OTTO VON GUERICKE UNIVERSITÄT MAGDEBURG

Team Project

Examining the Robustness and Resilience of AI-Generated Text Detectors

Project Supervisors:

Prof. Dr.-Ing. Andreas Nürnberger
M.Sc. Marcus Thiel

Matrkl. Nr	Name
244815	Supriya Pandurangacharya Upadhyaya
244267	Aarathi Vijayachandran
244684	Shashankh Mysore Girish

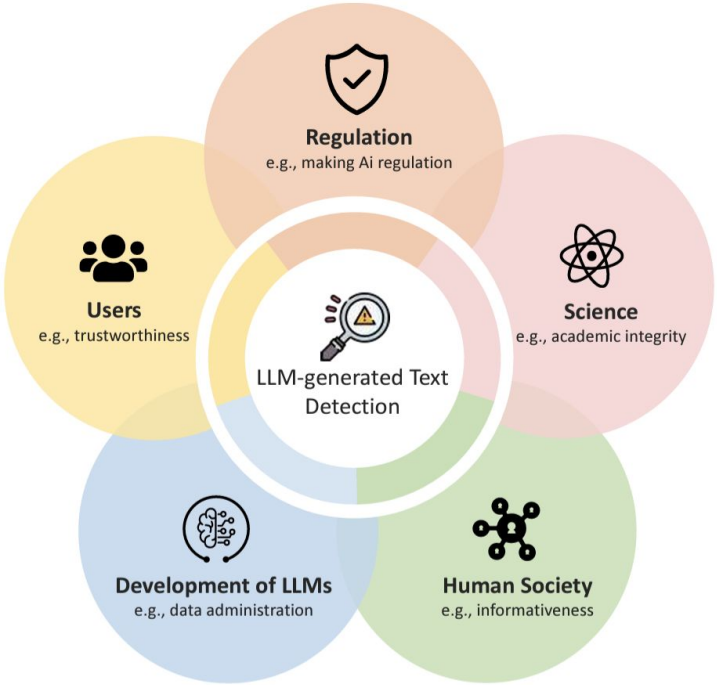
Agenda

- Motivation
- Research Aim
- Literature survey
  - Existing detectors
  - Similar studies
- Experimental setup
  - Dataset
  - Detectors under study
  - Experimental setup
- Results and observations
- Limitation and future work

Motivation

Motivation

- The LLM-generated text flood - rapid integration into daily life
- Need for LLM text detectors
  - Mitigate misuse & harm
  - Combat misinformation
  - Prevent abuse
  - Protect Users
  - Maintain Trust
  - Responsible AI
- Detector performance assertions



Source: Wu et al. (2023)

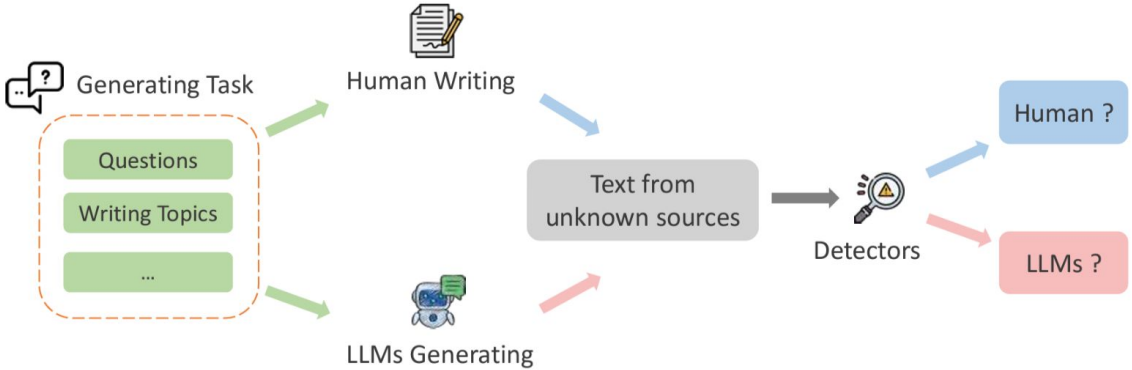
Research Aim

Research Aim

Main Objective: To critically evaluate the robustness and resilience of popular AI-generated text detectors across different LLM-generated text, linguistic styles, domains, and adversarial scenarios.

Key Questions:

1. How do detectors perform when faced with text from diverse LLMs, different domains, adversarial or prompt attacks?
2. What is minimal fine-tuning data required to enhance the performance of detector for generations from a new language model?



Source: Wu et al. (2023)

Literature Survey - Similar Studies

Similar Studies

- **Adversarial Vulnerabilities:** AI detectors, including watermarking methods, are susceptible to attacks like paraphrasing, leading to evasion challenges. *(Sadasivan et al., 2023)*
- **Methodological Challenges:** Out-of-distribution issues, model ambiguity, and evolving LLM capabilities hinder detection reliability, emphasizing the need for adaptable systems. *(Wu et al., 2023)*
- **Bias in Detection Tools:** Academic detection tools frequently misclassify AI-generated text as human, exposing limitations in their effectiveness. *(Weber et al., 2023)*
- **Adversarial Attacks:** Minor perturbations can significantly compromise detector performance, highlighting robustness challenges. *(Zhou et al., 2023)*
- **Innovative Solutions:** The Siamese Calibrated Reconstruction Network (SCRN) improves detection accuracy by 6.5–18.25% under adversarial conditions. *(Huang et al., 2023)*

Research Gaps

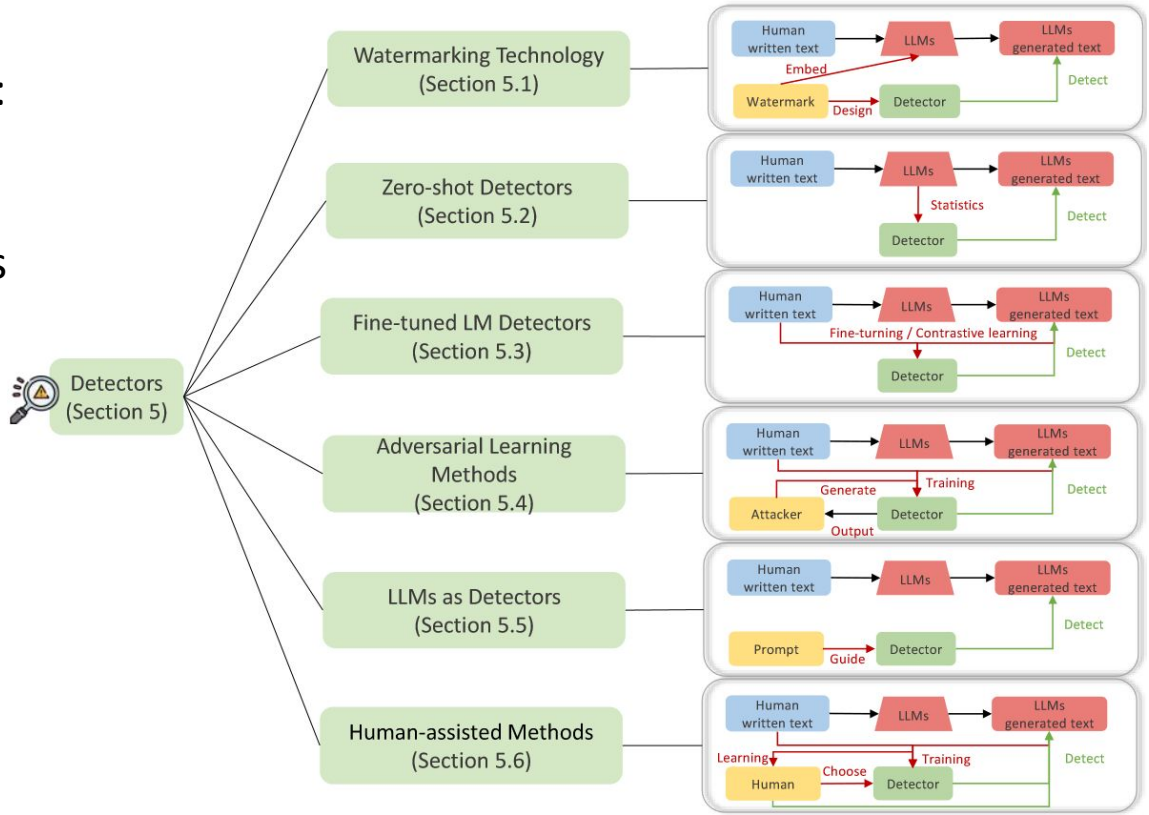
- **Limitations in Detection Models:** Existing approaches rely heavily on either LLM-based or traditional ML-based methods, but their combined effectiveness remains underexplored.
- **Subset Training Efficiency:** The effectiveness of minimal training subsets in enhancing detector performance for new language models remains uncertain.
- **Quality Analysis Deficiency:** Existing studies lack comprehensive benchmarks for evaluating detection methods and fail to analyze the underlying reasons for detection outcomes, such as identifying which features contribute most to classifier predictions

Literature Survey - Detectors

AI Generated Text Detectors

In our study we have considered:

- Fine-tuned LM Detectors
- Machine Learning Detectors (Classifiers)



Source: Wu et al. (2023)

Methodology - Datasets

M4: Multi-Generator, Multi-Domain, and Multi-Lingual Black-Box Machine-Generated Text Detection Dataset

Source/ Domain	Data License	Language	Total Human	Parallel Data							
				Human	Davinci003	ChatGPT	GPT4	Cohere	Dolly-v2	BLOOMz	Total
Wikipedia	CC BY-SA-3.0	English	6,458,670	3,000	3,000	2,995	3,000	2,336	2,702	3,000	20,033
Reddit ELI5	Huggingface	English	558,669	3,000	3,000	3,000	3,000	3,000	3,000	3,000	21,000
WikiHow	CC-BY-NC-SA	English	31,102	3,000	3,000	3,000	3,000	3,000	3,000	3,000	21,000
PeerRead	Apache license	English	5,798	5,798	2,344	2,344	2,344	2,344	2,344	2,344	19,862
arXiv abstract	CC0-public domain	English	2,219,423	3,000	3,000	3,000	3,000	3,000	3,000	3,000	21,000
Arabic-Wikipedia	CC BY-SA-3.0	Arabic	1,209,042	3,000	–	3,000	–	–	–	–	6,000
True & Fake News	MIT License	Bulgarian	94,000	3,000	3,000	3,000	–	–	–	–	9,000
Baike/Web QA	MIT license	Chinese	113,313	3,000	3,000	3,000	–	–	–	–	9,000
id_newspapers_2018	CC BY-NC-SA-4.0	Indonesian	499,164	3,000	–	3,000	–	–	–	–	6,000
RuATD	Apache 2.0 license	Russian	75,291	3,000	3,000	3,000	–	–	–	–	9,000
Urdu-news	CC BY 4.0	Urdu	107,881	3,000	–	3,000	–	–	–	–	6,000
Total				35,798	23,344	32,339	14,344	13,680	14,046	14,344	147,895

Source: Wang, Yuxia et al. (2023)

M4: Multi-Generator, Multi-Domain, and Multi-Lingual Black-Box Machine-Generated Text Detection Dataset

gpt-3.5-turbo


```
{
  "prompt": "Generate a 150-220-word abstract for work with title
    : Polymer Quantum Mechanics and its Continuum Limit. Describe
    ...",
  "human_text": "  A rather non-standard quantum representation
    of the canonical commutation\nrelations of quantum mechanics
    systems, known as the polymer representation has\ngained some
    attention in recent years, due to its possible relation with
    Planck\nscale physics. In...",
  "machine_text": "In this paper, we investigate the continuum
    limit of polymer quantum mechanics. The aim of our work is to
    better understand the quantum behavior of polymers and its
    relationship to the classical theory. Our research attempts
    to clarify the nature of the continuum limit and determine
    the correctness of the approximations made. Specifically, we
    study the connectivity of the polymer strands. We find that a
    continuous path can be defined in the polymer configuration
    space, which is essential for the polymer to display
    continuum behavior. We show that the approximation of
    continuous randomness in the spatial variable can certainly
    be used when the polymer is stretched out over a significant
    distance. For a coil-like polymer, our results demonstrate
    that its behavior is non-classical.\n...",
  "model": "gpt-3.5-turbo",
  "source": "arxiv",
  "source_ID": 704.0007
}
```


bigscience/bloomz


```
{
  "source_id": 704.0007,
  "title": "Polymer Quantum Mechanics and its Continuum Limit",
  "abstract": "  A rather non-standard quantum representation of
    the canonical commutation\nrelations of quantum mechanics
    systems, known as the polymer representation has\ngained
    some attention in recent years, due to its possible relation
    with Planck\nscale physics. In...",
  "machine_abstract": "The present work is devoted to the study
    of quantum mechanics in polymer chains, which are modeled as
    one-dimensional systems with nearest-neighbor interactions
    between particles.  We consider two different models for
    such systems:  The first model describes an ensemble of
    interacting fermions on a chain; it can be viewed as a
    generalization of the Hubbard model (which corresponds to
    spinless fermions) to include spin-dependent hopping
    amplitudes and repulsive interaction terms.  In this case we
    show that there...",
  "prompt": "Write a long abstract of a scientific paper from
    arXiv.org. Use approximately 200-400 words.\nTitle:
    \"{title}\".\n...]",
  "model": "bigscience/bloomz",
  "source": "arxiv"
}
```


Source: Wang, Yuxia et al. (2023)

Pre-processing M4 Dataset

New line characters, LaTeX code, tab characters, non-printable characters, and Unicode symbols removed

Original

```
We introduce the notion of Landau (\\Gamma, \\chi)-automorphic functions of magnitude $\\nu$ for any integer $\\nu \\geq 0$ and show that they are holomorphic sections of certain line bundles over the complex flag manifold $\\mathbb{C}^{N} / \\Lambda^{(\\nu)}_{\\mathbb{N}} = \\mathrm{SL}_N(\\mathbb{C})/\\mathrm{Sp}_{2N - N_0}(\\mathbb{C})$. We also prove an analogue of the Riemann-Roch theorem in this setting which allows us to compute the dimension of these spaces explicitly as a function of $\\nu$. Finally
```

```
In this paper, we present our X-ray timing observations of PSR J1930+1852 in the Crab-like supernova remnant G54.1+0.3. Our main motivation for this research is to study the timing properties of this pulsar and infer its physical characteristics as well as the characteristics of its environment. \n\nWe used data from the Chandra X-ray Observatory to carry out an in-depth study of the pulsar's timing properties. Our analysis revealed that PSR J1930+1852 has a stable rotation period with a spin-down rate of (9.1\u00b10.7)\u00d710^\u221211 s/s. We also detected significant pulse profile variations over time, suggesting the presence of magnetospheric emission modulated by the rotation of the pulsar. \n\nFurthermore, we found a correlation between the pulse profile and the X-ray
```

Pre-processed

```
We introduce the notion of Landau (Γ, χ)-automorphic functions of magnitude ν for any integer ν≥ 0 and show that they are holomorphic sections of certain line bundles over the complex flag manifold ℂ^N / Λ^(ν)_ℕ = SL_N(ℂ)/Sp_2N - N_0(ℂ).We also prove an analogue of the Riemann-Roch theorem in this setting which allows us to compute the dimension of these spaces explicitly as a function of ν. Finally we give some examples of explicit bases for these spaces. This is joint work with Jens Franke. The results presented here were obtained while I was at
```

```
In this paper, we present our X-ray timing observations of PSR J1930+1852 in the Crab-like supernova remnant G54.1+0.3. Our main motivation for this research is to study the timing properties of this pulsar and infer its physical characteristics as well as the characteristics of its environment.We used data from the Chandra X-ray Observatory to carry out an in-depth study of the pulsar's timing properties. Our analysis revealed that PSR J1930+1852 has a stable rotation period with a spin-down rate of (9.1±0.7)×10^-11 s/s. We also detected significant pulse profile variations over time, suggesting the presence of magnetospheric emission modulated by the rotation of the pulsar.Furthermore, we found a correlation between the pulse profile and the X-ray luminosity, with a
```

Source: Wang, Yuxia et al. (2023)

Data Splits

Total 6000 abstracts = 3000 human-written abstracts + 3000 corresponding machine-generated abstracts								
Training - 64%			**Validation - 16%**			**Test - 20%**		
Total	Human	Machine	Total	Human	Machine	Total	Human	Machine
3840	1920	1920	960	480	480	1200	600	600

Evaluation Metrics

	Predicted Machine-Generated	**Predicted Human-Produced**
True Machine - Generated	True Positives (TP)	False Negatives (FN)
True Human - Produced	False Positives (FP)	True Negatives (TN)

- ➔ Accuracy
- ➔ Precision
- ➔ Recall
- ➔ F1-Score

Methodology - Detectors Under Study

Baseline: LLM-Based Detectors

RoBERTa-Academic-Detector and BLOOMZ-560M-Academic-Detector

(Bentzen Winje and Sivesind (2023))

Base-model	RoBERTa-base	Bloomz-560m	Bloomz-1b7	Bloomz-3b
Wiki	roberta-wiki	Bloomz-560m-wiki	Bloomz-1b7-wiki	Bloomz-3b-wiki
Academic	roberta-academic	Bloomz-560m-academic	Bloomz-1b7-academic	Bloomz-3b-academic
Mixed	roberta-mixed	Bloomz-560m-mixed	Bloomz-1b7-mixed	Bloomz-3b-mixed

Datasets

The models were trained on selections from the GPT-wiki-intros and ChatGPT-Research-Abstracts, and are separated into three types, **wiki**-detectors, **academic**-detectors and **mixed**-detectors, respectively.

In-domain peformance of academic-detectors:

Base model	Accuracy	Precision	Recall	F1-score
Bloomz-560m	0.964	0.963	0.965	0.964
Bloomz-1b7	0.946	0.941	0.951	0.946
Bloomz-3b	*0.984	*0.983	0.985	*0.984
RoBERTa	0.982	0.968	*0.997	0.982

Baseline - ML Model

XGBoost Classifier (Desaire et al. (2023b))

- Using 20 text features from 4 categories
  - Paragraph complexity
  - Sentence-level diversity in length,=
  - Differential usage of punctuation marks
  - Different 'popular words'
- **99% accuracy** claimed

Table 1. Features in the model

Feature number	Feature type (1–4)[a]	Short description	Greater in
1	1	sentences per paragraph	human
2	1	words per paragraph	human
3	2	")" present	human
4	2	"-" present	human
5	2	";" or ":" present	human
6	2	"?" present	human
7	2	"'" present	ChatGPT
8	3	standard deviation in sentence length	human
9	3	length difference for consecutive sentences	human
10	3	sentence with <11 words	human
11	3	sentence with >34 words	human
12	4	contains "although"	human
13	4	contains "However"	human
14	4	contains "but"	human
15	4	contains "because"	human
16	4	contains "this"	human
17	4	contains "others" or "researchers"	ChatGPT
18	4	contains numbers	human
19	4	contains 2 times more capitals than "."	human
20	4	contains "et"	human

[a]Feature types: 1, paragraph complexity; 2, punctuation marks; 3, diversity in sentence length; and 4, popular words or numbers.

Source: Desaire et al. (2023b), Desaire et al. (2023a)

Methodology - Experimental Setup

Experiment : Cross-Model Robustness

- **Aim:** Test detectors on texts generated by models other than those used for training.
- **Dataset:**
  - M4 arXiv abstract dataset (ChatGPT, BLOOMZ, Davinci, Cohere, Flan-T5)
- Detectors evaluated before and after fine-tuning

Experiment : Cross-Domain Robustness

- **Aim:** Evaluate detector models' ability to generalize across different domains.
- **Datasets:**
  - **New ML-Llama-3 Dataset**
    - Extracted 3000 human-written ML abstracts from **ArxivPapers dataset** (Kardas et al. (2020)).
    - Generated machine-generated abstracts using **Llama-3**.
    - **Final Dataset**: Contains 2973 human-machine abstract pairs.
  - M4 arXiv abstract dataset (ChatGPT, BLOOMZ)
  - **Extension of M4 dataset** with machine-generated texts from **Llama-3** articles dataset
- Conducted single and multiple levels of fine-tuning and evaluation

Experiment : Estimation of Minimum Fine-tuning Data Subset

- **Aim:** Determine the minimum training data required for optimal performance.
- **Dataset:** M4 arXiv abstract dataset (Cohere)
- Conducted only for RoBERTa and BLOOMZ Academic Detectors
- Fine-tuned models on subsets of training data in geometric progression
  → (1, 2, 4, 8, 10, 15, 20, 25, 30, 35, 40, 45, 50, 60, 70, 80, 90, 100)%

Experiment : Robustness Against Adversarial Attacks

- **Aim:** Evaluate the model's ability to correctly classify adversarially manipulated texts as machine-generated.
- **Dataset:** M4 arXiv abstract dataset (ChatGPT, BLOOMZ)
- Used **TextAttack Framework** (Morris et al. (2020)) to apply adversarial attacks

Attack Type	**Transformation**
deepwordbug (Gao et al. (2018))	Character Insertion, Deletion, Swap, Substitution
pruthi (Pruthi, Dhingra, and Lipton (2019))	Character Insertion, Deletion, Swap
pwws (Ren et al. (2019))	Synonym Swap
textfooler (Jin et al. (2020))	Word Embedding Swap

Experiment : Robustness Against Prompt Attacks

- **Aim:** Evaluate resistance to input alterations using prompt engineering
- **Prompt attack types:** spelling/punctuation/grammar errors, change tense/voice, synonyms, informal language or slang, sentence restructuring, idiomatic expressions
- **Dataset:**
  - **Extension of M4 dataset**
    - **Machine-generated texts from Llama-3.1:**
      - Generated 150-250 word abstracts using arXiv paper titles as prompts.
      - Selected 100 random human + machine abstract pairs.
    - **Prompt-attack-oriented machine-generated text:**
      - For each prompt-attack type: LLaMA-3.1-generated abstracts + attack-instruction prompts → **Gemini-1.5-Flash** model → **prompt-attacked machine-generated abstracts**

Experiment : Robustness Against Prompt Attacks

We investigate gas-grain chemistry in cold interstellar cloud cores, where chemical reactions occur on dust grain surfaces and within the surrounding gas phase. Our research aims to address the challenge of accurately describing these complex processes using a microscopic Monte Carlo approach. This method allows us to simulate the stochastic nature of surface reactions, incorporating factors such as dust grain size distribution, temperature fluctuations, and the impact of cosmic rays. By employing this approach, we are able to investigate the formation and destruction of molecular species on grain surfaces and their subsequent desorption into the gas phase. Our results provide new insights into the chemical richness and diversity present in cold interstellar cloud cores, shedding light on the origins of complex organic molecules found in these environments. Furthermore, our work contributes to a deeper understanding of the role that dust grains play in regulating the chemical composition of the surrounding gas, with implications for the formation of stars and planets.

Machine-generated M4 arXiv abstract from LLaMA-3.1

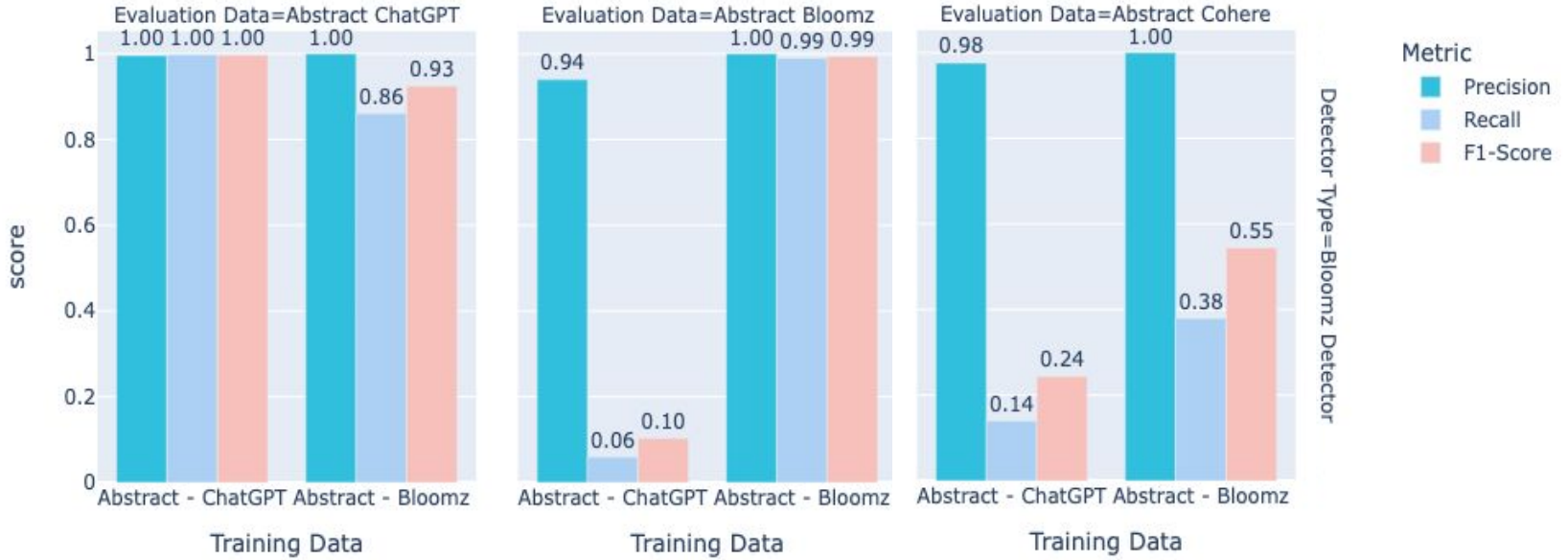
We examine gas-grain chemistry in frigid interstellar cloud cores, where chemical processes take place on dust grain surfaces and within the surrounding gaseous medium. Our investigation aims to tackle the challenge of accurately portraying these intricate processes using a microscopic Monte Carlo method. This technique permits us to simulate the random nature of surface reactions, incorporating factors such as dust grain size distribution, temperature variations, and the influence of cosmic rays. By utilizing this approach, we are able to explore the formation and breakdown of molecular species on grain surfaces and their subsequent release into the gaseous phase. Our findings offer novel insights into the chemical abundance and variety present in cold interstellar cloud cores, illuminating the origins of complex organic molecules found in these environments. Moreover, our work contributes to a deeper comprehension of the function that dust grains play in controlling the chemical makeup of the surrounding gas, with implications for the genesis of stars and planets.

Prompt-attacked machine-generated M4 arXiv abstract from Gemini-1.5-Flash

Example of prompt attack on a machine-generated M4 arXiv abstract, where words are replaced with their synonyms without changing the overall meaning of the text.

Results and Observations

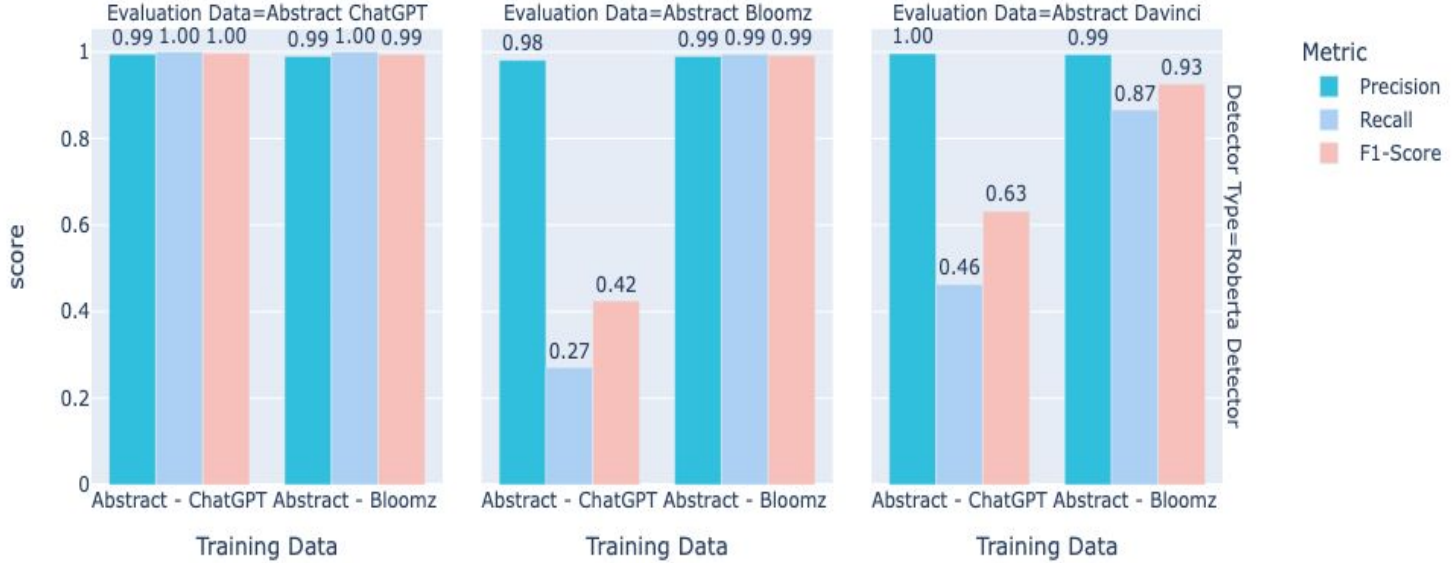
Cross-Model Evaluation - Bloomz Detector



Bloomz detector has poor cross-model performance unless fine-tuned on cross-model generations

Exception is ChatGPT generations since pre-trained Bloomz detector has already seen ChatGPT research abstracts

Cross-Model Evaluation - Roberta Detector



Roberta detector fine-tuned on Bloomz training set has good cross-model performance across all the evaluation sets

Roberta detector fine-tuned on ChatGPT training set does NOT hold good cross-model performance

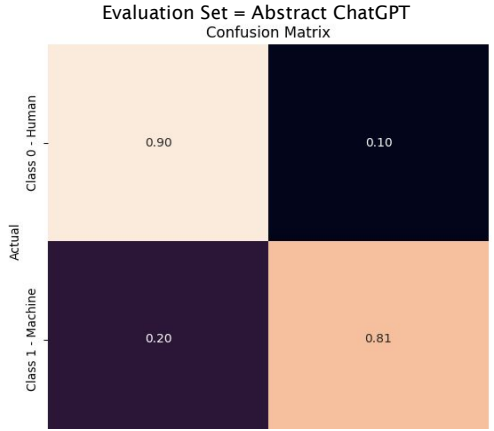
Cross-Model Evaluation - XGBoost Detector

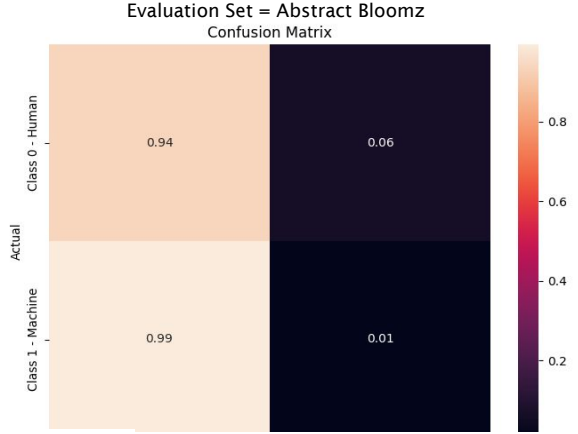
XGBoost Detectors performance is dependent on the hand crafted features generations and hence cross-model performance is poor

Cross-Model Evaluation - XGBoost

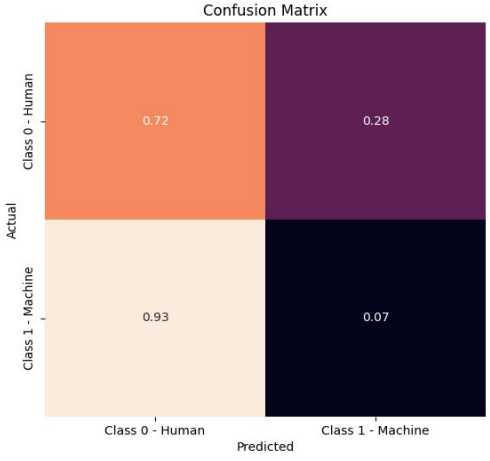
XGBoost detector always misclassifies “Machine Text” as “Human Text”

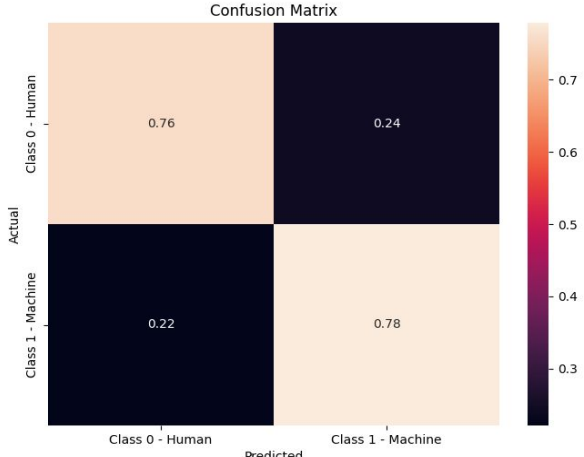
Training Set = Abstract ChatGPT





Training Set = Abstract Bloomz

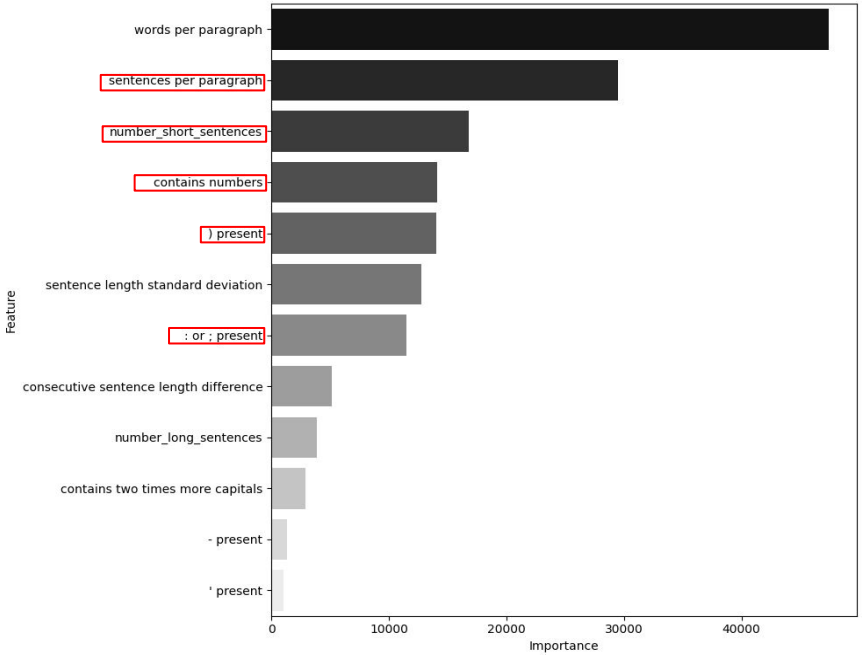




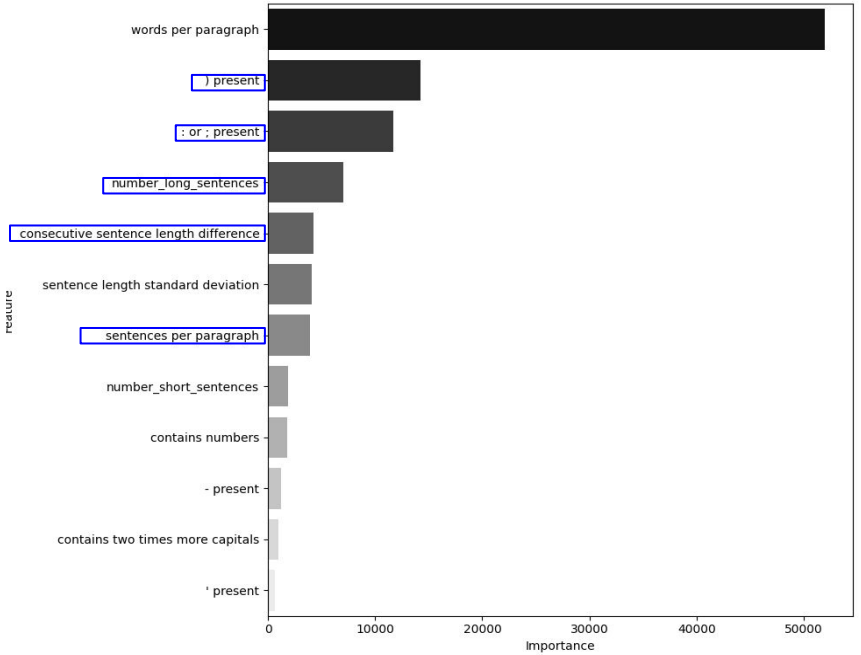
Cross-Model Evaluation - XGBoost

Feature Importance

XGBoost Trained on ChatGPT Abstracts



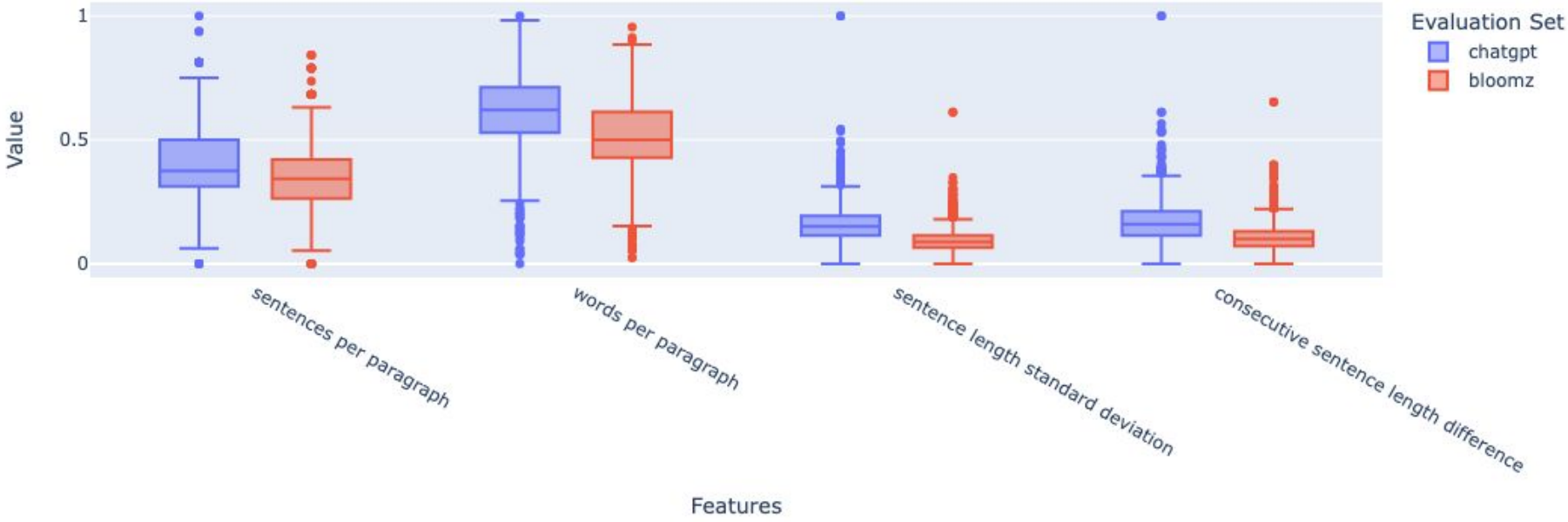
XGBoost Trained on Bloomz Abstracts



OTTO VON GUERICKE UNIVERSITÄT MAGDEBURG

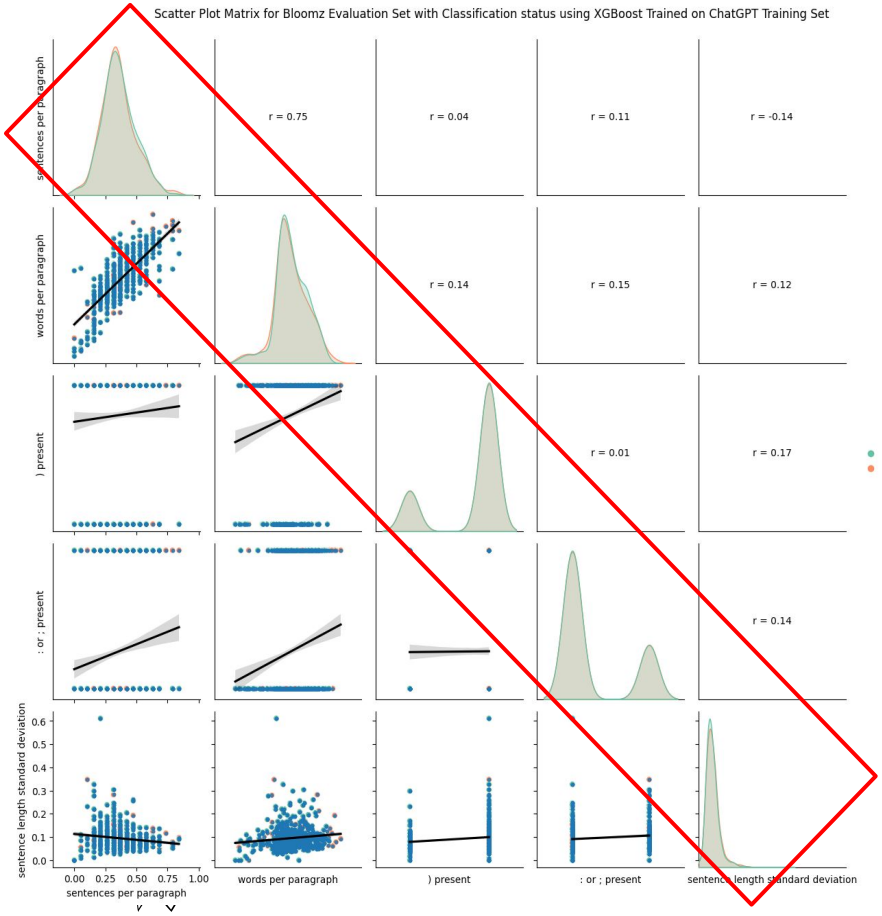
Cross-Model Evaluation - XGBoost

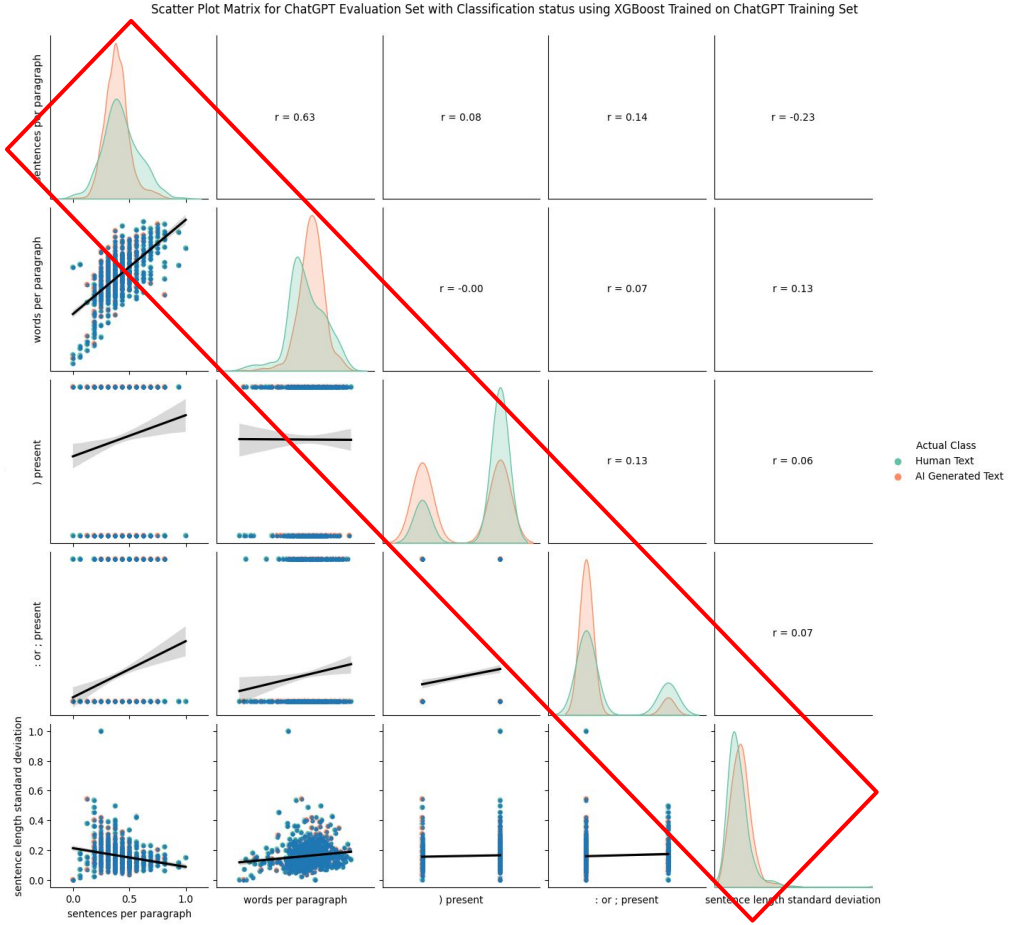
Evaluation dataset text Style features of ChatGPT Vs Bloomz generated abstract



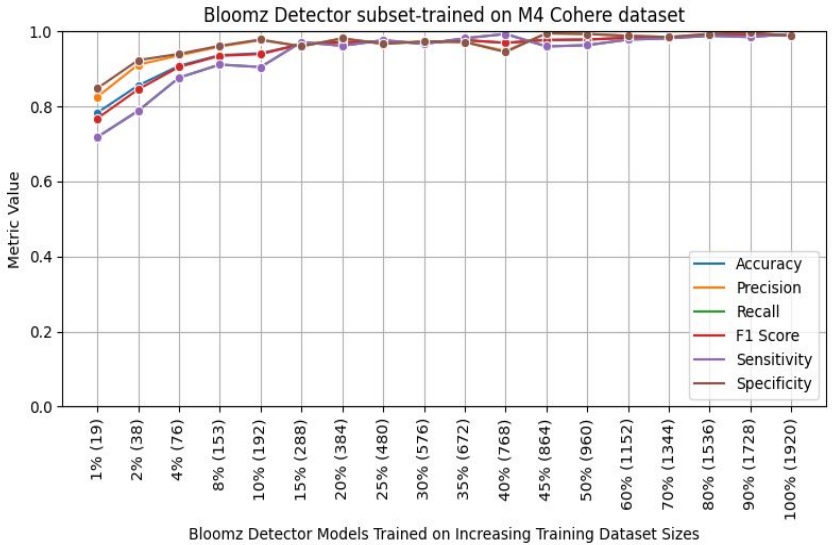
Cross-Model Evaluation - XGBoost

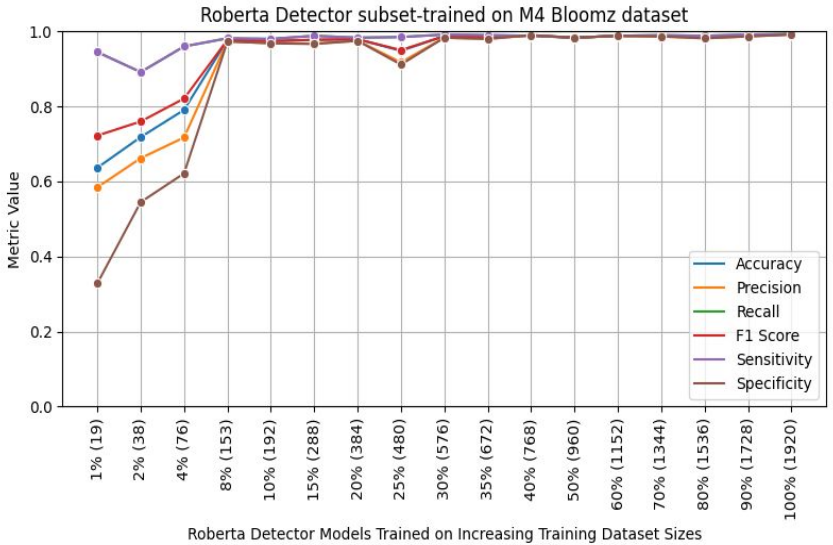
Comparison of Bloomz Vs ChatGPT Text Features distribution for “Human Text” and “Machine Text” Classes





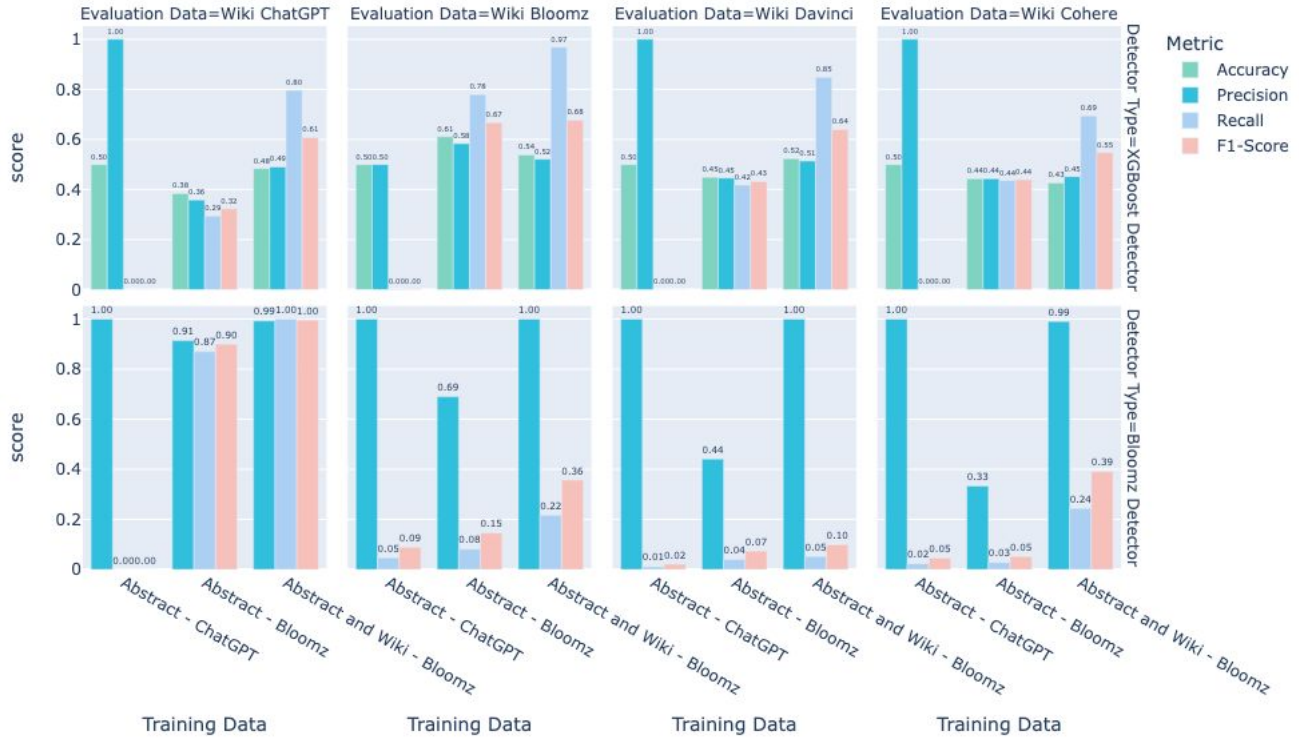
Estimation of Minimum Fine-tuning Data Subset



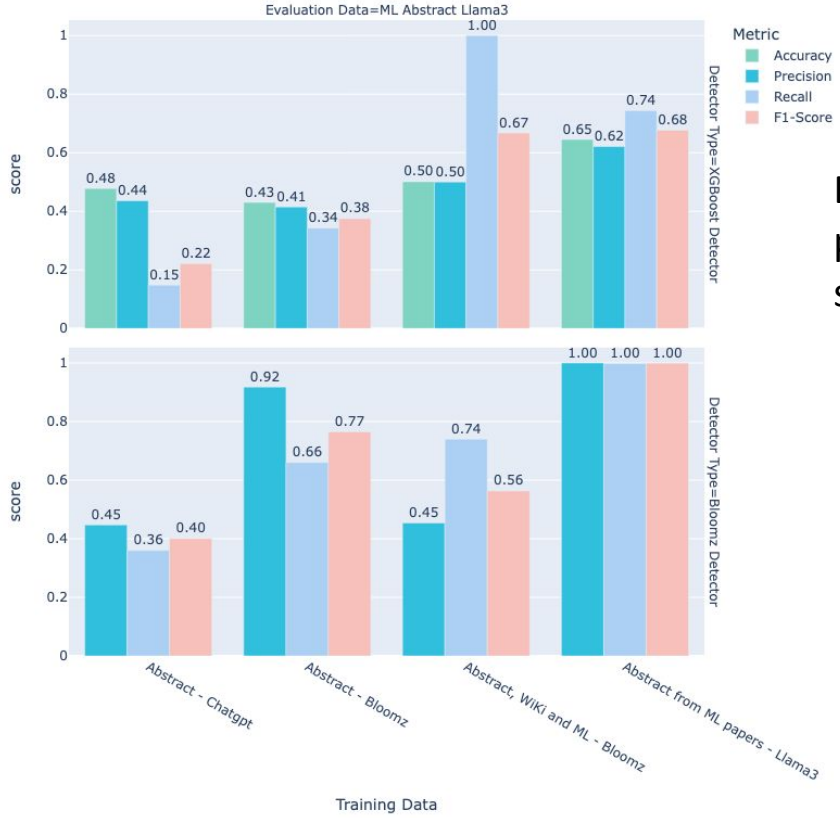


At least 8% to 15% (153 to 288) of generations from new model is required for LLM-based detectors to adapt

Cross-Domain Evaluation - XGBoost and Bloomz Detectors

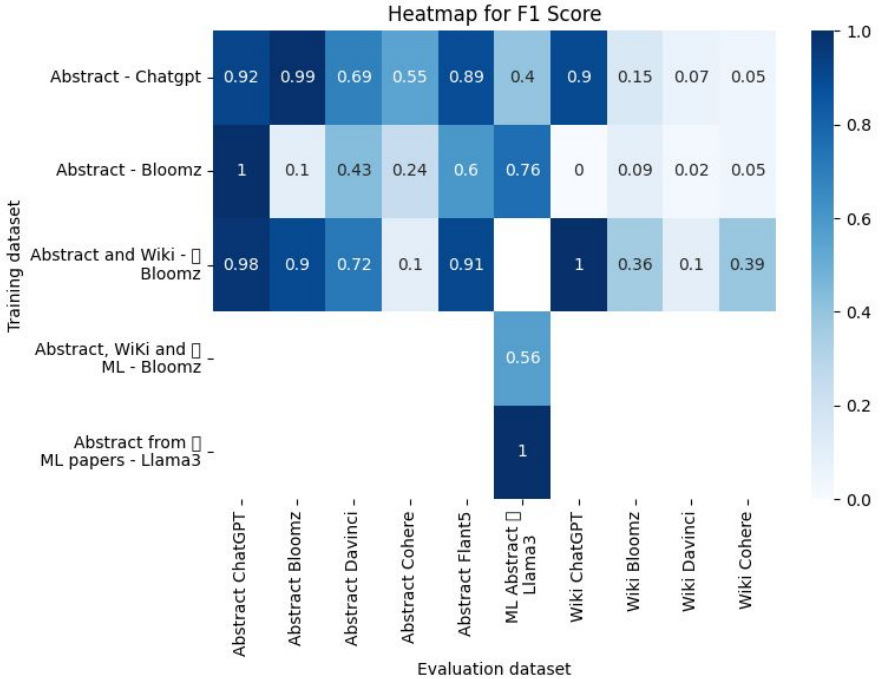


Cross-Domain Evaluation - Fine-tune on Multiple Domains

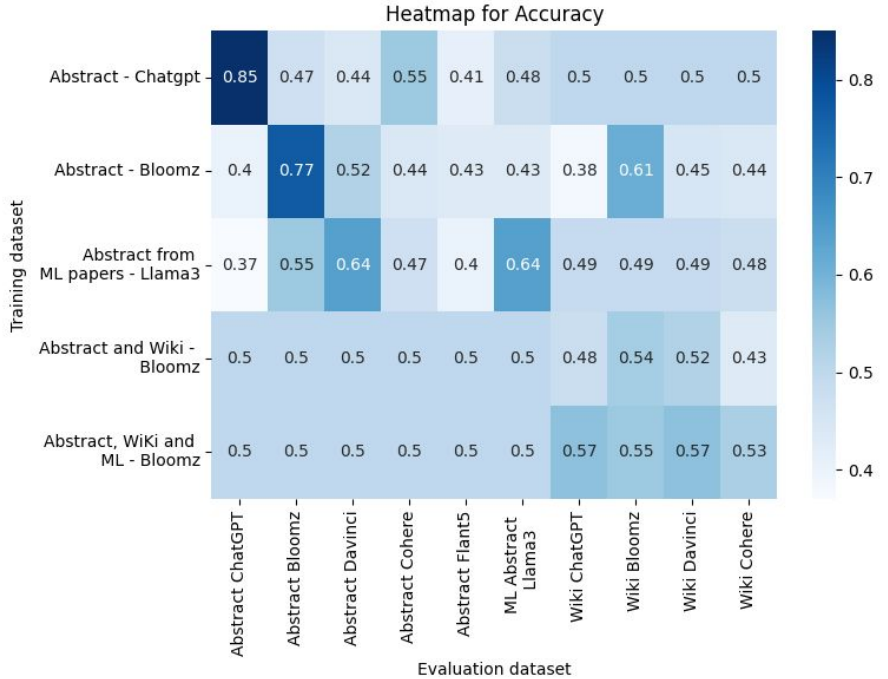


Both XGBoost detector and Bloomz detector perform best when trained/fine-tuned on a single domain

Evaluation metrics for Bloomz and XGBoost based AI generated text detectors

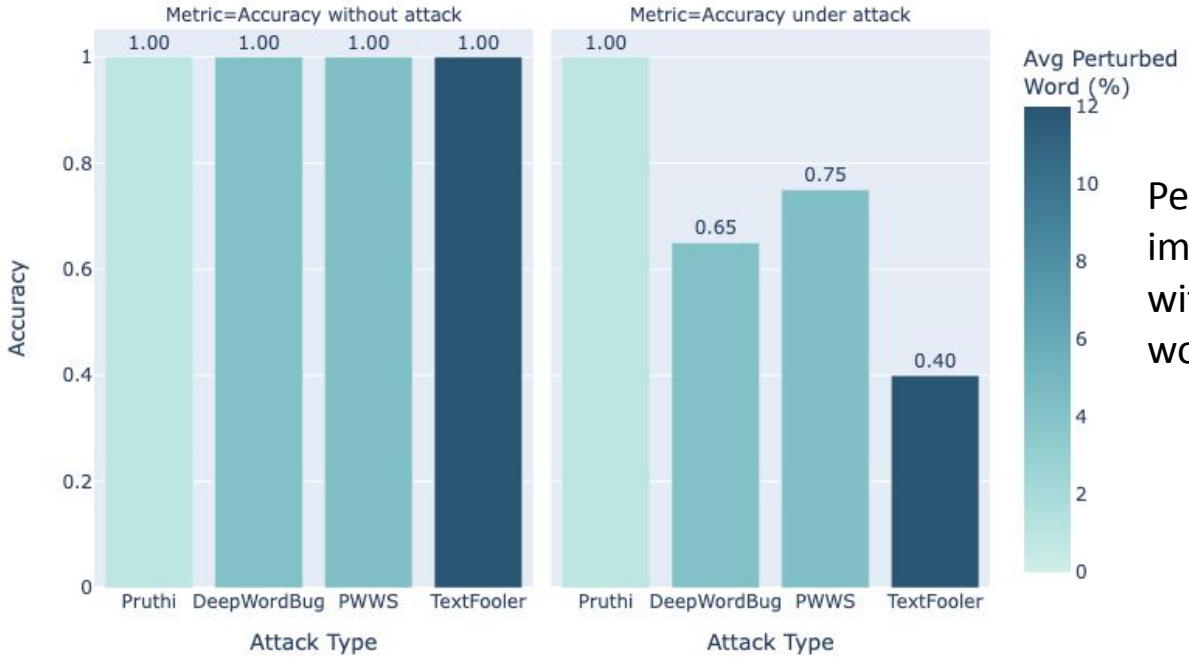


F1 Score heatmap for Bloomz based AI generated text detectors



Accuracy heatmap for XGBoost Classifier based AI generated text detectors

Adversarial Attack



Performance of detectors impacted by adversarial text with the number of perturbed word as small as ~4%

Adversarial Attack

The nature of the perturbations suggests that these tests may not fully reflect practical adversarial scenarios

- Disrupt the original meaning
- Unrealistic and nonsensical text

original_text	perturbed_text
We address the problem of constructing high-accuracy, faithful analytic waveforms describing the gravitational wave signal emitted by inspiralling and coalescing binary black holes. We work within the Effective-One-Body (EOB) framework and propose a methodology for improving the current (waveform)implementations of this framework based on understanding, element by element, the physics behind each feature of the waveform, and on systematically comparing various EOB-based waveforms with ``exact'' waveforms obtained by numerical relativity approaches. The present paper focuses on small-mass-ratio non-spinning binary systems, which can be conveniently studied by Regge-Wheeler-Zerilli-type methods. Our results include: (i) a resummed, 3PN-accurate description of the inspiral waveform, (ii) a better description of radiation reaction during the plunge, (iii) a refined analytic expression for the plunge waveform, (iv) an improved treatment of the matching between the plunge and ring-down waveforms. This improved implementation of the EOB approach allows us to construct complete analytic waveforms which exhibit a remarkable agreement with the ``exact'' ones in modulus, frequency and phase. In particular, the analytic and numerical waveforms stay in phase, during the whole process, within $\pm 1.1 %$ of a cycle. We expect that the extension of our methodology to the comparable-mass case will be able to generate comparably accurate analytic waveforms of direct use for the ground-based network of interferometric detectors of gravitational waves.	We address the problem of constructing high-accuracy, faithful analytic waveforms describing the gravitational wave signal emitted by inspiralling and coalescing binary Negro holes. We work within the Effective-One-Body (EOB) framework and propose a methodology for improving the current (waveform)implementations of this framework based on understanding, element by element, the physics behind each feature of the waveform, and on systematically comparing various EOB-based waveforms with ``exact'' waveforms obtained by numerical relativity approaches. The acquaint paper focuses on small-mass-ratio non-spinning binary systems, which can be conveniently studied by Regge-Wheeler-Zerilli-type methods. Our results include: (i) a resummed, 3PN-accurate description of the inspiral waveform, (ii) a better description of radiation reaction during the plunge, (iii) a refined analytic expression for the plunge waveform, (iv) an improved discussion of the oppose between the plunge and ring-down waveforms. This improved implementation of the EOB approach allows us to build stark analytic waveforms which exhibit a remarkable understanding with the ``exact'' ones in modulus, frequency and phase. In particular, the analytic and numerical waveforms stay in phase, during the hale process, within $\Pm 1.1 %$ of a cycle. We expect that the extension of our methodology to the comparable-mass event will be able to give comparably accurate analytic waveforms of direct use for the ground-based network of interferometric detectors of gravitational waves.

(a) Example of pwws attack

original_text	perturbed_text
We address the problem of constructing high-accuracy, faithful analytic waveforms describing the gravitational wave signal emitted by inspiralling and coalescing binary black holes. We work within the Effective-One-Body (EOB) framework and propose a methodology for improving the current (waveform)implementations of this framework based on understanding, element by element, the physics behind each feature of the waveform, and on systematically comparing various EOB-based waveforms with ``exact'' waveforms obtained by numerical relativity approaches. The present paper focuses on small-mass-ratio non-spinning binary systems, which can be conveniently studied by Regge-Wheeler-Zerilli-type methods. Our results include: (i) a resummed, 3PN-accurate description of the inspiral waveform, (ii) a better description of radiation reaction during the plunge, (iii) a refined analytic expression for the plunge waveform, (iv) an improved treatment of the matching between the plunge and ring-down waveforms. This improved implementation of the EOB approach allows us to construct complete analytic waveforms which exhibit a remarkable agreement with the ``exact'' ones in modulus, frequency and phase. In particular, the analytic and numerical waveforms stay in phase, during the whole process, within $\pm 1.1 %$ of a cycle. We expect that the extension of our methodology to the comparable-mass case will be able to generate comparably accurate analytic waveforms of direct use for the ground-based network of interferometric detectors of gravitational waves.	We address the problem of constructing high-accuracy, faithful analytic waveforms describing the gravitational wave signal emitted by inspiralling and coalescing binary black holes. We work within the Effective-One-Body (OB) framework and propose a methodology for improving the current (wavform)implementations of this framework based on understanding, element by element, the physics behind each feature of the waveform, and on systematically comparing various EOB-baesd waveforms with ``exact'' waveforms obtained by numerical relativity approaches. The present paper focuses on small-mass-ratHo non-pinning binary systems, which can be conveniently studied by Regge-Wheeler-Berilli-type methods. Our results include: (i) a resummed, 3Pd-accurate description of the inspirTal waveform, (i) a better description of radiation reaction during the lunge, (iii) a refined anafytic expression for the plunge waveform, (iv) an improved treatment of the matching between the plunge and ring-down waveforms. TThis improved implementation of the EuB approach allows us to construct compelte analytic waveforms which exhibit a remarkable agreement with the ``exact'' ones in modulus, frequency and phase. In particular, the analytic and numerical waveforms stay in phase, during the hwole process, within $\pl 1.1 %$ of a cycle. We expect that the extension of our methodology to the comparable-mass case will be ble to generate comparably accurate analytic waveforms of direct use for the ground-based network of interfermoetric detectors of gravitational waves.

(b) Example for deepwordbug attack

Example of original text and perturbed text for adversarial attacks

Prompt Attack

RoBERTa Detector				
Prompt Attack Using Gemini-1.5-Flash \ Metric	Precision	Recall	Accuracy	F1 Score
Synonyms	1.0	1.0	1.0	1.0
Informal language or slang	1.0	0.86	0.93	0.92
Sentence restructuring	1.0	1.0	1.0	1.0
Idiomatic expressions	1.0	1.0	1.0	1.0
Grammar mistakes	1.0	1.0	1.0	1.0

Roberta model proved to be resilient to most of the prompt attacks except for ‘Informal language or slang’

Conclusion

Conclusion

- **Cross-model and cross-domain challenges** show strong within-model performance but significant drops in accuracy for cross-model and cross-domain
- **Fine-tuning** with minimal additional data (8–15%, ~153-288 samples) improves cross-model performance.
- **Adversarial attacks** degrade detection accuracy, highlighting weaknesses in robustness. However, these attacks are found to be practically questionable
- **Prompt attacks** though not found to significantly impact the detector performance it has scope for further investigation using advanced LLMs

Limitations and Future Work

Limitations

- Limited prompt attack evaluations
- No qualitative analysis of LLM-based detectors
- Dataset is limited to M4 and it’s extensions
- Restricted to lower parameter versions of the detectors

Future Work

- Incorporation of handcrafted features like syntactic, semantic, and statistical metrics
- Strategies for improved cross-domain generalization
- Real-world evaluation on diverse applications
- Exploration of ensemble learning for enhanced robustness


LIVE
Thank you for listening.
BET

References

[1] Heather Desaire, A. E. Chua, M.-G. Kim, and D. Hua. “Accurately detecting AI text when Chat-GPT is told to write like a chemist”. In: Cell Reports Physical Science 4.101672 (2023).
[2] Vinu Sankar Sadasivan, A. Kumar, S. Balasubramanian, W. Wang, and S. Feizi. “Can AI-generated text be reliably detected?” In: arXiv preprint arXiv:2303.11156v3 (2023). url: https://arxiv.org/abs/2303.11156v3.
[3] Junchao Wu, Shu Yang, Runzhe Zhan, Yulin Yuan, Derek F. Wong, and Lidia S. Chao. “A survey on LLM-generated text detection: Necessity, methods, and future directions”. In: arXiv preprint arXiv:2310.14724v2 (2023). url: https://arxiv.org/abs/2310.14724v2.
[4] Debora Weber-Wulff, Alla Anohina-Naumeca, Sonja Bjelobaba, Tomáš Foltýnek, Jean Guerrero-Dib, Olumide Popoola, Petr Šigut, and Lorna Waddington. “Testing of detection tools for AI-generated text”. In: International Journal for Educational Integrity 19.26 (2023).
[5] Chenghao Zhou, Yao Wang, and Jiahao Liu. “Adversarial Attacks on AI Text Detectors: A Study on Robustness”. In: arXiv preprint arXiv:2404.01907v1 (2023). url: https://arxiv.org/abs/2404.01907.
[6] Wei Huang, Bing Liu, and Shuming Zhang. “Siamese Calibrated Reconstruction Network for Robust AI Text Detection”. In: arXiv preprint arXiv:2406.01179v2 (2023). url: https://arxiv.org/abs/2406.01179.

References

[7] Andreas Bentzen Winje and Nicolai Sivesind. “Turning Poachers into Gamekeepers: Detecting Machine-Generated Text in Academia Using Large Language Models”. PhD thesis. June 2023.
[8] Colin B. Clement, Matthew Bierbaum, Kevin P. O’Keeffe, and Alexander A. Alemi. On the Use of ArXiv as a Dataset. 2019. arXiv: 1905.00075 [cs.IR].
[9] Heather Desaire, Aleesa E Chua, Madeline Isom, Romana Jarosova, and David Hua. “Distinguishing academic science writing from humans or ChatGPT with over 99% accuracy using off-the-shelf machine learning tools”. In: Cell Reports Physical Science 4.6 (2023).
[10] John X Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. “Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp”. In: arXiv preprint arXiv:2005.05909 (2020).
[11] Ji Gao, Jack Lanchantin, Mary Lou Soffa, and Yanjun Qi. “Black-box generation of adversarial text sequences to evade deep learning classifiers”. In: 2018 IEEE Security and Privacy Workshops (SPW). IEEE. 2018, pp. 50–56.
[12] Danish Pruthi, Bhuwan Dhingra, and Zachary C Lipton. “Combating adversarial misspellings with robust word recognition”. In: arXiv preprint arXiv:1905.11268 (2019).
[13] Shuhuai Ren, Yihe Deng, Kun He, and Wanxiang Che. “Generating natural language adversarial examples through probability weighted word saliency”. In: Proceedings of the 57th annual meeting of the association for computational linguistics. 2019, pp. 1085–1097.

References

[14] Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. “Is bert really robust? a strong baseline for natural language attack on text classification and entailment”. In: Proceedings of the AAAI conference on artificial intelligence. Vol. 34. 05. 2020, pp. 8018–8025.
[15] Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, and et al. “M4: Multi-Generator, Multi-Domain, and Multi-Lingual Black-Box Machine-Generated Text Detection”. In: arXiv preprint arXiv:2305.14902 (2023). url: https://arxiv.org/abs/2305.14902.
[16] Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. “Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense”. In: Advances in Neural Information Processing Systems 36 (2024).
[17] Marcin Kardas, Piotr Czapla, Pontus Stenetorp, Sebastian Ruder, Sebastian Riedel, Ross Taylor, and Robert Stojnic. “Axcell: Automatic extraction of results from machine learning papers”. In: arXiv preprint arXiv:2004.14356 (2020).