performance, irrespective of the writer's attitudes. This benchmark[2] serves as a valuable tool for identifying the most suitable detectors for specific text styles and tasks.

The paper is organized as follows: Section 2 gives an overview of related work, focusing on detector systems and analog datasets. Section 3 details the specifics of our proposed dataset and the generators used. Section 4 explains the metrics used for comparing different generators and the used detector systems, while section 5 presents our results. In section 6 we present two ablation studies and a discussion on the implications of our results is given in section 7. The paper concludes with section 8, where we summarize our contributions and make suggestions for future research.

2 Related Work

Following, we present related work on generated text detectors and datasets for MGT.

Generated Text Detectors MGT detection has evolved significantly over the years. Initial efforts focused on detecting differences in linguistic features and using statistical measures such as perplexity, relative entropy, and style similarities (Beresneva, 2016). Gehrmann et al. (2019) provided a suite of baseline statistical methods known as "GLTR" to help humans detect MGT without prior training. Recent developments like DetectGPT (Mitchell et al., 2023) define a curvature-based criterion using log-likelihoods computed by the model in question, circumventing the need to train a separate classifier, collect a dataset of real or generated passages, or explicitly watermark the generated text. Similarly, Su et al. (2023) presented DetectLLM, a zero-shot approach detecting a range of LLMs using log-rank information.

Another prevalent strategy in this area involves the detection of MGT as a classification task. In this context, various pre-trained language models such as Vicuna (Chiang et al., 2023) or RoBERTa (Guo et al., 2023) are fine-tuned on datasets that combine human-written with their corresponding MGT examples. Notably, many of these models are specialized for specific tasks, as GROVER (Zellers et al., 2019). Others are tailored to specific domains (Koike et al., 2023) or aim to detect output from specific generator models (Guo et al., 2023).

LLM generated text datasets The datasets used to evaluate systems capable of distinguishing between human-written and MGT vary considerably. Some developers of detector systems have chosen to create their own datasets for this purpose (Radford et al., 2019; Gehrmann et al., 2019; Zellers et al., 2019; Verma et al., 2023). However, publicly available datasets have often been created for very specific purposes. In particular, some are designed for specific generator models (Fagni et al., 2021; Guo et al., 2023). Other datasets are limited to a single task (Guo et al., 2023), while some are specific to a particular domain, such as the CHEAT dataset (Yu et al., 2023), which solely focuses on generating academic abstracts.

The TuringBench dataset (Uchendu et al., 2021) provides a comprehensive benchmark containing news articles generated by 19 different models, based on their titles, and juxtaposed with the corresponding human-authored articles. However, the text-generating capabilities of the models in this dataset are not on par with the advanced capabilities of current LLMs. Furthermore, Clark et al. (2021) presented a dataset aimed at investigating the human ability to distinguish between texts written by humans and machines. Notably, this dataset spans multiple domains and incorporates various models, though it is limited by having only 50 samples for each generator. More recently, datasets encompassing a wider variety of generator models have been introduced, yet they tend to specialize in a limited number of tasks. These include multilingual news writing (Macko et al., 2023), question answering (He et al., 2023), and article and review writing

[1] https://github.com/IDSIA-NLP/BUST
[2] https://bust.nlp.idsia.ch

Source Dataset	Task	Human	Machine	**Total**
Dolly Databricks	Brainstorming	250	1734	**1984**
Dolly Databricks	Classification QA	250	1724	**1974**
Dolly Databricks	Closed QA	250	1707	**1957**
Dolly Databricks	Creative Writing	250	1738	**1988**
Dolly Databricks	General QA	250	1724	**1974**
Dolly Databricks	Information Extraction	250	1722	**1972**
Dolly Databricks	Open QA	250	1719	**1969**
Dolly Databricks	Summarization	250	1698	**1948**
Amazon Reviews	Review writing	898	6251	**7149**
Twitter Post	Stance writing	282	1969	**2251**
Total		**3180**	**21986**	**25166**

Table 1: Statistical information on our parallel human-written and machine-generated dataset, divided into the different source datasets and tasks.

(Wang et al., 2023). In comparison, our BUST dataset incorporates diverse tasks, enabling a more comprehensive analysis of detection systems' capabilities. A comprehensive overview of the available datasets in this domain is given in Appendix Tab. 5. In contrast to other studies, our research focuses exclusively on instruction-tuned models, which the public is increasingly using through prompting interfaces.

3 Dataset

This section describes in detail the datasets used in our research. For a comprehensive overview of these datasets, see Tab. 1. Our collected comparison corpus includes paired entries of partially contextualized instructions, as well as both human-written and MGT responses. Multiple MGT responses are available for each instruction, generated by different models, as shown in Fig. 1. The specific models from which these responses are derived vary in complexity and range in size from 7B to over 100B parameters. Details of the text generation models can be found in Appendix Tab. 10.

For a comprehensive comparison and robust analyses, we included models in different configurations: standard models, models we further trained ourselves, and models with or without Parameter-efficient Fine-tuning (PEFT).

3.1 Dataset creation

This study uses a dataset obtained from a subset of the publicly available Databricks Dolly dataset (Conover et al., 2023). The dataset is characterized by a large variety of human writing styles with 5000 annotators. In addition, the dataset is divided into 8 different tasks, which allows for differential benchmarking across these categorical variations.

In constructing our benchmark dataset, we initially selected 2000 samples from the Databricks Dolly dataset using a stratified random sampling technique to ensure representativeness across different tasks. Then we prompted 7 distinct Large Language Models (LLMs) with the instructions of these examples to generate the MGT responses.

To further examine detectors' adaptability to distributional variations, we enhanced our dataset by integrating two additional datasets, specifically repurposed for this study, which will be described in the following section.

3.1.1 Repurposed review and stance datasets

While repurposing the reviews dataset for our study, we strategically leveraged existing datasets, emphasizing those with the potential inclusion of emotional or biased statements to evaluate how detectors perform in such specific scenarios. We utilized an Amazon review dataset (Ni et al., 2019) focused on magazine subscriptions, spanning from 2014 to 2018, amounting to 900 reviews after meticulous selection and linkage with product descriptions. Additionally, we incorporated the CovidLies dataset (Hossain et al., 2020; Chen et al., 2020), featuring tweets expressing stances on specific topics. In the case of the review dataset, the processing pipeline involved applying sentiment prediction and keyword extraction. We mapped 5-star review scores to a textual scale, ranging from "Very unsatisfied" to "Very satisfied." Using sentiment, keywords, mapped satisfaction levels, and product descriptions, we generated hypothetical instructions formatted similarly to the Databricks Dolly dataset, which underwent the same steps of synthetic generation. In the case of CovidLies, because the dataset already provided the Tweet's stance, we extracted mentions and hashtags to provide them as guiding information within the prompt.[3]

Similar to the Databricks Dolly dataset, the repurposed datasets include a wide variety of authors, as the majority of reviews and tweets are written by different users. By refraining from quality-based filtering, the datasets ensure coverage of the full spectrum of human expression and provide a broad, realistic picture of different language styles and writing methods.

4 Methodology

4.1 Dataset meta-analysis

Our research is primarily concerned with providing a careful evaluation and comparison of human-

[3]Appendix Tab. 9 provides a glimpse of a few examples and Appendix Fig. 5 shows an overview of the process.

written and MGT responses. To ensure the completeness and reliability of our results, we used multiple evaluation metrics, each tailored to specific aspects of the data. This section explains the methods and metrics used during our evaluation process. **Linguistic features** and statistical differences serve as basic markers for evaluating the generated responses' quality, fluency, and coherence. The features used for our assessment are listed in Appendix Tab. 6. **Readability** indices are an empirical measure of the ease with which a reader can understand a written text. We use various readability indices with different emphases listed in Appendix Tab. 7. **Writer's attitudes** offer indirect assessments of the author's attitude by categorizing the text based on dimensions like convincingness, persuasiveness, irony, and emotions. The objective is to gain interpretable insights into whether the attitudes expressed in the texts influence the behavior of detectors. In our approach, we utilized off-the-shelf models to infer attitudes.[4] **LIWC-based features** are linguistic and psychometric features obtained using LIWC[5] (Boyd et al., 2022), a well-known academically developed software, providing quantitative insights of not only linguistic and grammar (such as particular punctuation, time orientation etc.) but also other more sophisticated aspects of written text such as cognition, perception, social, motivation.

The rationale for selecting linguistic, readability, writer's attitude, and LIWC-based features was to bring explainability to detector results. These features were chosen for their ability to reflect diverse perspectives on texts and their source.

4.2 Detectors

In our benchmark, we evaluated a diverse set of black-box detector models, including both publicly available and proprietary systems, as well as those developed for commercial and non-commercial purposes. Each detector was tested in a real-world zero-shot scenario in which they were asked to distinguish between human-written and MGT, without any prior fine-tuning on specific examples. The XLMR_ChatGPT (Antoun et al., 2023) and ChatGPT_QA (Guo et al., 2023) detectors are based on RoBERTa (Liu et al., 2019) and were fine-tuned using the HC3 dataset. In contrast to ChatGPT_QA, the XLMR_ChatGPT detector was also trained on out-of-domain data to improve its resilience against common attack methods. Meanwhile, the training of the ChatGPT_QA model was limited to question-answer pairs extracted from the full text of the HC3 corpus to obtain a distinct knowledge domain. Additionally, we used the RADAR_Vicuna7B (Hu et al., 2023) model developed for robust MGT detection through adversarial learning. This learning process simulates an adversarial game with two players: a paraphraser that attempts to generate realistic text that can evade the detector and a detector that attempts to thwart such attempts. In contrast, the LLMDet (Wu et al., 2023) detector specializes in identifying the origin of a given text and efficiently distinguishing between several LLMs, such as GPT-2, OPT, LLaMA, and their human-written counterparts. It uses the next-token probabilities of salient n-grams and proxy perplexity to make its decisions. Lastly, we looked at the widely used commercial system GPTZero[6]. Since it is a closed-source platform, the details of GPTZero's model architecture, training datasets, and methods are not publicly available, adding a layer of complexity to evaluating its detection efficacy.

4.2.1 Detectors evaluation

Detector performance is evaluated using standard classification metrics, considering LLM-generated as the positive class. Due to the imbalanced nature of our dataset, we used F1-macro as the primary metric supplemented by the Matthews Correlation Coefficient (MCC). F1-macro provides a balanced assessment of precision and recall across all classes, and MCC is robust to skewed distributions.[7] The differing outputs of the various detectors are handled as follows: The LLMDet model outputs probabilities for the text being generated by several specific LLMs, which we merged into a single machine-generated category or human-generated; here, we took "1 - (score of human-generated)". For the Radar_Vicuna7B model, which outputs the probability that a text is machine-generated, we labeled text as machine-generated if the probability exceeded 0.5, using the probability as a value. ChatGPT_QA and XLMR_ChatGPT detectors make binary decisions, using the provided score if the detector labels the text as MGT or 1-score in the opposite case. The evaluation is performed on different levels, considering different generative models, tasks and detectors. Additionally, we perform correlation analysis and build surrogate models to

[4]The models used are listed in Appendix Tab. 8.

[5]https://www.liwc.app/

[6]https://www.gptzero.me

[7]Additional metrics are reported in Appendix A.

simulate detector behavior.

Correlations between detectors' predictions and dataset features We conducted a multifaceted analysis to comprehensively evaluate detector performance and gain insights into their behavior across various scenarios. Firstly, we computed correlations between detector systems and dataset features. This approach allows us to uncover patterns in performance. Our goal was not only to assess raw performance but also to understand better how detectors respond in different contextual settings. Furthermore, we aimed to unveil interpretable associations between detector performance and text features at varying levels, ranging from lexico-syntactic attributes to nuanced expressions of writers' attitudes.

In our categorical correlation analysis, we employed Chi-tests to examine the relationships between detector predictions (human or MGT) and the categorized attitudes of writers. For a more detailed exploration, we also included correlation assessments on continuous features. This involved evaluating relationships between textual features (such as length and word count), softmax outputs representing writers' attitudes versus the prediction scores generated by the detectors.

Surrogate models As noted, detectors are not only black-box models but often also proprietary, meaning many aspects of their inner workings are not publicly disclosed. We simulate detector behavior using surrogate models to understand different detector outcomes on our benchmark dataset. Each surrogate model is a classifier that ingests all mentioned linguistic, attitudinal and psychometric features and uses a particular detector label (human vs. MGT) as the target. To make the surrogate model explainable, we use the Gradient Boosting classifier as it provides insight into variable importance. We construct multiple surrogate models while also considering different tasks and generative models.

5 Results

5.1 Dataset Meta-analysis

Readability scores of texts produced by generative models vary greatly depending on the chosen readability metric, although, in general, for the majority of readability metrics, texts generated by Falcon-7b-Dolly obtain on average quite different scores from those of other generative models.

Linguistic features The greatest variations between generators could be observed in the generated text length, where the longest texts are generated by Guanaco-7b, LLaMA-2-7b*, and GPT-4.

LIWC features The greatest differences between generated texts could be noticed in terms of authentic, analytic, tone and clout scores, with GPT-4 texts having the highest analytic and authentic scores. Less pronounced are differences in perception, cognition, drives and social aspects. While GPT-4 texts still achieve the highest perception scores, the highest drive scores are observed in GPT-3.5. With respect to cognition, GPT-4 is closely followed by LLaMA-2-70b* and LLaMA-2-7b*, while in terms of social aspects, the latter two take supremacy over the competitors. Minor differences were noticed in conversation scores where LLaMA-2-7b* comes closest to humans. Additional figures for readability, linguistic, and LIWC features are reported in the Appendix A.5.

Writers attitude Analyzing irony and emotional expression in human-written text versus language models (LLMs) reveals interesting patterns. Notably, human text tends to contain more irony than LLM-generated texts, with GPT-4 exhibiting the least ironic undertones (see Appendix Fig. 17). In terms of emotions, a comparison shows distinct trends. LLaMA-2-7b* prominently expresses curiosity, while Falcon-7b-Dolly emerges as the model conveying the highest levels of remorse, making it the "saddest" among all the models. Remorse and admiration are notably over-expressed by Falcon-7b-Dolly and GPT-3.5. GPT-4 and Guanaco-7b, on the other hand, stand out for conveying excitement, while Guanaco-7b surpasses both humans and other models in overexpressing the love emotion (see Appendix Fig. 15).

Turning to convincingness, in general, LLMs produce arguments perceived with higher quality (being relevant, clear, and impactful) compared to humans. Notably, GPT-3.5 deviates the most from the human baseline in this regard (see Appendix Fig. 14). Regarding persuasiveness, all generators tend to generate text identified as persuasive more frequently than human-generated text. GPT-4 closely aligns with the persuasive trends observed in human-generated content (see Appendix Fig. 18). Though a direct correlation between these results and the ability of various LLMs to replicate human expression's attitude elements is not established, these results could shed light on detectors' performance variations when assessing texts generated by diverse generators.

Task	ChatGPT_QA	LLMDet	Radar_Vicuna7B	XLMR_ChatGPT
* (All-tasks)	0.607	**0.626**	0.549	0.547
brainstorming	**0.686**	0.647	0.519	0.662
classification	0.509	**0.577**	0.474	0.487
closed_qa	0.537	**0.597**	0.476	0.543
creative_writing	**0.713**	0.611	0.661	0.644
general_qa	**0.781**	0.608	0.585	0.745
information_extraction	0.428	**0.556**	0.493	0.436
open_qa	**0.723**	0.616	0.522	0.685
review_generation	0.652	**0.701**	0.612	0.504
stance_generation	0.439	**0.574**	0.471	0.388
summarization	**0.593**	0.562	0.513	0.544

Table 2: Detectors performance in general (**ALL**) and **per task** in terms of F1-macro.

5.2 Detectors evaluation

Among the detectors evaluated (see Tab. 2), LLMDet emerges as the top performer in terms of F1-macro, showcasing not only the best overall performance but also remarkable stability across various tasks. Following closely is ChatGPT_QA, securing the second position, particularly excelling in general_qa, open_qa, and creative_writing tasks, outperforming LLMDet by 28%, 17%, and 16%, respectively. Notably, XLMR_ChatGPT mirrors the performance pattern of ChatGPT_QA, except for struggles in the stance and review_generation tasks. On the other end, Radar_Vicuna7B ranks as the least effective detector overall, but its strengths surface in information_extraction, review, and stance_generation tasks. Shifting the focus to assessing the performance depending on the source text generator, ChatGPT_QA stands out with a clear advantage in texts generated by Guanaco-7b, Falcon-7b-Dolly, and GPT-3.5 compared to other detectors. Conversely, Radar_Vicuna7B appears to struggle the most to detect GPT-4 generated texts (see Fig. 2 and Appendix Fig. 21).

5.3 Correlation analysis

On the categorical correlation between detector predictions and writers' attitudes, we did not observe significant correlations when considering the entire dataset. However, on a task-specific level (see Appendix Fig. 22), Radar_Vicuna7B is the most correlated detector. It significantly correlated with at least two writer's attitudes in 8 out of 10 tasks. Despite its high correlation, Radar_Vicuna7B exhibited the poorest performance. Conversely, LLMDet, the least correlated overall, demonstrated stability across tasks. It only showed significant correlations with two or more attitudes in 3 of the 10 tasks, and half of the tasks exhibited no correlation at all. Persuasiveness was the most correlated attitude among detectors, particularly in information_extraction, closed_qa, creative_writing, and review_generation.

Assessment of the continuous correlation between detector prediction scores and linguistic, attitudinal, and psychometric variables. Spearman correlation scores between different detector prediction scores and input variables are generally low. ChatGPT_QA has absolute correlation scores >0.2 with the highest number of input variables, mostly linguistic and LIWC features. On the contrary, LLMDet has absolute correlation scores >0.2 with only 5 input variables, out of which 4 are linguistic. In particular, 3 linguistic features, number of words, number of sentences and percentage of unique words, are correlated with all detector scores; however, they are not in the same direction. Apart from linguistic and LIWC features, XLMR_ChatGPT detector scores are also correlated with approval and convincingness features (see Appendix Fig. 19).

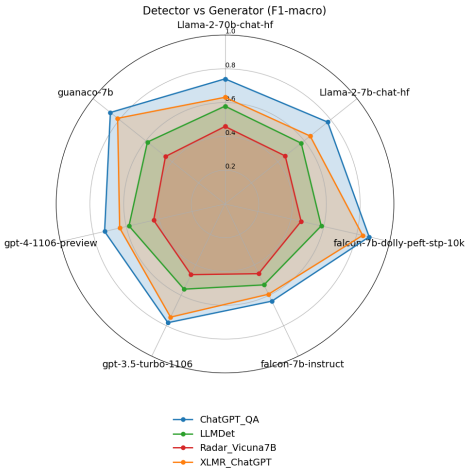

Figure 2: Detectors performance on detecting text produced by different generators (F1-macro)

5.4 Surrogate models results

Overall performance Our surrogate models were implemented using XGBoost library with 10 repetitions to ensure the robustness of the results. The performance of surrogate models on the whole dataset (considering all tasks and all generative models) depends on the performance metric (see Tab. 3). Regarding F1-macro and MCC scores, which are also more important given the existing class imbalance, the best surrogate is the one for ChatGPT_QA, while the second best is the surrogate for XLMR_ChatGPT. The surrogates for the other two detectors reflect remarkable fluctuations in the F1-macro score and, hence, are deemed less reliable, although the surrogate for Radar_Vicuna7B performs best in terms of F1-weighted and F1-micro. Looking at the features

Detector	F1-weighted	F1-macro	F1-micro	ROC AUC	MCC
ChatGPT_QA	0.871 (0.003)	**0.854 (0.003)**	0.862 (0.003)	**0.936 (002)**	**0.711 (0.005)**
XLMR_ChatGPT	0.842 (0.001)	*0.840 (0.001)*	0.842 (0.001)	0.923 (0.002)	*0.680 (0.003)*
Radar_Vicuna7B	**0.930 (0.002)**	0.664 (0.007)	**0.944 (0.001)**	*0.930 (0.004)*	0.396 (0.015)
LLMDet	*0.909 (0.002)*	0.703 (0.006)	*0.922 (0.002)*	0.915 (0.004)	0.446 (0.017)

Table 3: XGBoost classifier performance in terms of F1-score, ROC AUC and MCC across 10 runs. The best and second-best scores per detector surrogate and performance metric are denoted in bold and italics, respectively.

that the surrogate classifiers considered the most important across all 10 runs, we can see (Fig. 3) that linguistic features are very important for simulating detectors' behavior. In particular, the number of words (denoted as "stats_num_words") is considered an important variable in all 10 runs for surrogate models of each detector, with also quite a high average importance score across the runs.

Performance by task When simulating detectors' behavior on a task level, in terms of F1-macro and MCC score, the surrogate XGB model simulating the behavior of ChatGPT_QA performs the best in all tasks except for review and stance_generation. For review_generation, the surrogate for XLMR_ChatGPT outperforms the others, while for stance_generation simulating Radar_Vicuna7B is the most successful (also in terms of other metrics, see Appendix Tab. 12). Comparing each detector surrogate model performance across different tasks in terms of MCC, the surrogate for ChatGPT_QA performs the best on the information_extraction task, the surrogates for XLMR_ChatGPT and Radar_Vicuna7B perform the best on the stance_generation task and the surrogate for LLMDet performs the best on review_generation task. Looking at the most important features used by surrogates, we observe that linguistic features such as "stats_percent_unique_words" and "stats_num_words" are perceived as most important for different detectors and tasks, although the surrogate for LLMDet tends to rely more on word length and punctuation.[8]

Performance by generative model When simulating detector behavior with respect to the underlying generator model (see Appendix Tab. 13), in terms of F1-macro and MCC scores, we had the most success for ChatGPT_QA detector for all generators except human, Falcon-7b* and GPT-4. In the case when human or Falcon-7b* generated the text, the best performance was obtained for the Radar_Vicuna7B surrogate, while in the case of GPT-4, it was the surrogate for XLMR_ChatGPT.

[8]See Fig. 31 in Appendix A.

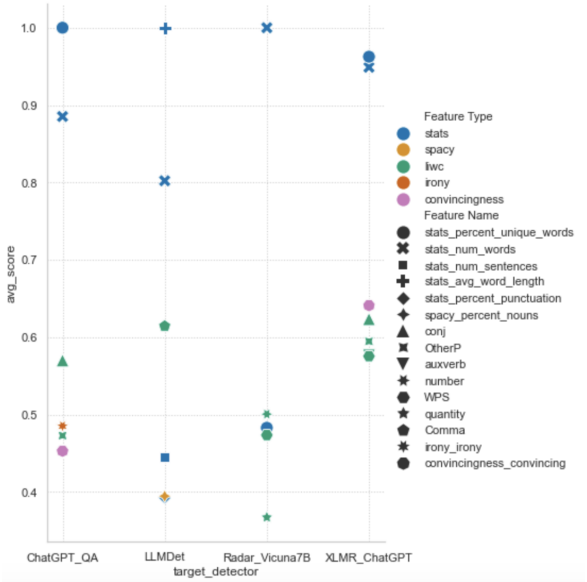

Figure 3: Features resulting as top 10 most important within each of 10 runs per different detector XGBoost surrogate classifiers, together with their average importance score.

Regarding feature importance (see Appendix Fig. 32), we observe a consistent trend where surrogates generally depend on the "number of words" feature. However, the LLMDet surrogate additionally relies on "average word length" and punctuation (comma, apostrophe), and the XLMR_ChatGPT surrogate even considers various emotions.

6 Ablation studies

Analyzing the detector performance across the ablation scenarios (see Fig. 4), we found GPT-Zero as the top performer consistently, with scores of 0.683, 0.692, and 0.652 in the uncontrolled, temperature-min, and length scenarios, respectively. Following closely is ChatGPT_QA, which suffered a noticeable 17% drop in performance, particularly when length control is applied. XLMR_ChatGPT, in the third position, exhibits a similar pattern. Interestingly, the introduction of temperature control does not significantly affect detector performance. However, length control appears to be a critical factor, leading to performance reduction in all detectors except Radar_Vicuna7B, which demonstrates a slight improvement under this constraint.[9] Extending the ablation analysis to the text and the corresponding generators, we made the following observations.

In the Length-based scenario, focusing on text characteristics revealed a notable increase in the overall percentage of nouns, coupled with a decrease in the percentage of verbs (see Appendix

[9]See Appendix Tab. 11 for detailed detector performance on the different ablation scenarios.

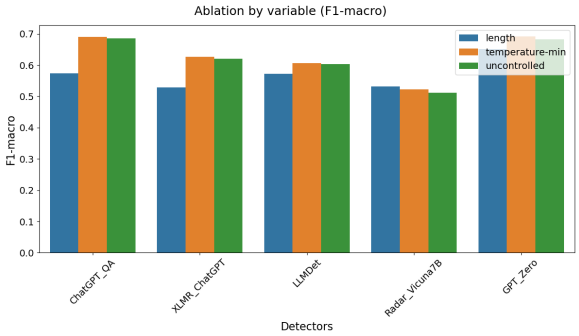

Figure 4: Detectors performance on the ablation set **per controlled variable** (F1-macro)

Fig. 25). Additionally, the emphasis on time orientation was reduced (see Appendix Fig. 27). We did not observe significant changes in writer's attitudes. However, when evaluating the performance of detectors on texts generated by different models, a noticeable change was identified for ChatGPT_QA. It struggled more to correctly detect text generated by GPT-4 (see Appendix Fig. 24).

In the Temperature-min scenario, an analysis of text characteristics showed a drastic increase in the average sentence length for LLaMA-2-7b* (see Appendix Fig. 26). In relation to writer's attitudes, LLaMA-2-7b* was significantly affected in a positive manner. It played a crucial role in tempering overexpressed emotions, such as curiosity, and contributed to enhanced convincingness (see Appendix Fig. 13). Despite this, the text generated by LLaMA-2-7b* was not harder to detect in this scenario (see Appendix Fig. 24).

7 Discussion

Analyzing detectors' performance across tasks (see Tab. 4), focusing on F1-macro, we noted similarities between ChatGPT_QA and XLMR_ChatGPT. However, XLMR_ChatGPT decreases performance, particularly in stance and review_generation tasks, indicating limitations with less probable data scenarios not extensively covered during training. This suggests that XLMR_ChatGPT might be less adept at handling unseen tasks, potentially lacking generalizability.

Furthermore, when evaluating models with pronounced emotions (Falcon-7b-Dolly, Guanaco-7b, and GPT-3.5), we identified a significant advantage of ChatGPT_QA over LLMDet, shown in the detector vs. generator F1-macro comparison. This observation aligns with the correlations observed between ChatGPT_QA predictions and writers' attitudes, contrasting with LLMDet, which exhibits minimal correlation with writers' attitudes (see Appendix Fig. 22). Although Radar_Vicuna7B is the overall least performing model, it demonstrates competitive performance in stance and review_generation tasks. These tasks were generated by repurposing datasets less likely covered during LLMs training, which could suggest that Radar_Vicuna7B is less overfitted. In general, LLMDet demonstrates more consistent performance across the tasks, while ChatGPT_QA ranks as the second-best performer in certain tasks but falls short in others. Specifically, ChatGPT_QA performs well in QA-related tasks as anticipated, yet it underperforms in tasks such as information_extraction.

When comparing detector versus surrogate model performances at the task level, we can notice that for brainstorming, general_qa, open_qa, creative_writing, surrogate model performance mirrors detector behavior (meaning that the good performance of a detector, typically ChatGPT_QA, could also be well mimicked by the respective surrogate model) while this is not the case for the other tasks (stance_generation, information_extraction, closed_qa, classification, review_generation, summarization). Particularly when LLMDet best performs in a task (e.g., stance_generation, see Tab. 2 and Appendix Fig. 30), its behavior is hard to explain with the surrogate model. This means that the considered features are very useful for simulating ChatGPT_QA behavior and less so for LLMDet, probably leading the LLMDet surrogate to rely on unexpected features like punctuation

Notably, variations in the best-performing model across evaluations arise from differences in how the data was grouped. Figure 2 and Table 4 assess performance concerning the generator, using a subset of human-generated and the corresponding MGT, resulting in a balanced dataset. Conversely, Table 2 evaluates detectors on full or task-specific subsets, disregarding the generator, based on human and corresponding MGTs from multiple models, often leading to an imbalanced dataset that may bias F1-macro scores in favor of LLMDet.

8 Conclusion

In this study, we present a benchmark for evaluating LLM-generated text detectors. Our dataset comprises 3180 instructions with corresponding human responses across 10 tasks from 3 diverse sources. Synthetic responses were generated by prompting 7 different LLMs, resulting in a dataset exceeding

	LLaMA-2 70b*			GPT-4			GPT-3.5			LLaMA-2-7b*			Falcon 7b*			Guanaco 7B			Falcon-7b Dolly		
	F1-m	MCC	ROC	F1-m	MCC	ROC	F1-m	MCC	ROC	F1-m	MCC	ROC	F1-m	MCC	ROC	F1-m	MCC	ROC	F1-m	MCC	ROC
CLOSED QA																					
ChatGPT_QA	**0.548**	**0.230**	**0.591**	**0.768**	**0.562**	**0.772**	**0.723**	**0.489**	**0.730**	**0.636**	**0.358**	**0.657**	**0.614**	**0.326**	**0.640**	**0.838**	**0.681**	**0.838**	**0.880**	**0.760**	**0.880**
XLMR_ChatGPT	0.493	0.005	0.502	0.700	0.405	0.701	0.719	0.447	0.721	0.521	0.055	0.527	0.577	0.158	0.578	0.740	0.495	0.743	0.770	0.569	0.775
LLMDet	0.530	0.151	0.564	0.567	0.285	0.610	0.545	0.204	0.583	0.545	0.204	0.583	0.553	0.232	0.593	0.569	0.293	0.613	0.585	0.359	0.632
Radar_Vicuna7B	0.378	0.023	0.505	0.357	-0.121	0.466	0.372	-0.031	0.493	0.357	-0.121	0.466	0.352	-0.142	0.458	0.358	-0.114	0.468	0.380	0.035	0.507
CLASSIFICATION																					
ChatGPT_QA	0.568	0.336	0.621	0.820	**0.676**	0.824	0.588	0.361	0.634	**0.727**	**0.542**	**0.741**	0.472	**0.209**	0.560	**0.904**	**0.815**	**0.904**	**0.855**	**0.733**	**0.857**
XLMR_ChatGPT	0.426	**0.068**	0.520	**0.824**	0.665	**0.826**	**0.635**	**0.378**	**0.661**	0.618	0.354	0.647	0.486	0.165	**0.556**	0.861	0.729	0.862	0.836	0.685	0.837
LLMDet	**0.600**	0.351	**0.638**	0.567	0.233	0.598	0.485	0.000	0.500	0.599	0.344	0.636	**0.515**	0.078	0.536	0.584	0.289	0.618	0.576	0.264	0.609
Radar_Vicuna7B	0.352	0.064	0.507	0.326	-0.184	0.455	0.352	0.064	0.507	0.349	0.000	0.500	0.352	0.064	0.507	0.346	-0.030	0.496	0.343	-0.066	0.489
OPEN QA																					
ChatGPT_QA	**0.859**	**0.721**	**0.859**	**0.863**	**0.731**	**0.864**	**0.818**	**0.636**	**0.818**	**0.841**	**0.683**	**0.841**	**0.656**	**0.343**	**0.664**	**0.884**	**0.776**	**0.884**	**0.841**	**0.683**	**0.841**
XLMR_ChatGPT	0.725	0.477	0.730	0.746	0.531	0.752	0.752	0.548	0.759	0.688	0.389	0.691	0.638	0.278	0.639	0.770	0.595	0.777	0.759	0.566	0.766
LLMDet	0.574	0.361	0.627	0.577	0.379	0.632	0.552	0.265	0.600	0.570	0.344	0.623	0.482	0.026	0.511	0.570	0.344	0.623	0.561	0.303	0.611
Radar_Vicuna7B	0.427	0.052	0.516	0.375	-0.163	0.436	0.454	0.229	0.557	0.432	0.076	0.523	0.449	0.193	0.550	0.433	0.084	0.525	0.442	0.139	0.539
INFORMATION EXTRACTION																					
ChatGPT_QA	0.476	**0.139**	**0.547**	**0.647**	**0.387**	**0.668**	0.600	**0.321**	0.632	**0.548**	**0.245**	**0.594**	**0.504**	**0.182**	**0.565**	**0.759**	**0.555**	**0.765**	**0.803**	**0.625**	**0.805**
XLMR_ChatGPT	0.430	-0.069	0.471	0.596	0.214	0.603	**0.630**	0.275	**0.635**	0.496	0.045	0.520	0.460	-0.015	0.493	0.715	0.431	0.715	0.733	0.466	0.733
LLMDet	**0.481**	0.074	0.529	0.494	0.123	0.547	0.493	0.117	0.545	0.503	0.157	0.558	0.486	0.092	0.536	0.526	0.263	0.590	0.526	0.263	0.590
Radar_Vicuna7B	0.402	0.088	0.520	0.392	0.009	0.502	0.397	0.045	0.511	0.398	0.055	0.513	0.372	-0.101	0.469	0.373	-0.095	0.471	0.384	-0.040	0.489
BRAINSTORMING																					
ChatGPT_QA	**0.839**	**0.679**	**0.839**	**0.815**	**0.634**	**0.815**	**0.802**	**0.610**	**0.803**	**0.882**	**0.764**	**0.882**	**0.693**	**0.427**	**0.702**	**0.884**	**0.769**	**0.884**	**0.832**	**0.667**	**0.833**
XLMR_ChatGPT	0.766	0.532	0.766	0.766	0.532	0.766	0.766	0.532	0.766	0.787	0.576	0.788	0.653	0.321	0.657	0.842	0.696	0.843	0.819	0.644	0.820
LLMDet	0.602	0.391	0.646	0.604	0.399	0.648	0.597	0.368	0.639	0.593	0.353	0.635	0.523	0.112	0.549	0.591	0.346	0.633	0.609	0.423	0.655
Radar_Vicuna7B	0.444	0.126	0.536	0.365	-0.200	0.418	0.451	0.173	0.547	0.441	0.109	0.532	0.450	0.163	0.545	0.441	0.109	0.532	0.429	0.040	0.513
GENERAL QA																					
ChatGPT_QA	**0.830**	**0.688**	**0.833**	**0.819**	**0.661**	**0.821**	**0.825**	**0.677**	**0.828**	**0.810**	**0.640**	**0.812**	**0.730**	**0.460**	**0.730**	**0.843**	**0.721**	**0.846**	**0.817**	**0.656**	**0.819**
XLMR_ChatGPT	0.772	0.599	0.779	0.740	0.514	0.746	0.774	0.604	0.781	0.740	0.514	0.746	0.708	0.436	0.712	0.774	0.604	0.781	0.769	0.593	0.777
LLMDet	0.560	0.359	0.621	0.557	0.341	0.616	0.538	0.251	0.592	0.555	0.332	0.614	0.463	-0.010	0.496	0.558	0.350	0.618	0.550	0.306	0.607
Radar_Vicuna7B	0.507	0.196	0.569	0.459	0.011	0.504	0.535	0.346	0.607	0.492	0.132	0.549	0.527	0.297	0.596	0.525	0.288	0.594	0.532	0.326	0.603
SUMMARIZATION																					
ChatGPT_QA	**0.576**	**0.171**	**0.583**	**0.764**	**0.532**	**0.765**	**0.750**	**0.502**	**0.750**	**0.664**	**0.331**	**0.665**	**0.570**	**0.161**	**0.578**	**0.795**	**0.601**	**0.796**	**0.818**	**0.657**	**0.820**
XLMR_ChatGPT	0.468	-0.059	0.471	0.676	0.378	0.682	0.649	0.313	0.653	0.565	0.131	0.566	0.504	0.010	0.505	0.724	0.501	0.733	0.735	0.533	0.745
LLMDet	0.450	0.063	0.522	0.483	0.246	0.570	0.473	0.183	0.556	0.467	0.145	0.546	0.458	0.102	0.534	0.482	0.235	0.568	0.487	0.270	0.575
Radar_Vicuna7B	0.478	0.134	0.546	0.422	-0.077	0.468	0.455	0.039	0.515	0.452	0.026	0.510	0.439	-0.018	0.493	0.450	0.019	0.507	0.497	0.231	0.573
CREATIVE WRITING																					
ChatGPT_QA	**0.830**	**0.664**	**0.831**	**0.767**	**0.534**	**0.767**	**0.836**	**0.678**	**0.837**	**0.828**	**0.659**	**0.828**	**0.721**	**0.448**	**0.722**	**0.845**	**0.696**	**0.845**	**0.830**	**0.664**	**0.831**
XLMR_ChatGPT	0.790	0.581	0.790	0.682	0.374	0.684	0.773	0.547	0.773	0.750	0.500	0.750	0.680	0.370	0.682	0.834	0.675	0.835	**0.830**	0.666	**0.831**
LLMDet	0.537	0.348	0.608	0.531	0.310	0.600	0.529	0.301	0.597	0.526	0.284	0.593	0.483	0.094	0.536	0.529	0.301	0.597	0.529	0.301	0.597
Radar_Vicuna7B	0.625	0.410	0.661	0.552	0.166	0.574	0.623	0.403	0.659	0.613	0.362	0.646	0.618	0.382	0.653	0.629	0.424	0.665	0.627	0.417	0.663
REVIEW WRITING																					
ChatGPT_QA	**0.814**	**0.672**	**0.818**	**0.629**	**0.424**	**0.665**	**0.895**	**0.802**	**0.896**	**0.830**	**0.697**	**0.834**	**0.691**	**0.502**	**0.713**	**0.982**	**0.965**	**0.982**	**0.971**	**0.942**	**0.971**
XLMR_ChatGPT	0.627	0.429	0.665	0.372	0.074	0.512	0.816	0.679	0.821	0.600	0.396	0.646	0.586	0.379	0.636	0.922	0.851	0.922	0.950	0.902	0.950
LLMDet	0.683	0.505	0.709	0.686	0.515	0.712	0.637	0.351	0.656	0.671	0.462	0.695	0.617	0.292	0.633	0.682	0.501	0.707	0.666	0.442	0.688
Radar_Vicuna7B	0.488	0.220	0.568	0.498	0.286	0.582	0.501	0.308	0.587	0.500	0.302	0.585	0.501	0.308	0.587	0.501	0.308	0.587	0.501	0.305	0.586
STANCE WRITING																					
ChatGPT_QA	**0.694**	**0.521**	**0.717**	**0.580**	**0.385**	**0.634**	**0.561**	0.363	0.621	**0.766**	**0.614**	**0.777**	0.474	**0.260**	**0.569**	**0.637**	**0.452**	**0.674**	**0.845**	**0.727**	**0.848**
XLMR_ChatGPT	0.570	0.368	0.627	0.525	0.314	0.598	0.643	**0.454**	**0.678**	0.602	0.406	0.649	0.460	0.233	0.560	0.610	0.415	0.654	0.807	0.670	0.813
LLMDet	0.476	0.154	0.551	0.475	0.148	0.549	0.494	0.256	0.576	0.469	0.117	0.540	**0.475**	0.148	0.549	0.502	0.309	0.587	0.501	0.299	0.585
Radar_Vicuna7B	0.349	0.085	0.507	0.349	0.085	0.507	0.349	0.085	0.507	0.348	0.057	0.505	0.349	0.085	0.507	0.300	-0.295	0.409	0.343	-0.039	0.495

Table 4: Detectors performance per task and per generator model in terms of F1-macro (F1-m), ROC AUC (ROC), and Matthews Correlation Coefficient (MCC). The best scores per task and performance metric are displayed in bold-face

25k texts. Evaluating 5 detectors revealed substantial performance variance across tasks. Our meta-analysis explored dataset characteristics, guiding the analysis of detector performance. Surrogate models highlighted the difficulty in explaining the most performant detector, frequently relying on unexpected textual features. Emotions, in particular, aided some detectors in specific tasks, but the most stable and effective detector exhibited consistent performance irrespective of writer's attitudes.

These distinctive characteristics position our benchmark as a valuable resource for end-users seeking to select a detector that best suits their target text style and use case. Notably, we have integrated all analyses into an automated pipeline, which enhances accessibility and usability for researchers and practitioners in the field.

Limitations

Limited Amount of Datasets. A notable limitation of our study is that we rely on a limited number of datasets. Although these datasets were selected for relevance and quality, they represent only a fraction of the available textual sources. This limitation could lead to a lack of diversity in languages, language styles, topics and linguistic nuances, which could affect the generalizability of our results. Users should be aware that the results and findings from our dataset may not fully reflect the complexity and variations that occur in practice. Consequently, interpretations and applications of our research should be made with these limitations in mind.

Limited Selection of Generator Models. Our research includes a selected set of text generation models, which is a notable limitation. This limited

selection means that the dataset may not represent the full range of text generation capabilities currently available. Advanced or emerging models may not be adequately represented, particularly those using state-of-the-art techniques or architectures. As a result, the effectiveness of our dataset in recognizing text from these unrepresented models may be reduced.

Limited Selection of Detector Models. In the present study, we use a specific selection of detector models, which is a significant limitation. This limited selection may not cover the full range of detection methods and technologies available in practice. Consequently, this could lead to a biased assessment of the reliability of our dataset, as it is evaluated against a potentially not fully representative sample of detectors. This limitation highlights the need to be cautious when generalizing the results and suggests that it is important to test the dataset with a wider variety of detector models for a more comprehensive evaluation.

Ethics Statement

Intended Use Our research, which focuses on the collection of a new dataset for machine-generated text detection, is designed to advance the field responsibly and ethically. This dataset, drawn from publicly available resources, integrates various publicly available text generation models and detectors. Our primary goal is to provide the research community with a robust benchmark for evaluating and improving text recognition algorithms. We anticipate that this dataset will help to improve the accuracy and reliability of recognition methods and thereby gain a broader understanding of the nuances in MGT. We emphasize that our dataset is intended to be used for academic and research purposes only. It is intended to facilitate the development of more sophisticated recognition tools that can distinguish between human and MGT to ensure the integrity and authenticity of digital content. In this way, we aim to support ongoing efforts to maintain transparency and trustworthiness in digital communications. We strongly discourage any use of this dataset that violates ethical standards or promotes malicious activity, such as the creation of misleading or harmful content. Our adherence to ethical guidelines reflects our commitment to promoting positive impact in the field of natural language processing and beyond.

Biases. In our study, we acknowledge the potential for inherent biases within our dataset and the broader implications that these biases may have in the field of MGT detection. We recognize that any dataset, especially when derived from publicly available sources, may reflect historical and societal biases in the data it contains. These biases include, but are not limited to, linguistic, cultural, gender or demographic preferences. We are particularly cautious about how these biases might affect the performance and interpretation of detection algorithms and potentially lead to biased or unfair results.

Misuse Potential. Our data set, which was developed for the evaluation of MGT detection, has significant potential for positive impact. However, we are aware of the possibility of its misuse. In particular, there is a risk that the dataset could be used to develop methods to create more convincing MGT, which could be used for misleading purposes such as misinformation or spam. To prevent this, we prohibit the use of our dataset for unethical practices, including but not limited to the creation of misleading or harmful content. We encourage users to adhere to a high standard of ethical responsibility and ensure that the data set is only used to enhance the authenticity and trustworthiness of digital communications.

Reliability conditions. The reliability of our dataset for the evaluation of MGT detection models depends on the performance and development of the generator systems. It is important to recognize that the accuracy and effectiveness of these tools can vary significantly depending on the design, training data, and underlying algorithms. Consequently, the usefulness of our dataset as a benchmark is reliable on the condition that these models are regularly updated and refined to adapt to advances in text generation technologies. Users should note that the results obtained may not be universally applicable to all generator and detector systems, and that continuous validation against a variety of models is essential to maintain the relevance and effectiveness of the dataset in an evolving technological landscape.

Acknowledgements

The work described in this paper has been supported by the project "*Detection of Artificially Generated and Misleading Text in Social Media* (DAGMTSM)" (CYD-C-2021016, armasuisse S&T).

References

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Merouane Debbah, Etienne Goffinet, Daniel Heslow, Julien Launay, Quentin Malartic, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. Falcon-40B: an open large language model with state-of-the-art performance.

Wissam Antoun, Virginie Mouilleron, Benoît Sagot, and Djamé Seddah. 2023. Towards a robust detection of language model generated text: Is chatgpt that easy to detect? *arXiv preprint arXiv:2306.05871*.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online. Association for Computational Linguistics.

Daria Beresneva. 2016. Computer-generated text detection using machine learning: A systematic review. In *Natural Language Processing and Information Systems: 21st International Conference on Applications of Natural Language to Information Systems, NLDB 2016, Salford, UK, June 22-24, 2016, Proceedings 21*, pages 421–426. Springer.

Ryan L Boyd, Ashwini Ashokkumar, Sarah Seraj, and James W Pennebaker. 2022. The development and psychometric properties of liwc-22. *Austin, TX: University of Texas at Austin*, pages 1–47.

Emily Chen, Kristina Lerman, Emilio Ferrara, et al. 2020. Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus twitter data set. *JMIR public health and surveillance*, 6(2):e19273.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

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. *arXiv preprint arXiv:2107.00061*.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. 2023. Free dolly: Introducing the world's first truly open instruction-tuned llm.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A Dataset of Fine-Grained Emotions. pages 4040–4054.

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

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

Shai Gretz, Roni Friedman, Edo Cohen-Karlik, Assaf Toledo, Dan Lahav, Ranit Aharonov, and Noam Slonim. 2019. A Large-scale Dataset for Argument Quality Ranking: Construction and Analysis.

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

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

Tamanna Hossain, Robert L. Logan IV, Arjuna Ugarte, Yoshitomo Matsubara, Sean Young, and Sameer Singh. 2020. COVIDLies: Detecting COVID-19 Misinformation on Social Media.

Xiaomeng Hu, Pin-Yu Chen, and Tsung-Yi Ho. 2023. Radar: Robust ai-text detection via adversarial learning. *arXiv preprint arXiv:2307.03838*.

Maurice Jakesch, Jeffrey T Hancock, and Mor Naaman. 2023. Human heuristics for ai-generated language are flawed. *Proceedings of the National Academy of Sciences*, 120(11):e2208839120.

Jong Wook Kim. Openai/gpt-2-output-dataset: Dataset of gpt-2 outputs for research in detection, biases, and more.

Ryuto Koike, Masahiro Kaneko, and Naoaki Okazaki. 2023. Outfox: Llm-generated essay detection through in-context learning with adversarially generated examples. *arXiv preprint arXiv:2307.11729*.

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

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

Dominik Macko, Robert Moro, Adaku Uchendu, Jason Samuel Lucas, Michiharu Yamashita, Matúš Pikuliak, Ivan Srba, Thai Le, Dongwon Lee, Jakub Simko, et al. 2023. Multitude: Large-scale multilingual machine-generated text detection benchmark. *arXiv preprint arXiv:2310.13606*.

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

Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying Recommendations using Distantly-Labeled Reviews and Fine-Grained Aspects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 188–197, Hong Kong, China. Association for Computational Linguistics.

R OpenAI. 2023. Gpt-4 technical report. *arXiv*, pages 2303–08774.

Yikang Pan, Liangming Pan, Wenhu Chen, Preslav Nakov, Min-Yen Kan, and William Yang Wang. 2023. On the risk of misinformation pollution with large language models. *arXiv preprint arXiv:2305.13661*.

Alessandro Pegoraro, Kavita Kumari, Hossein Fereidooni, and Ahmad-Reza Sadeghi. 2023. To chatgpt, or not to chatgpt: That is the question! *arXiv preprint arXiv:2304.01487*.

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

Jinyan Su, Terry Yue Zhuo, Di Wang, and Preslav Nakov. 2023. Detectllm: Leveraging log rank information for zero-shot detection of machine-generated text. *arXiv preprint arXiv:2306.05540*.

Ruixiang Tang, Yu-Neng Chuang, and Xia Hu. 2023. The science of detecting llm-generated texts. *arXiv preprint arXiv:2303.07205*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

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. *arXiv preprint arXiv:2109.13296*.

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

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Chenxi Whitehouse, Osama Mohammed Afzal, Tarek Mahmoud, Alham Fikri Aji, et al. 2023. M4: Multi-generator, multi-domain, and multi-lingual black-box machine-generated text detection. *arXiv preprint arXiv:2305.14902*.

Kangxi Wu, Liang Pang, Huawei Shen, Xueqi Cheng, and Tat-Seng Chua. 2023. Llmdet: A large language models detection tool. *arXiv preprint arXiv:2305.15004*.

Peipeng Yu, Jiahan Chen, Xuan Feng, and Zhihua Xia. 2023. Cheat: A large-scale dataset for detecting chatgpt-written abstracts. *arXiv preprint arXiv:2304.12008*.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. *Advances in neural information processing systems*, 32.

A Appendix

A.1 Related Work

Dataset	LLMs	Domain/Task
GPT-2 Output[10] (Radford et al., 2019)	GPT-2	Various
HC3 (Guo et al., 2023)	ChatGPT	QA
Neural Fake News (Zellers et al., 2019)	Grover	News
TweepFake (Fagni et al., 2021)	GPT2	Tweets
GPT2-Output (Kim)	GPT2	WebText
TURINGBENCH (Uchendu et al., 2021)	GPT1,2,3	News
ChatGPTorNot (Pegoraro et al., 2023)	ChatGPT	QA (medical and finance)
CHEAT (Yu et al., 2023)	ChatGPT	Abstracts
MULTITuDE (Macko et al., 2023)	Alpaca-Lora, GPT3.5/4, LLama, OPT, Davinci, Vicuna	News
ArguGPT (Liu et al., 2023)	GPT2/3.5, Babbage, Curie, Davinci	Essay writing
M4 (Wang et al., 2023)	Davinci003, ChatGPT, Cohere, Dolly-v2, BLOOMz, FlanT5, LLaMA	(News) Article writing, QA, Review writing
MGTBench (He et al., 2023)	ChatGPT, ChatGLM, Dolly, ChatGPT-turbo, GPT4Al,l StableLM	QA

Table 5: Overview of related benchmark datasets (QA is the acronym for Question Answering).

A.2 Dataset Description

Features	LLaMA-2-70b*	Guanaco-7b	LLaMA-2-7b*	Falcon-7b*	Falcon-7b-Dolly	Human	GPT-3.5	GPT-4
Avg. word length	5.64	5.07	5.44	5.41	7.19	5.44	5.48	5.45
Avg. sentence length	92.00	122.68	137.19	82.92	172.32	87.25	91.76	103.50
Avg. words per sentence	14.96	23.38	22.38	14.00	25.48	14.22	15.08	17.16
Percent of vowels	28.40	28.95	28.83	29.13	29.35	29.17	29.62	29.34
Percent of consonants	43.69	44.13	44.35	44.36	45.12	44.12	45.19	45.22
Percent of punctuation	4.54	3.61	4.29	4.30	3.47	3.51	3.56	4.01
Percent of stopwords	30.83	40.54	33.73	30.91	37.32	29.78	33.34	34.25
Num. words	167.16	326.42	265.24	91.79	198.09	71.50	96.27	226.63
Num. sentences	12.33	24.40	17.55	6.52	15.03	5.02	7.35	15.87
Percent of unique words	69.95	56.68	64.61	80.56	60.77	82.23	76.29	68.63
Percent of long words	24.96	18.97	23.45	23.01	22.49	21.52	24.42	24.51

Table 6: Overview of different linguistic features and their average values across different generators. Notation used for the above features: 'stats_avg_word_length', 'stats_avg_sentence_length', 'stats_avg_words_per_sentence', 'stats_percent_vowels', 'stats_percent_consonants', 'stats_percent_punctuation', 'stats_percent_stopwords', 'stats_num_words', 'stats_num_sentences', 'stats_percent_unique_words', 'stats_percent_long_words'

Readability score	LLaMA-2-70b*	Guanaco-7b	LLaMA-2-7b*	Falcon-7b*	Falcon-7b-Dolly	Human	GPT-3.5	GPT-4
flesch_kincaid_score	10.63	10.36	10.19	11.66	14.45	12.01	10.63	9.47
flesch_score	50.33	54.83	51.61	48.27	39.77	49.14	45.58	52.74
gunning_fog_score	12.82	12.66	12.33	14.00	16.60	14.39	12.67	11.17
coleman_liau_score	10.96	9.51	10.94	10.75	11.14	10.76	12.22	10.64
dale_chall_score	9.87	9.51	9.90	10.02	10.67	9.66	10.31	10.10
ari_score	10.77	10.36	10.20	12.01	15.76	12.83	10.58	9.20
linsear_write_score	12.00	12.61	11.35	14.26	18.69	15.31	10.45	9.56

Table 7: Avg. Readability

Attitude	Model	Description
Emotions[11] (Demszky et al., 2020)	SamLowe/roberta-base-go_emotions	Trained on the Go-Emotions dataset to classify over 28 emotions
Persuasiveness[12]	paragon-analytics/roberta_persuade	Binary classification (Non-persuasive, Persuasive)
Irony[13] (Barbieri et al., 2020)	cardiffnlp/twitter-roberta-base-irony	Roberta base model pretrained on tweets and finetuned for irony detection with the TweetEval benchmark
Convincingness[14] (Gretz et al., 2019)	jakub014/bert-base-uncased-IBM-argQ-30k-finetuned-convincingness-IBM	Trained on the assumption that convincingness is linked to high quality arguments which are typically clear, relevant, and with high impact.

Table 8: Off-the-shelf models used to infer attitudes in texts.

A.3 Repurposed review and stance datasets

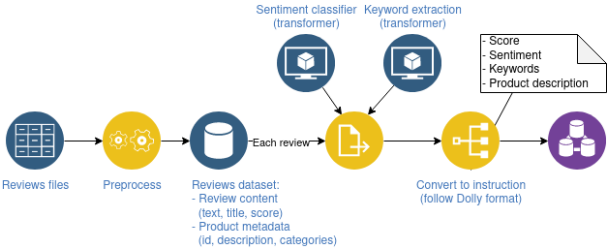

Figure 5: Process to repurpose review datasets into instructions.

Dataset	Instruction	Context	Response	Generator
Amazon reviews	Based in the provided product description, generate a **positive** product review from a client who is **Very satisfied**. Consider the following keywords: **DIY, DIY, house, fashion magazines**	Product description: Tackle any home improvement project or repair with help from The Family Handyman. Guarantee success with step-by-step instructions and how-to photos, plus save time and money with handy hints, workshop tips and new product reviews. No matter how little experience you have, The Family Handyman makes it easy! Family Handyman publishes 11 issues per year which includes a special issue that counts as 2 issues in your subscription.	I'm a so-so DIY'er. I try to DIY pretty much everything around the house, not everything turns out too well. So that's where I'm coming from, and my opinion of this magazine is that it's a lot of fun to read. Interesting, helpful, fun ideas and articles. Some good tips too. Nothing earth shattering but not bad. I feel like this is what my wife must feel like when she reads her fashion magazines.	Human
			I'm beyond pleased with The Family Handyman! As someone who loves DIY projects around the house, this magazine is a game-changer. The step-by-step instructions and how-to photos are incredibly helpful, making every home improvement project a breeze. The handy hints and workshop tips have saved me both time and money. It's like having a personal expert by my side. I especially appreciate the new product reviews, guiding me towards the best tools for the job. It's become my go-to resource for all things DIY. Couldn't be happier with my subscription!	GPT-3.5
	Based in the provided product description, generate a **positive** product review from a client who is **Unsatisfied**. Consider the following keywords: **bike magazine, bike magazine, bike magazine, bi**	Product description: Cycle World, America's largest motorcycle media brand, is the leading resource for in-depth, quality information on all types of motorcycles, accessories and riding gear.	I used to love this magazine, particularly in the 1970s. Unfortunately, the writing is now bland and there is little incisive coverage. This used to be the premier bike magazine. I won't renew, that's for sure.	Human
			Initially dissatisfied, but pleasantly surprised! The Cycle World magazine exceeded my expectations with its comprehensive coverage of motorcycles, accessories, and riding gear. The quality information provided is unmatched, making it a valuable resource for any bike enthusiast. Highly recommended for a deeper understanding of the biking world!	GPT-3.5
CovidLies	Generate a tweet related to COVID-19 that have a stance in favor of **'We're very close to a vaccine.'** Include the following mentions: **[@SpeakerPelosi]**		@SpeakerPelosi I understand Israel will have a vaccine for coronavirus in a few weeks.	Human
			Encouraging news on the COVID-19 front! We're very close to a vaccine breakthrough that could bring us closer to ending this pandemic. Fingers crossed for a swift and successful rollout. [@SpeakerPelosi] #COVID19 #VaccineHope	GPT-3.5

Table 9: Repurposed Datasets examples

A.4 Text generation

Generators In our study, we evaluated a diverse set of models from the Falcon and LLaMA families, as well as two GPT models. Specifically, we considered two models from the Falcon family (Almazrouei et al., 2023): Falcon 7B Instruct[15] and our customized version, Falcon 7B Dolly[16], which utilizes LoRA as a Parameter-Efficient Fine-Tuning (PEFT) technique. Additionally, we incorporated three models from the LLaMA family (Touvron et al., 2023b,a): LLaMA-2 7B Chat[17], LLaMA-2 70B Chat[18], and PEFT model LLaMA Guanaco [19]. Finally, our analysis also included two models from the GPT series (OpenAI, 2023): GPT-3.5-turbo-1106 and GPT-4-turbo-1106-preview[20]. The base models that were instruction- or chat-tuned are marked with an asterisk (*) for clarity. The specific configuration can be found in Tab. 10. This selection of models, encompassing a range of complexities and specializations, allows for a comprehensive examination of capabilities of the text generation models.

[15] https://huggingface.co/tiiuae/falcon-7b-instruct
[16] https://huggingface.co/jco2/falcon-7b-dolly-peft-stp-10k
[17] https://huggingface.co/meta-llama/Llama-2-7b-chat-hf
[18] https://huggingface.co/meta-llama/Llama-2-70b-chat-hf
[19] https://huggingface.co/timdettmers/guanaco-7b
[20] https://platform.openai.com/docs/models

Generator	Model name	Max Length	Temperature	PEFT
Falcon-7b*	Falcon 7B instruct	2048	1	No
Falcon-7b-Dolly	Falcon 7B dolly	2048	1	LoRA
Guanaco-7B	LLaMA Guanaco	2048	0.7	QLoRA
LLaMA-2-7b*	LLaMA-2 7B chat	4096	0.6	No
LLaMA-2-70b*	LLaMA-2 70B chat	4096	0.6	No
GPT-3.5	GPT-3.5-turbo-1106	16285(context window)/4096 (max. output)	1	No
GPT-4	GPT-4-turbo-1106-preview	128K(context window)/4096 (max. output)	1	No

Table 10: Models used for the text generation

Prompt We have tried to reduce the prompt template to a minimum to ensure it is uniformly applicable to the different models (see Fig. 6). Only the Guanaco-7b and Falcon-7b-Dolly models have minor modifications in the template due to the use of a different instruction tuning scheme while fine-tuning, as shown in Fig. 7. An additional line with context was only added if such context was available; otherwise, this line was omitted completely.

Context: [GIVEN_CONTEXT]

Instruction: [GIVEN_INSTRUCTION]

Figure 6: Prompt template for the generator models.

Context: [GIVEN_CONTEXT]

Human: [GIVEN_INSTRUCTION] **### Assistant**:

Figure 7: Modified prompt template for the Guanaco-7b and Falcon-7b-Dolly generator models.

A.5 Meta analysis

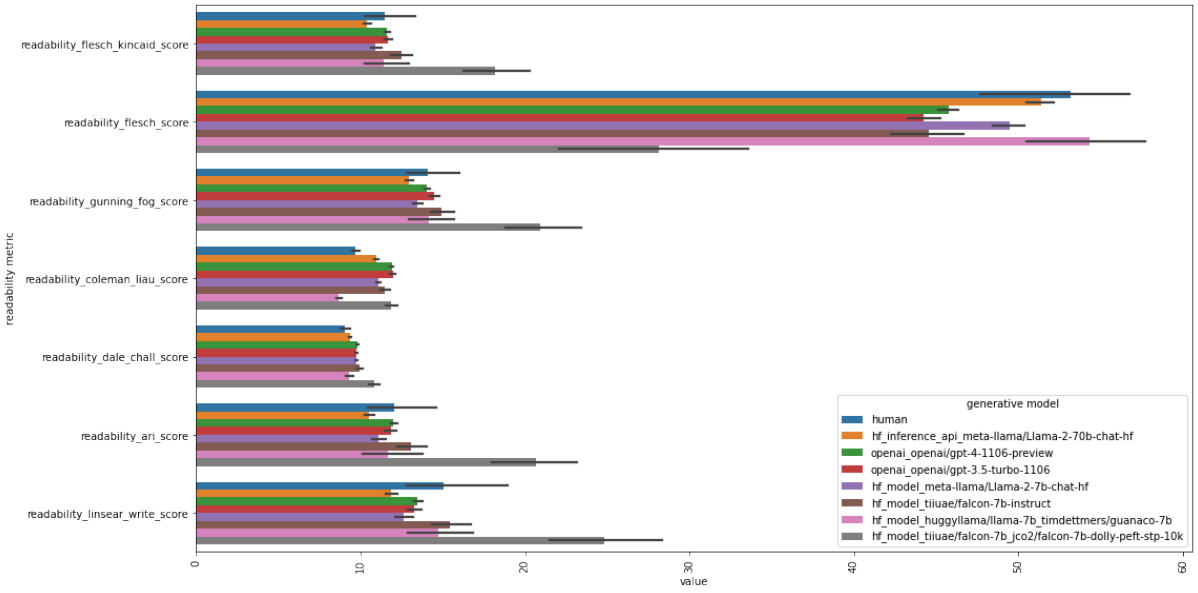

Figure 8: Readability scores across different generative models

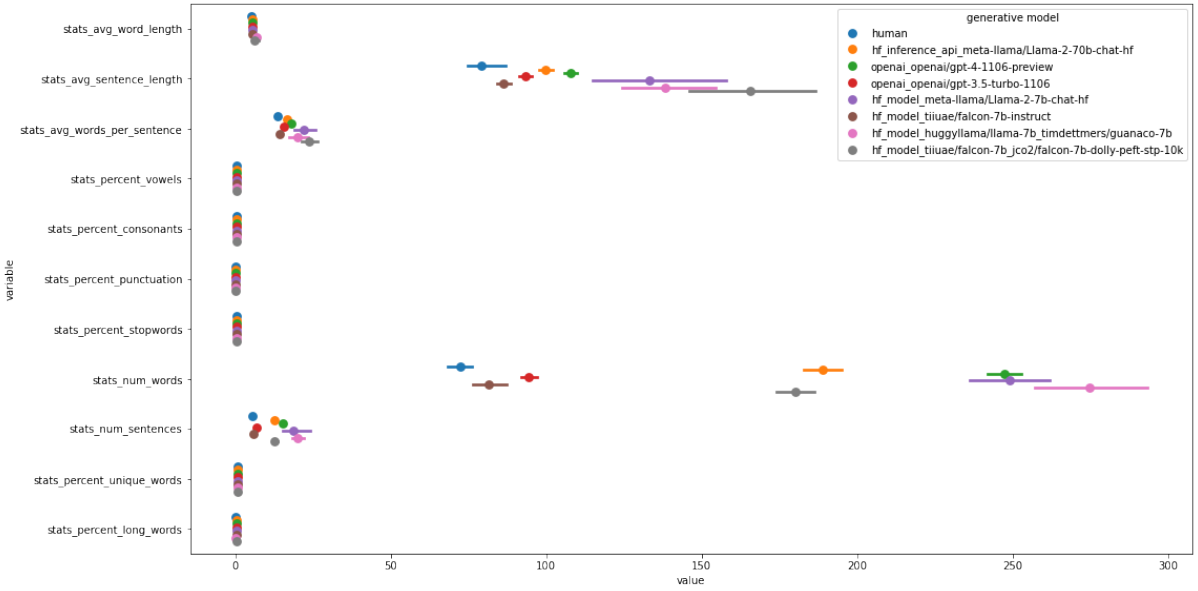

Figure 9: Linguistic features across different generative models

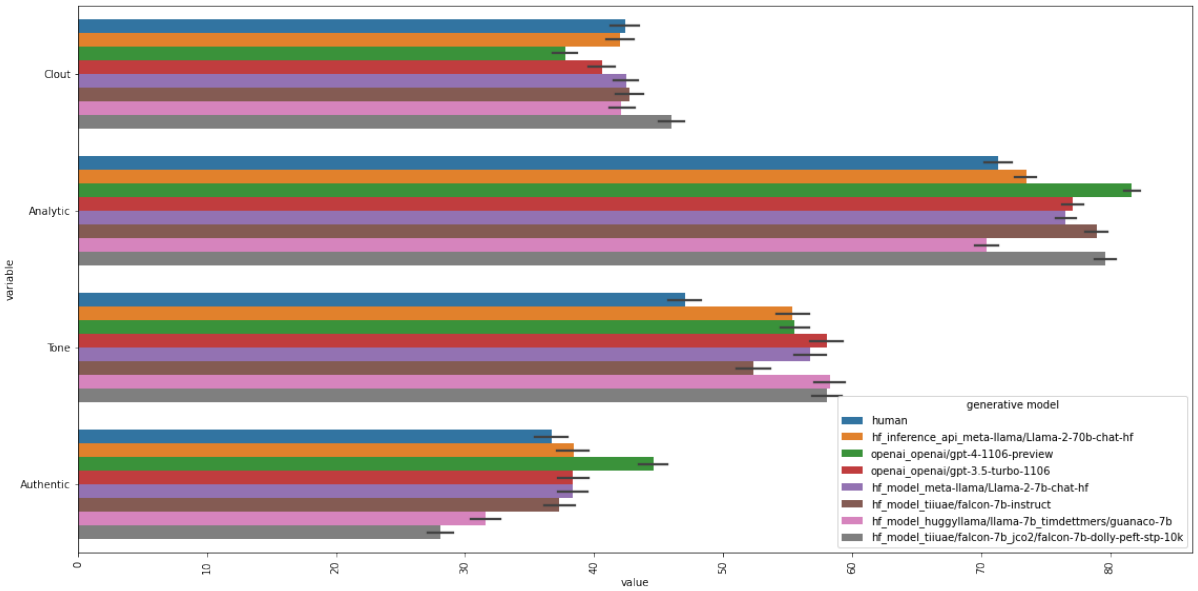

Figure 10: Scores of LIWC features Clout, Authentic, Tone and Analytic across different generative models

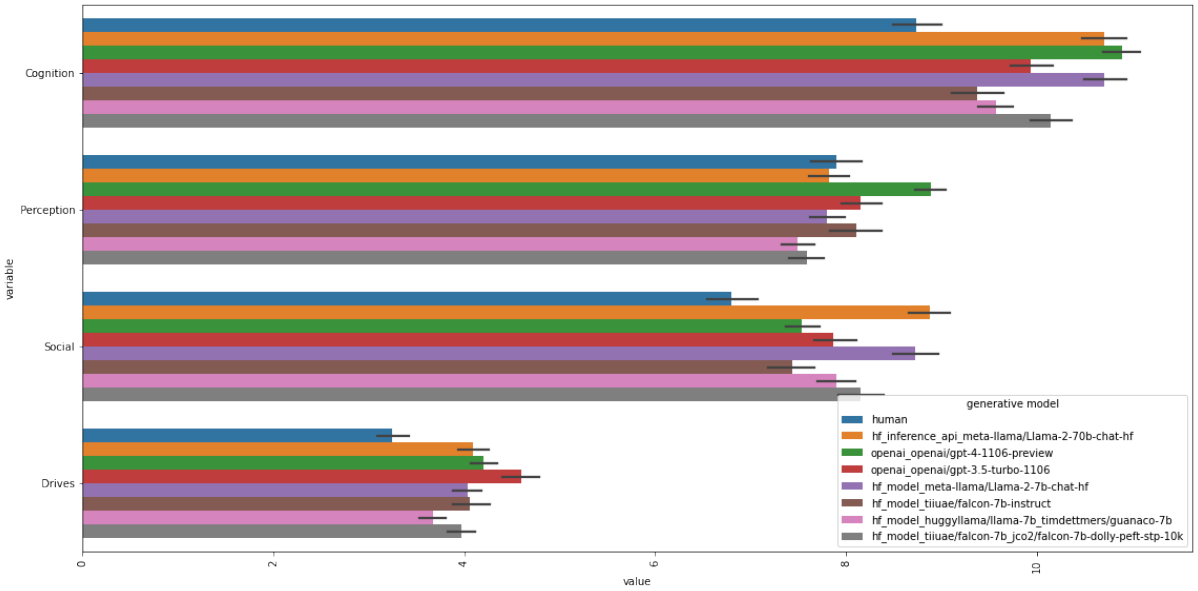

Figure 11: Scores of LIWC features Cognition, Perception, Social and Drives across different generative models

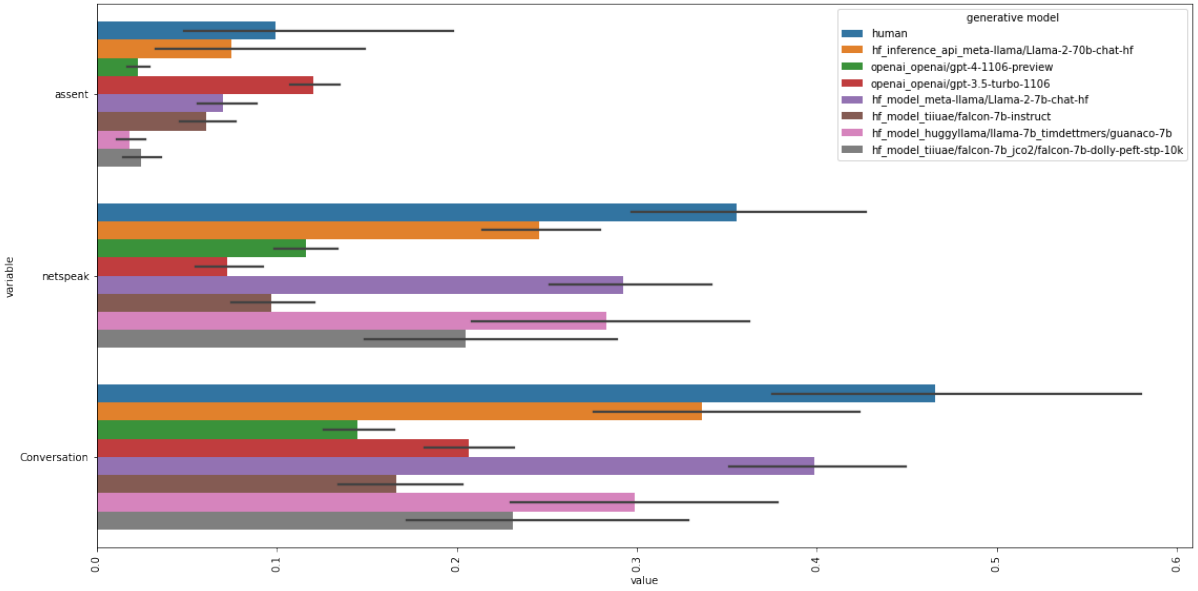

Figure 12: Scores of LIWC features Assent, Netspeak and Conversation across different generative models

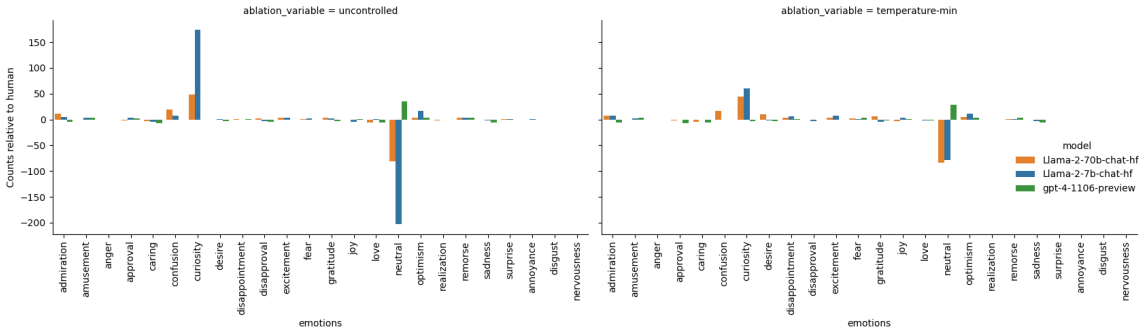

Figure 13: Writer's attitude Emotions counts relative to human, ablation controlling **temp-min**

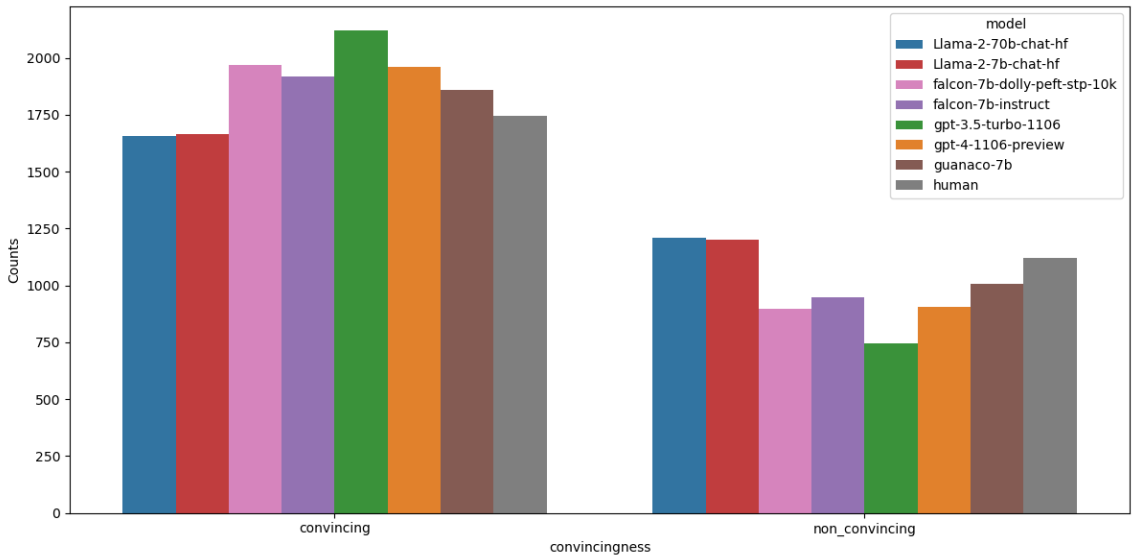

Figure 14: Writer's attitude Convincingness counts

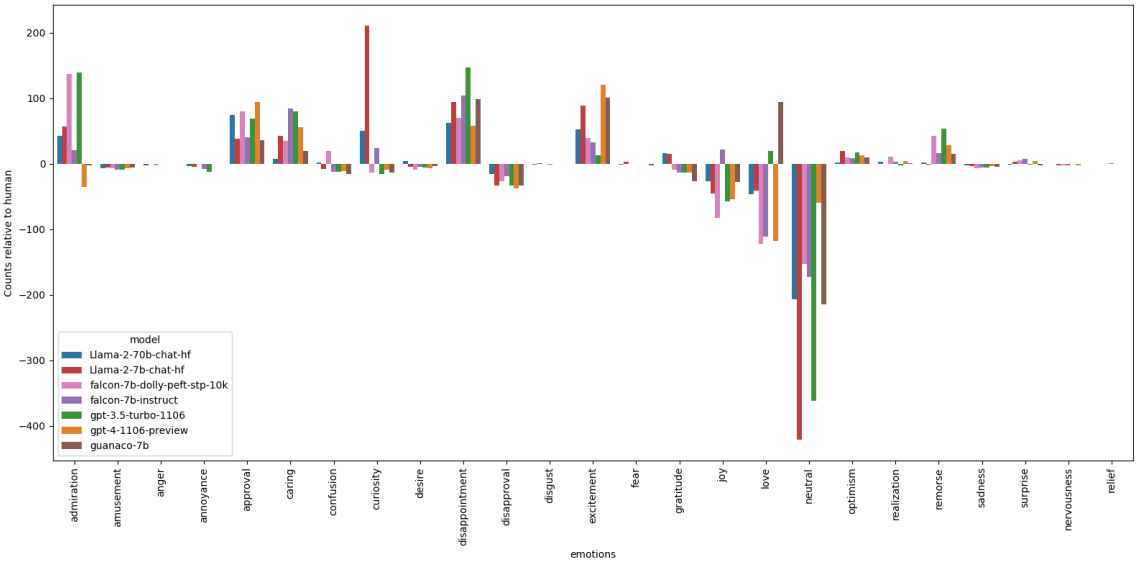

Figure 15: Writer's attitude Emotions counts relative to human

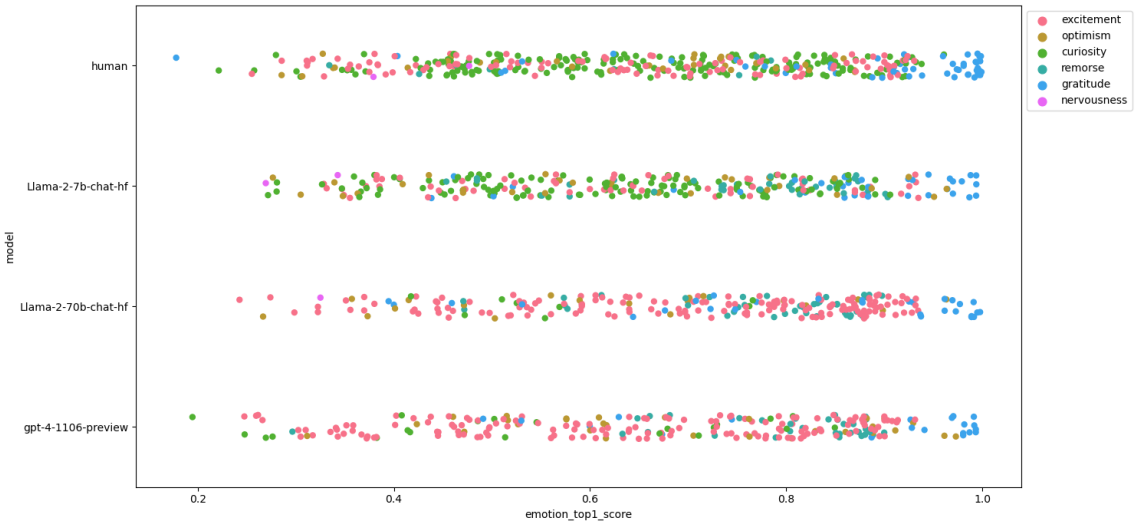

Figure 16: Writer's attitude contrasting expressed Emotions

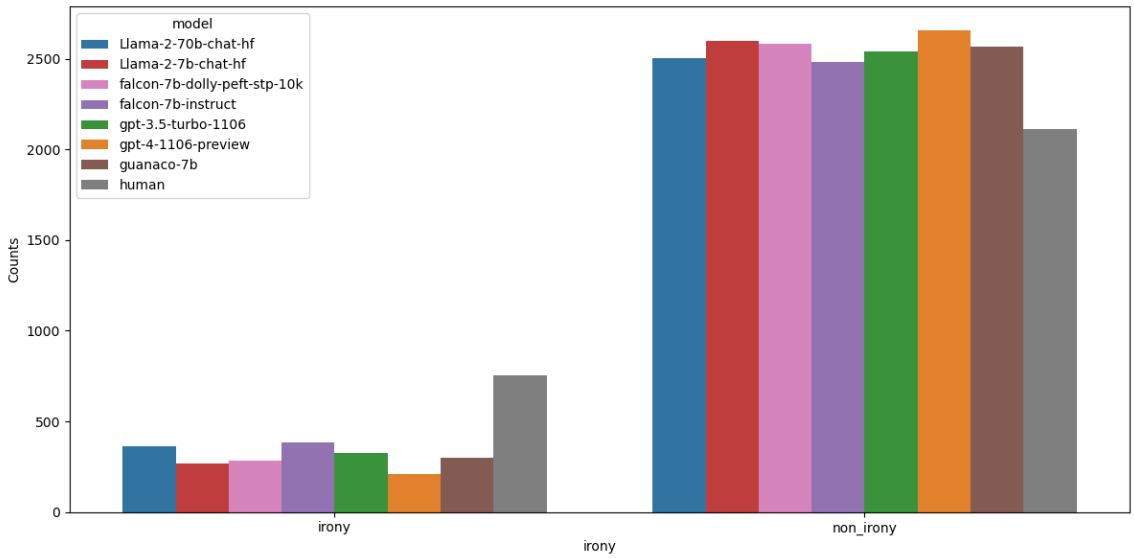

Figure 17: Writer's attitude Irony counts

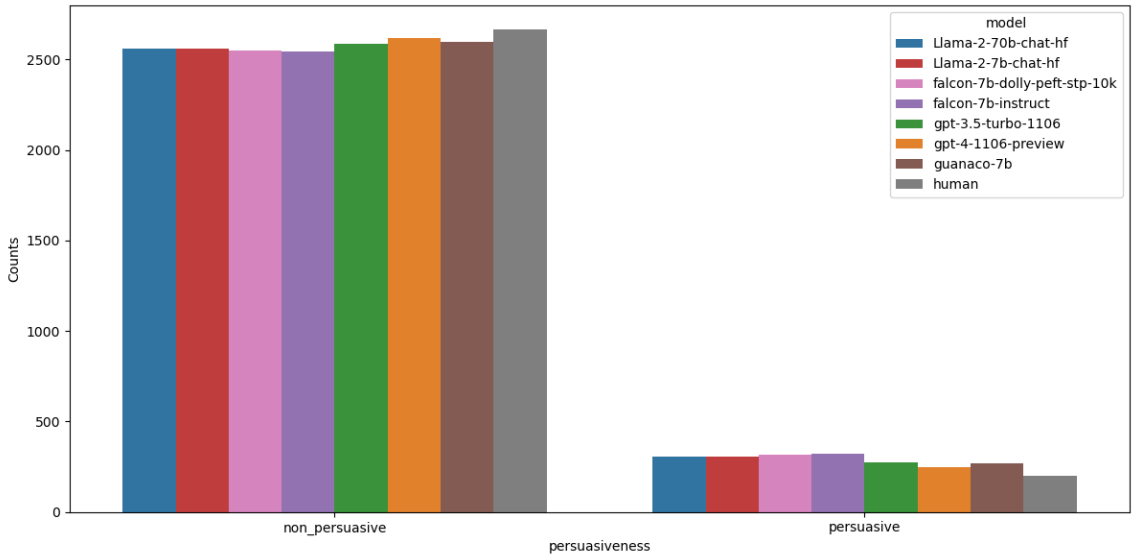

Figure 18: Writer's attitude Persuasiveness counts

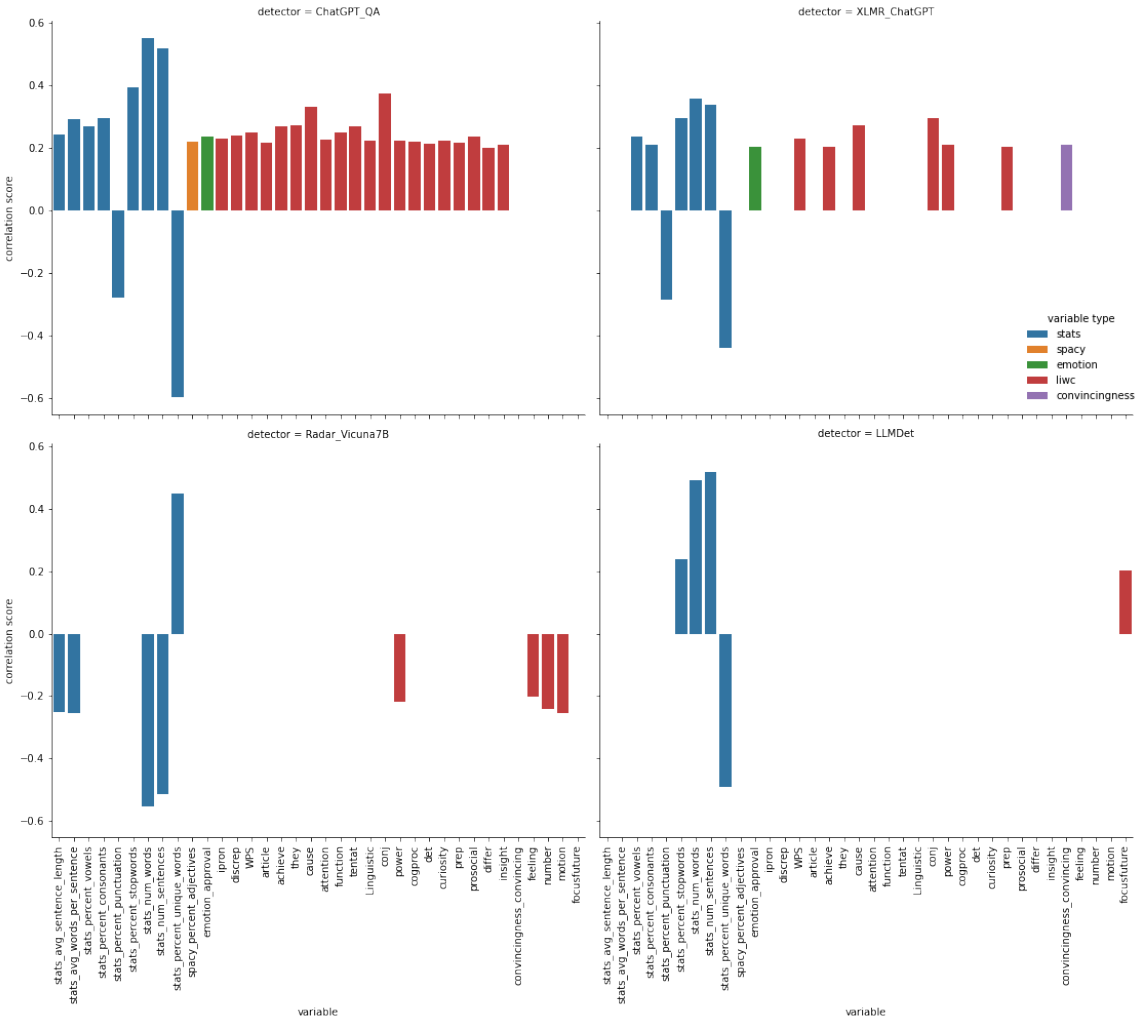

Figure 19: Correlation scores between different features and detectors scores higher than 0.2 (absolute values)

A.6 Detectors evaluation

This section provides additional figures of detector performance, including assessments across tasks based on F1-macro scores (Fig. 20) and evaluations of detection capability across various text generators using the Matthews Correlation Coefficient (MCC) metric (Fig. 21).

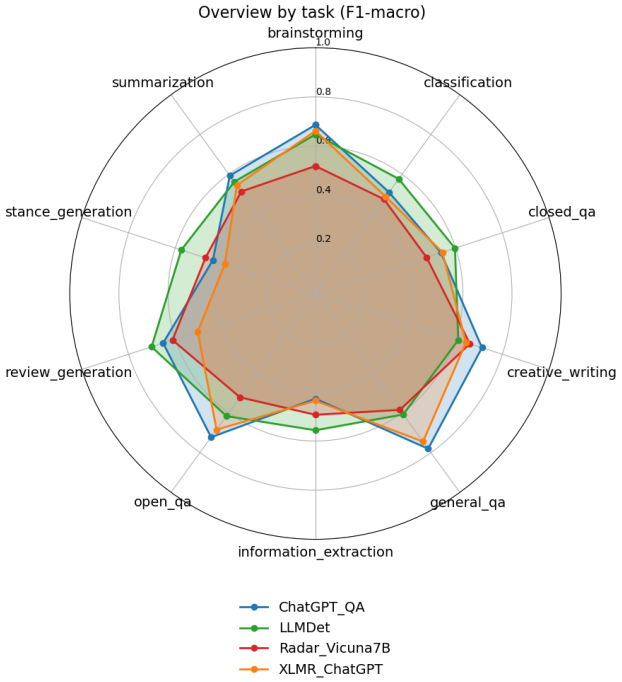

Figure 20: Detectors performance across tasks (F1-macro)

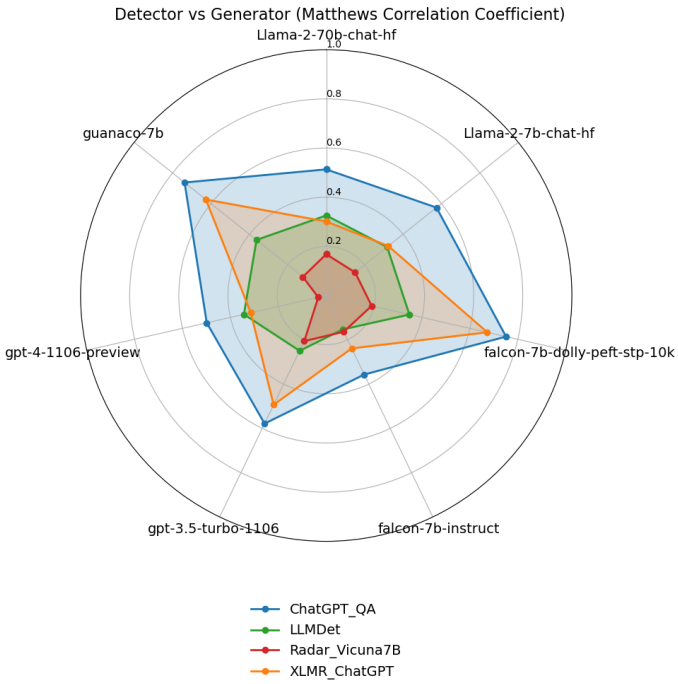

Figure 21: Detectors performance on detecting text produced by different generators (MCC)

A.7 Ablation

Figure 22: Categorical correlations between Detectors and Writers attitudes on the ablation set **per task** (Chi-test)

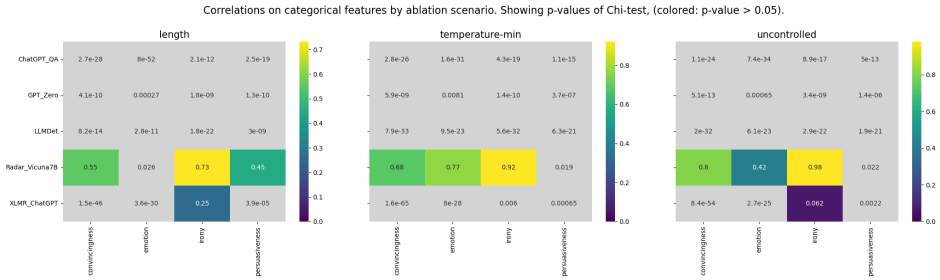

Figure 23: Categorical correlations between Detectors and Writers attitudes on the ablation set **per controlled variable** (Chi-test)

Variable	Detector	f1-weighted	f1-macro	roc_auc_ovr	mcc
uncontrolled	ChatGPT_QA	0.733	0.686	0.750	0.436
	GPT_Zero	0.770	0.683	0.671	0.372
	LLMDet	0.731	0.603	0.594	0.265
	Radar_Vicuna7B	0.666	0.512	0.521	0.057
	XLMR_ChatGPT	0.679	0.620	0.669	0.295
temperature-min	ChatGPT_QA	0.738	0.690	0.754	0.442
	GPT_Zero	0.779	0.692	0.678	0.395
	LLMDet	0.734	0.606	0.597	0.277
	Radar_Vicuna7B	0.676	0.522	0.531	0.089
	XLMR_ChatGPT	0.685	0.626	0.674	0.304
length	ChatGPT_QA	0.610	0.573	0.665	0.288
	GPT_Zero	0.742	0.652	0.648	0.304
	LLMDet	0.702	0.572	0.568	0.169
	Radar_Vicuna7B	0.687	0.533	0.542	0.128
	XLMR_ChatGPT	0.576	0.529	0.597	0.168

Table 11: Detectors performance on the ablation set **per controlled variable** in terms of F1-macro, F1-macro, ROC AUC and MCC.

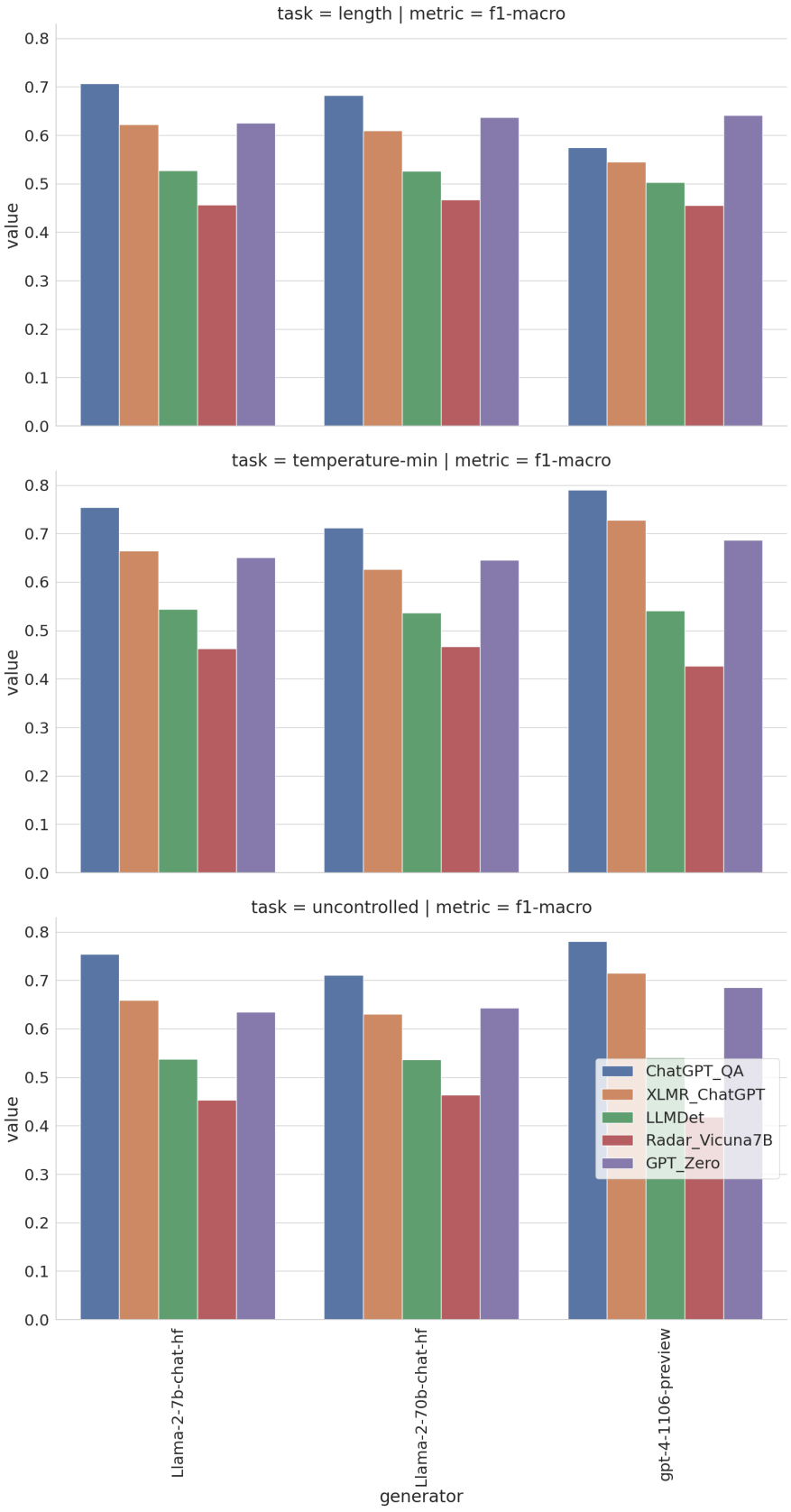

Figure 24: Detector performance by generator on the different ablation scenarios.

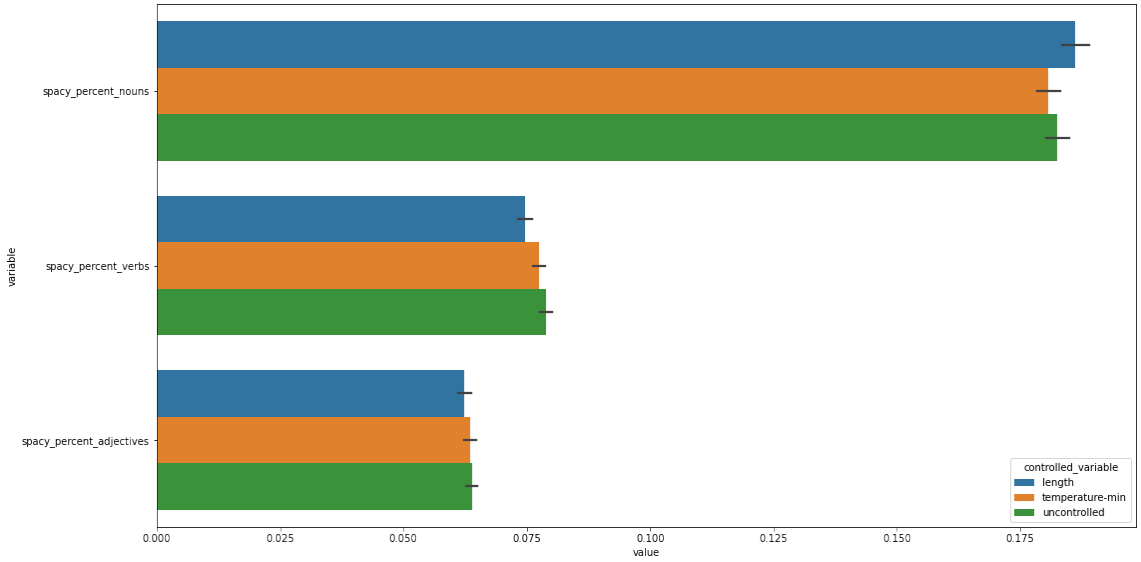

Figure 25: Distribution of spacy variables wrt the controlled variables (length and temperature).

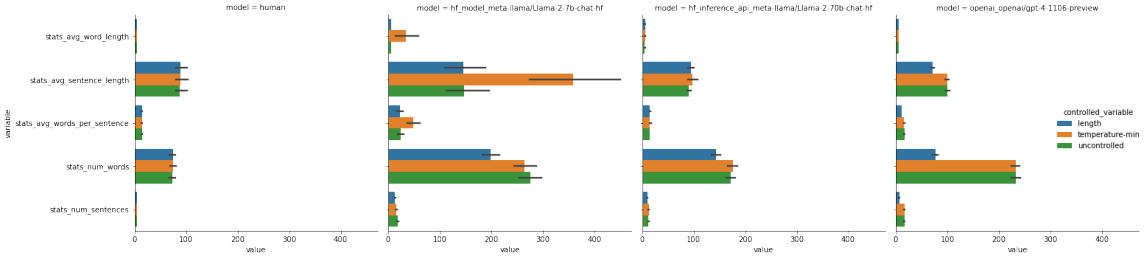

Figure 26: Distribution of stats variables per generative model wrt the controlled variables (length and temperature).

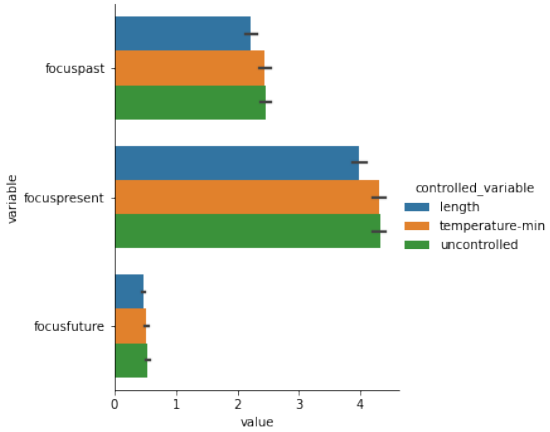

Figure 27: Distribution of time orientation variables wrt the controlled variables (length and temperature).

A.8 Surrogate models results

Due to a strong correlation between variable *stats_num_words* and LIWC variable *WC* (Spearman correlation coefficient = 1), variable *stats_percent_punctuation* and LIWC variable *AllPunc* (0.93) and variable *stats_percent_long_words* and LIWC variable *BigWords* (0.81), we eliminated those mentioned LIWC variables from (*WC*, *AllPunc*, *BigWords*) from further consideration (see Fig. 28). Also, readability features were not taken into consideration for surrogate model construction due to many missing values (limitation of readability metrics on short text).

Surrogate models were implemented using XGBoost library (with 10 repetitions). The training was done using 100 boosting rounds, logloss as an evaluation metric, 0.1 as eta (learning rate) and setting 5 as the maximum depth of each tree.

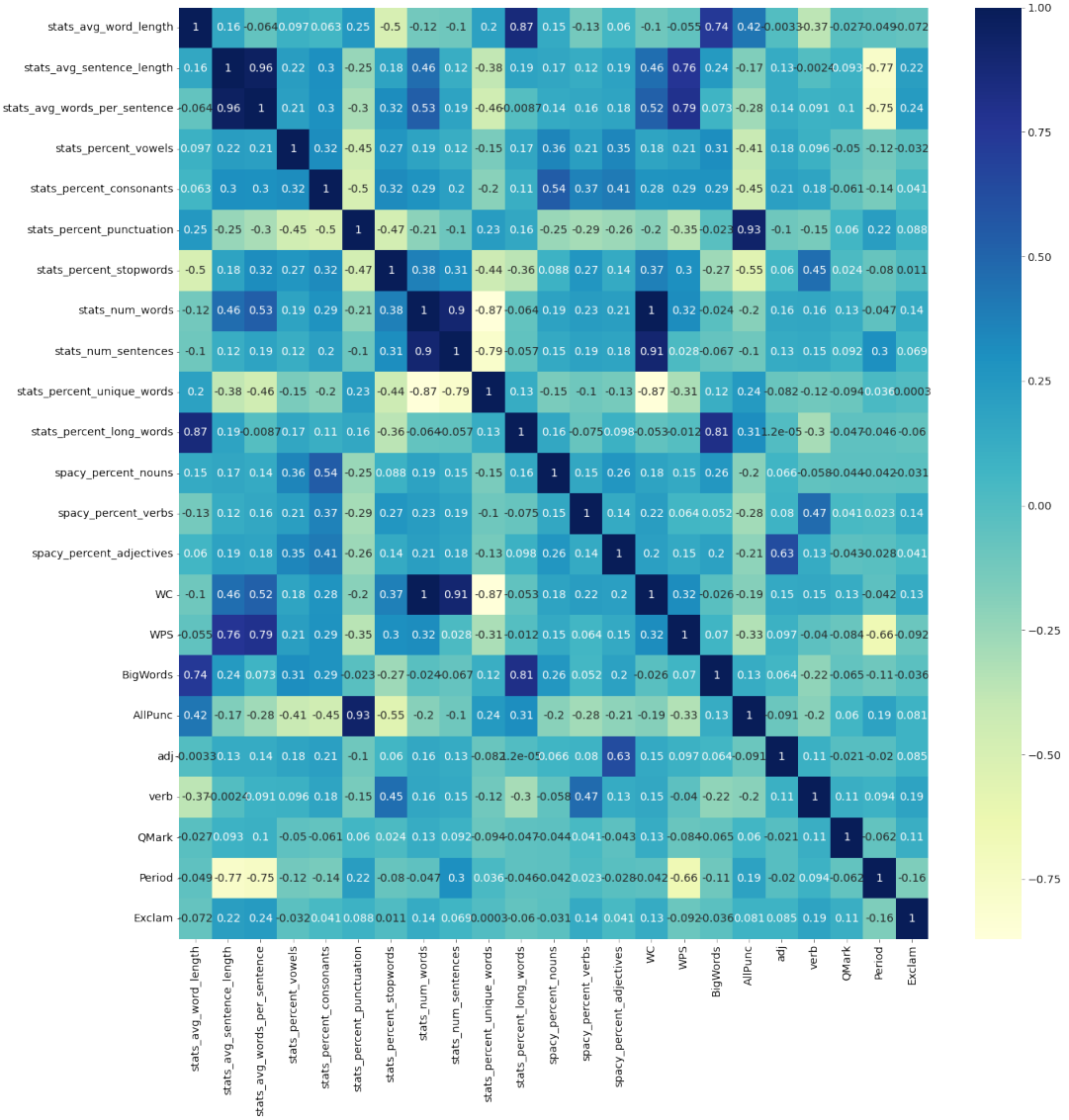

Figure 28: Correlation between input variables (excerpt))

The ROC curve for different surrogate classifiers in general can be seen in Fig. 29, while the ROC curve for the stance_generation (interesting due to different surrogate performance in terms of ROC AUC), can be found in Fig. 30.

As mentioned before, XGB surrogate performance by different tasks and generative models can be seen in Tab. 12 and Tab. 13 respectively, while the corresponding important features can be found in Fig. 31 and Fig. 32, respectively.

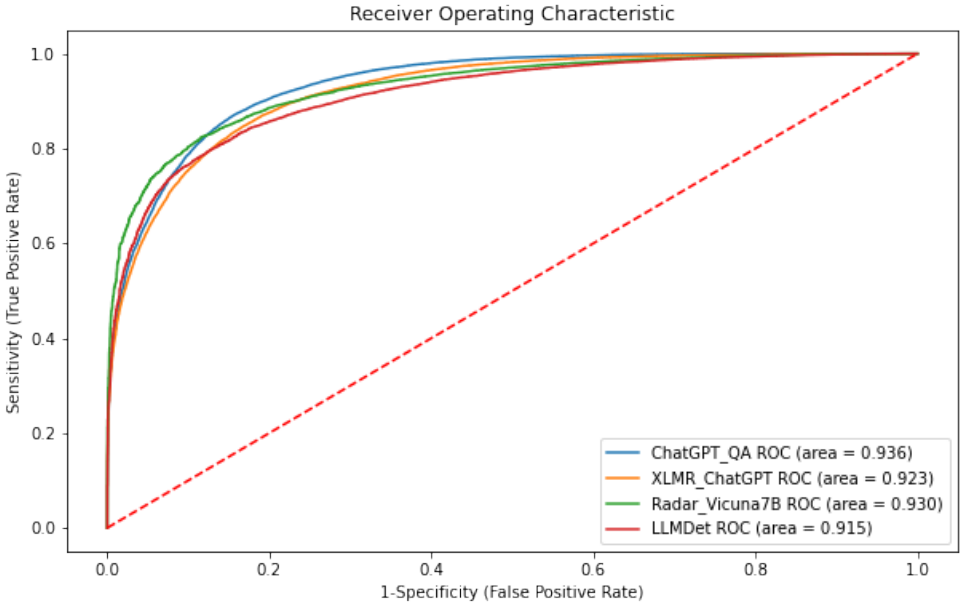

Figure 29: ROC curve for surrogate models (all tasks and generative models)

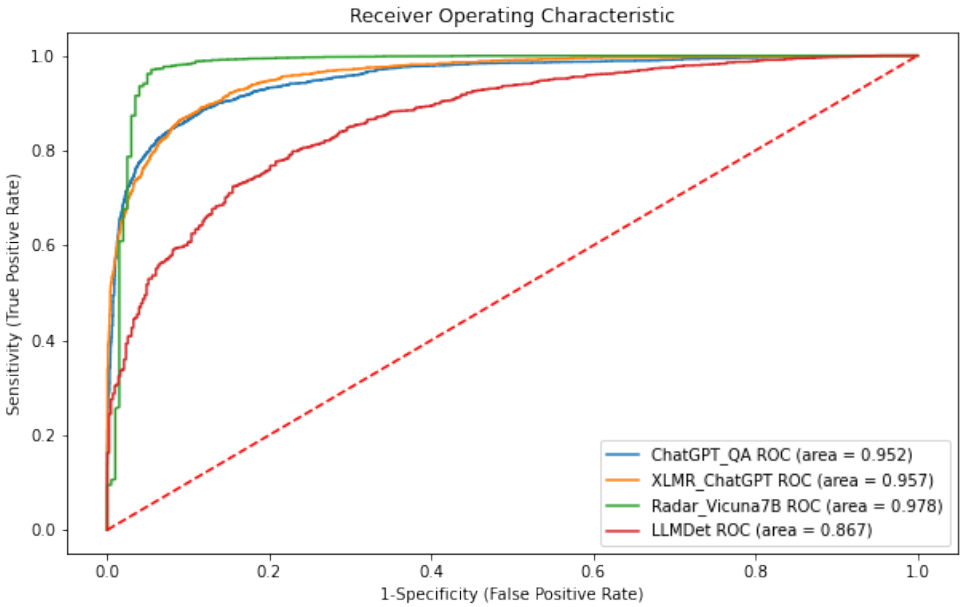

Figure 30: ROC curve for surrogate models for stance_generation task.

Task	Detector	F1-weighted	F1-macro	F1-micro	ROC AUC	MCC
brainstorming	ChatGPT_QA	0.875 (0.016)	**0.846 (0.018)**	0.878 (0.015)	**0.937 (0.007)**	**0.7 (0.034)**
	XLMR_ChatGPT	0.832 (0.01)	0.784 (0.013)	0.84 (0.01)	0.895 (0.014)	0.585 (0.03)
	Radar_Vicuna7B	0.891 (0.009)	0.64 (0.016)	0.912 (0.007)	0.863 (0.024)	0.349 (0.038)
	LLMDet	**0.908 (0.012)**	0.687 (0.038)	**0.922 (0.009)**	0.915 (0.016)	0.413 (0.069)
classification	ChatGPT_QA	0.878 (0.01)	**0.876 (0.009)**	0.877 (0.01)	**0.951 (0.007)**	**0.752 (0.019)**
	XLMR_ChatGPT	0.846 (0.011)	0.843 (0.011)	0.846 (0.011)	0.92 (0.012)	0.687 (0.022)
	Radar_Vicuna7B	**0.963 (0.009)**	0.568 (0.061)	**0.973 (0.006)**	0.899 (0.02)	0.194 (0.139)
	LLMDet	0.855 (0.019)	0.718 (0.029)	0.871 (0.015)	0.872 (0.012)	0.468 (0.053)
closed_qa	ChatGPT_QA	0.86 (0.011)	**0.86 (0.011)**	0.86 (0.011)	**0.935 (0.01)**	**0.721 (0.022)**
	XLMR_ChatGPT	0.802 (0.006)	0.787 (0.006)	0.804 (0.006)	0.875 (0.012)	0.579 (0.013)
	Radar_Vicuna7B	**0.904 (0.017)**	0.628 (0.04)	**0.923 (0.012)**	0.905 (0.014)	0.315 (0.074)
	LLMDet	0.866 (0.023)	0.663 (0.04)	0.886 (0.017)	0.867 (0.019)	0.373 (0.065)
creative_writing	ChatGPT_QA	0.836 (0.011)	**0.764 (0.014)**	0.844 (0.008)	0.879 (0.009)	**0.541 (0.022)**
	XLMR_ChatGPT	0.804 (0.022)	0.764 (0.027)	0.809 (0.021)	0.872 (0.019)	0.535 (0.055)
	Radar_Vicuna7B	0.881 (0.017)	0.607 (0.027)	0.908 (0.012)	0.86 (0.022)	0.306 (0.057)
	LLMDet	**0.927 (0.014)**	0.584 (0.032)	**0.943 (0.009)**	**0.904 (0.03)**	0.227 (0.075)
general_qa	ChatGPT_QA	0.881 (0.013)	**0.751 (0.026)**	0.892 (0.012)	**0.899 (0.015)**	**0.529 (0.056)**
	XLMR_ChatGPT	0.87 (0.014)	0.686 (0.016)	0.887 (0.012)	0.865 (0.015)	0.414 (0.035)
	Radar_Vicuna7B	0.887 (0.009)	0.546 (0.028)	0.917 (0.007)	0.84 (0.012)	0.192 (0.074)
	LLMDet	**0.901 (0.014)**	0.619 (0.032)	**0.919 (0.01)**	0.878 (0.024)	0.29 (0.065)
information_extraction	ChatGPT_QA	0.901 (0.012)	**0.891 (0.012)**	0.901 (0.012)	**0.954 (0.008)**	**0.783 (0.023)**
	XLMR_ChatGPT	0.83 (0.016)	0.826 (0.016)	0.83 (0.016)	0.907 (0.011)	0.653 (0.033)
	Radar_Vicuna7B	**0.910 (0.011)**	0.648 (0.03)	**0.924 (0.009)**	0.915 (0.012)	0.336 (0.059)
	LLMDet	0.874 (0.015)	0.689 (0.03)	0.89 (0.01)	0.862 (0.02)	0.414 (0.051)
open_qa	ChatGPT_QA	0.885 (0.013)	**0.84 (0.016)**	0.889 (0.012)	**0.923 (0.012)**	**0.688 (0.03)**
	XLMR_ChatGPT	0.855 (0.019)	0.747 (0.029)	0.867 (0.015)	0.88 (0.015)	0.518 (0.049)
	Radar_Vicuna7B	**0.894 (0.013)**	0.61 (0.03)	**0.916 (0.010)**	0.841 (0.025)	0.288 (0.064)
	LLMDet	0.891 (0.014)	0.606 (0.032)	0.914 (0.008)	0.892 (0.014)	0.28 (0.048)
review_generation	ChatGPT_QA	0.832 (0.009)	0.82 (0.008)	0.834 (0.008)	0.914 (0.007)	0.644 (0.016)
	XLMR_ChatGPT	0.843 (0.005)	**0.84 (0.005)**	0.844 (0.005)	0.926 (0.004)	**0.681 (0.01)**
	Radar_Vicuna7B	**0.976 (0.004)**	0.768 (0.025)	**0.980 (0.003)**	**0.945 (0.007)**	0.596 (0.039)
	LLMDet	0.926 (0.008)	0.783 (0.014)	0.933 (0.007)	0.938 (0.006)	0.585 (0.026)
stance_generation	ChatGPT_QA	0.898 (0.011)	0.885 (0.012)	0.898 (0.01)	0.952 (0.008)	0.771 (0.023)
	XLMR_ChatGPT	0.901 (0.011)	0.874 (0.013)	0.903 (0.011)	0.957 (0.007)	0.75 (0.026)
	Radar_Vicuna7B	**0.988 (0.004)**	**0.896 (0.042)**	**0.989 (0.004)**	**0.979 (0.02)**	**0.798 (0.083)**
	LLMDet	0.919 (0.011)	0.596 (0.033)	0.938 (0.007)	0.87 (0.016)	0.281 (0.075)
summarization	ChatGPT_QA	0.841 (0.01)	**0.821 (0.013)**	0.845 (0.01)	**0.907 (0.009)**	**0.649 (0.025)**
	XLMR_ChatGPT	0.804 (0.01)	0.774 (0.014)	0.809 (0.01)	0.878 (0.011)	0.555 (0.027)
	Radar_Vicuna7B	0.84 (0.022)	0.651 (0.03)	0.865 (0.014)	0.848 (0.018)	0.358 (0.047)
	LLMDet	**0.922 (0.012)**	0.643 (0.041)	**0.937 (0.009)**	0.89 (0.03)	0.343 (0.091)

Table 12: XGBoost classifier performance **per task** in terms of F1-score, ROC AUC and MCC across 10 runs. The best scores per task and performance metric are displayed in the bold-face.

Generator model	Detector	F1-weighted	F1-macro	F1-micro	ROC AUC	MCC
LLaMA-2-70b*	ChatGPT_QA	0.879 (0.009)	**0.873 (0.010)**	0.88 (0.009)	**0.948 (0.008)**	**0.748 (0.02)**
	XLMR_ChatGPT	0.854 (0.01)	0.853 (0.01)	0.854 (0.01)	0.935 (0.005)	0.708 (0.019)
	Radar_Vicuna7B	0.937 (0.008)	0.538 (0.028)	**0.956 (0.005)**	0.844 (0.036)	0.198 (0.068)
	LLMDet	**0.94 (0.007)**	0.634 (0.028)	0.952 (0.004)	0.914 (0.014)	0.321 (0.059)
Guanaco-7b	ChatGPT_QA	0.936 (0.007)	**0.875 (0.011)**	0.937 (0.006)	**0.965 (0.005)**	**0.753 (0.022)**
	XLMR_ChatGPT	0.901 (0.007)	0.83 (0.012)	0.905 (0.006)	0.944 (0.007)	0.666 (0.021)
	Radar_Vicuna7B	0.938 (0.007)	0.722 (0.024)	0.947 (0.005)	0.941 (0.01)	0.484 (0.041)
	LLMDet	**0.961 (0.008)**	0.584 (0.036)	**0.971 (0.006)**	0.92 (0.025)	0.243 (0.087)
LLaMA-2-7b*	ChatGPT_QA	0.836 (0.007)	**0.811 (0.007)**	0.84 (0.007)	**0.908 (0.007)**	**0.63 (0.014)**
	XLMR_ChatGPT	0.807 (0.013)	0.807 (0.012)	0.807 (0.012)	0.897 (0.008)	0.614 (0.025)
	Radar_Vicuna7B	0.921 (0.007)	0.507 (0.024)	**0.946 (0.005)**	0.825 (0.02)	0.086 (0.089)
	LLMDet	**0.929 (0.008)**	0.591 (0.025)	0.945 (0.004)	0.89 (0.013)	0.251 (0.048)
Falcon-7b*	ChatGPT_QA	0.783 (0.013)	0.778 (0.013)	0.783 (0.013)	0.874 (0.012)	0.556 (0.026)
	XLMR_ChatGPT	0.766 (0.014)	0.755 (0.014)	0.767 (0.013)	0.84 (0.013)	0.512 (0.027)
	Radar_Vicuna7B	**0.966 (0.009)**	**0.784 (0.049)**	**0.968 (0.007)**	**0.967 (0.01)**	**0.576 (0.091)**
	LLMDet	0.803 (0.013)	0.601 (0.022)	0.838 (0.01)	0.801 (0.008)	0.266 (0.045)
Falcon-7b-Dolly	ChatGPT_QA	0.94 (0.006)	**0.871 (0.011)**	0.942 (0.006)	**0.963 (0.007)**	**0.748 (0.021)**
	XLMR_ChatGPT	0.912 (0.007)	0.801 (0.014)	0.918 (0.005)	0.945 (0.008)	0.616 (0.025)
	Radar_Vicuna7B	**0.963 (0.008)**	0.691 (0.044)	**0.970 (0.006)**	0.945 (0.02)	0.429 (0.083)
	LLMDet	0.959 (0.007)	0.637 (0.05)	0.968 (0.005)	0.923 (0.017)	0.33 (0.1)
Human	ChatGPT_QA	0.869 (0.019)	0.612 (0.034)	0.892 (0.013)	0.877 (0.01)	0.27 (0.055)
	XLMR_ChatGPT	0.812 (0.016)	0.669 (0.019)	0.83 (0.011)	0.857 (0.011)	0.364 (0.031)
	Radar_Vicuna7B	**0.931 (0.011)**	**0.858 (0.018)**	**0.933 (0.011)**	**0.958 (0.008)**	**0.719 (0.036)**
	LLMDet	0.792 (0.015)	0.741 (0.018)	0.801 (0.013)	0.845 (0.012)	0.495 (0.031)
GPT-3.5	ChatGPT_QA	0.876 (0.013)	**0.857 (0.014)**	0.878 (0.012)	**0.933 (0.007)**	**0.719 (0.029)**
	XLMR_ChatGPT	0.809 (0.011)	0.774 (0.013)	0.815 (0.011)	0.885 (0.009)	0.558 (0.026)
	Radar_Vicuna7B	**0.967 (0.003)**	0.612 (0.035)	**0.976 (0.002)**	0.923 (0.02)	0.331 (0.075)
	LLMDet	0.892 (0.009)	0.64 (0.026)	0.911 (0.007)	0.869 (0.018)	0.337 (0.048)
GPT-4	ChatGPT_QA	0.823 (0.015)	0.817 (0.016)	0.823 (0.015)	0.905 (0.01)	0.635 (0.032)
	XLMR_ChatGPT	0.867 (0.008)	**0.867 (0.008)**	0.867 (0.008)	**0.937 (0.006)**	**0.735 (0.017)**
	Radar_Vicuna7B	0.874 (0.011)	0.655 (0.024)	0.892 (0.006)	0.878 (0.012)	0.353 (0.036)
	LLMDet	**0.941 (0.009)**	0.552 (0.036)	**0.956 (0.005)**	0.903 (0.02)	0.152 (0.075)

Table 13: XGBoost classifier performance **per generator model** in terms of F1-score, ROC AUC and MCC across 10 runs. The best scores per generator and performance metric are displayed in the bold-face.

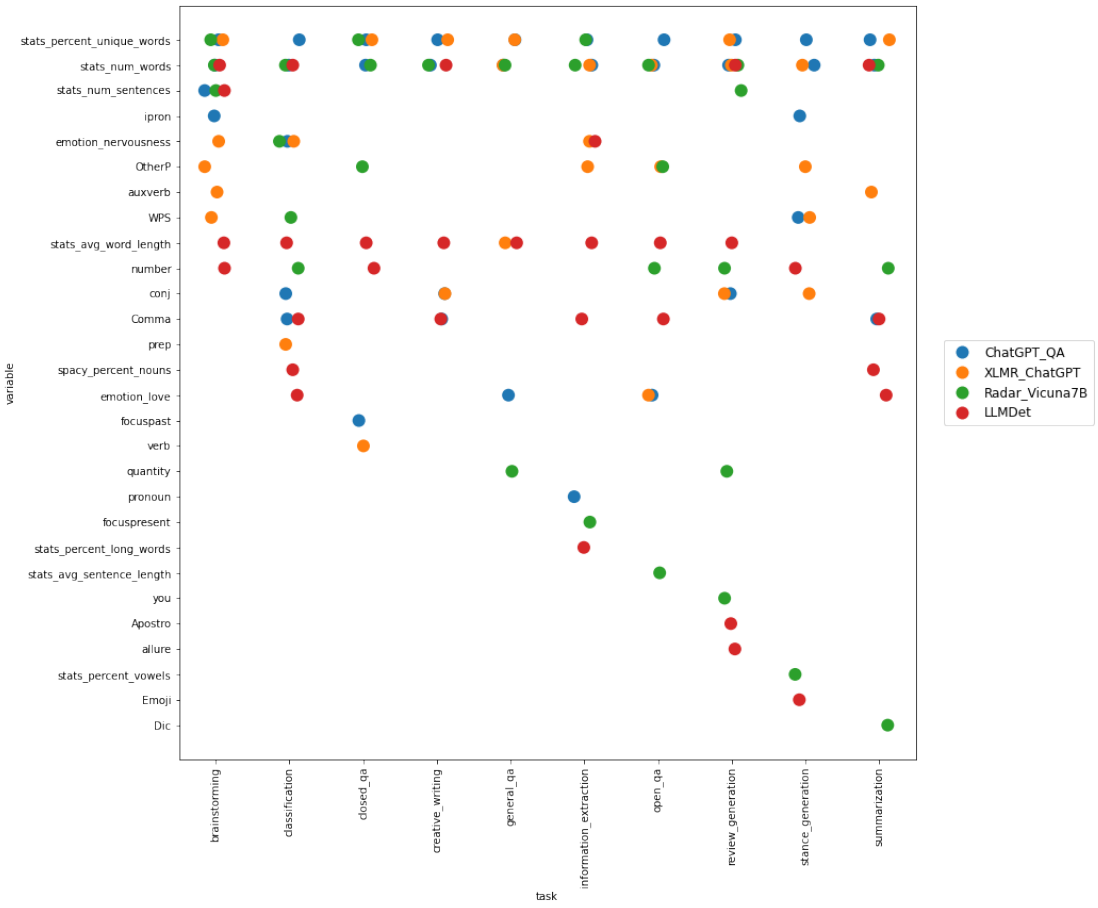

Figure 31: Feature importance for the classifiers by task

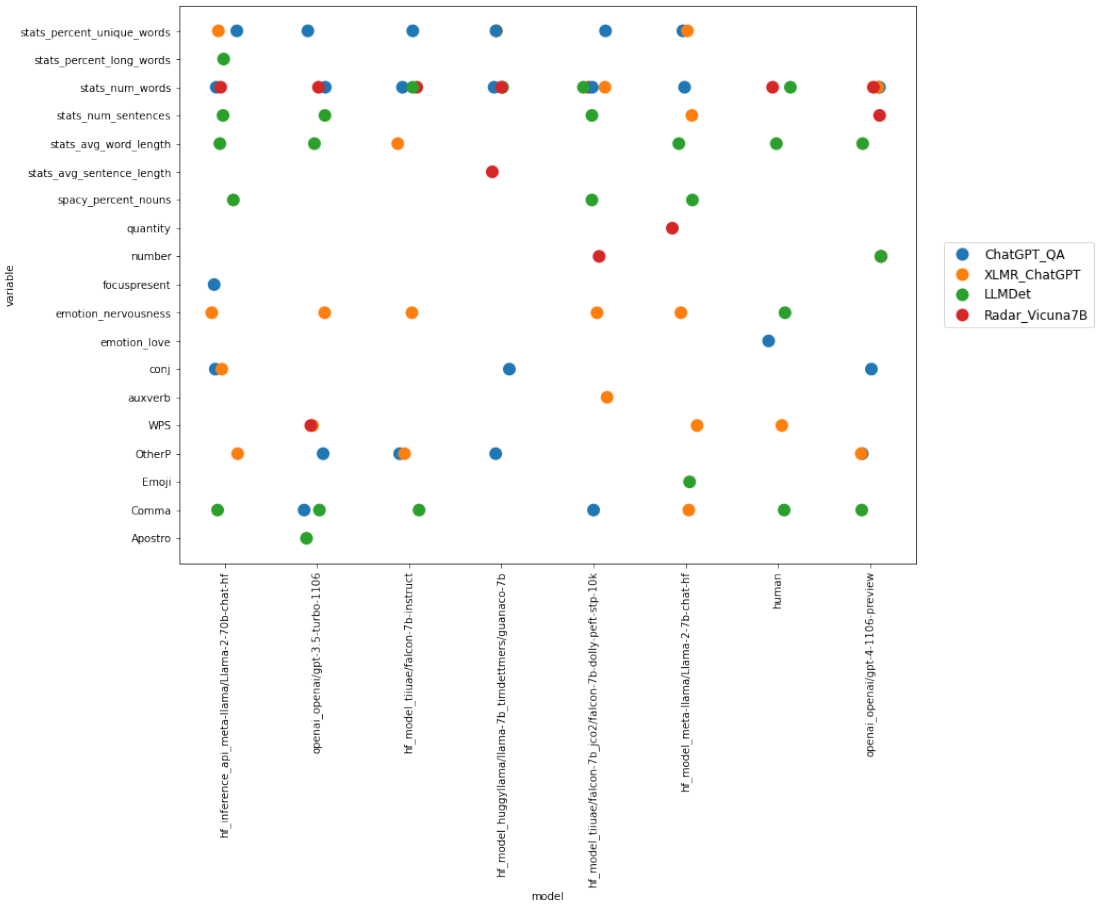

Figure 32: Feature importance for the classifiers by generative model