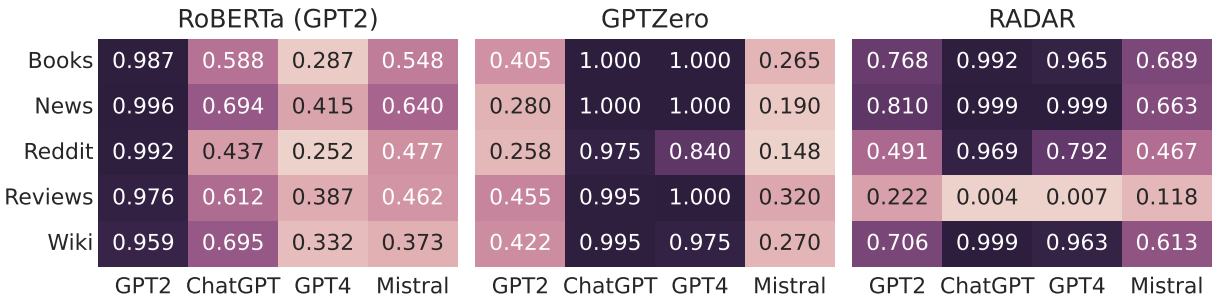

Figure 6: Heatmap measuring the accuracy of the RoBERTa-Large GPT2, GPTZero, and RADAR detectors across models and domains. We see a clear bias towards domains and models that the detectors have trained on.

	Open-Source								**Closed-Source**			
	Chat Models (llama-c, mistral-c, mpt-c)				**Non-Chat Models** (mistral, mpt, gpt2)				**Chat Models** (c-gpt, gpt4, cohere)		**Non-Chat Models** (cohere, gpt3)	
Dec. Strategy	greedy		sampling		greedy		sampling		greedy	sampling	greedy	sampling
Rep. Penalty?	✗	✓	✗	✓	✗	✓	✗	✓	✗	✗	✗	✗
R-B GPT2	84.1	52.3	77.9	26.2	98.6	44.1	60.5	35.4	70.9	41.7	65.1	52.5
R-L GPT2	79.7	41.1	71.4	19.5	98.5	43.0	67.2	53.4	61.4	34.7	61.1	48.6
R-B CGPT	80.2	63.3	75.0	39.3	53.3	26.4	14.9	1.7	59.1	38.1	46.5	39.0
RADAR	88.8	77.4	85.6	66.4	91.8	63.8	48.3	31.8	81.6	75.3	72.2	67.7
GLTR	89.8	67.5	83.9	38.3	99.6	56.9	44.5	0.5	80.7	54.3	75.6	63.7
F-DetectGPT	98.6	74.5	96.2	40.5	97.8	56.1	79.7	0.6	96.0	74.1	93.8	86.3
LLMDet	55.5	30.2	47.5	16.5	74.8	27.0	38.4	3.7	35.8	18.5	40.0	32.9
Binoculars	**99.9**	86.6	**99.7**	60.6	**99.9**	62.3	72.4	0.6	**99.2**	92.1	**99.0**	**95.0**
GPTZero	98.8	**93.7**	98.4	**82.5**	74.7	34.6	9.4	4.8	92.3	88.5	60.6	53.4
Originality	98.6	86.3	97.7	72.5	**99.9**	**64.1**	**89.0**	**51.2**	96.8	89.0	91.7	85.4
Winston	97.2	90.1	96.6	78.3	68.2	49.0	29.5	11.3	96.1	**93.7**	73.2	68.1
ZeroGPT(*)	95.4	80.7	90.5	54.9	85.1	57.2	16.0	0.3	92.1	65.8	83.4	72.7

Table 5: Accuracy Score at FPR=5% for all detectors across model groups and sampling strategies. Asterisks (*) indicate that the detector was unable to achieve the target FPR. We see that random sampling with a repetition penalty consistently makes output generations very difficult to detect, especially for open-source non-chat models.

alternative generation settings, especially if generative models are already trained on such outputs.

Finding 4: Seemingly strong, robust detectors can perform unexpectedly poorly The most accurate detectors—FastDetectGPT, Originality, Binoculars, etc.—may seem like reliable solutions to detection in general, but they sometimes deteriorate from perfect accuracy to complete failure (see Table 5). The changes in experimental settings were not particularly sophisticated either: simply changing the text generator, switching decoding strategies, or applying a repetition penalty was enough to introduce up to 95+% error rate. Our findings show that detectors tend not to generalize across different models or generation settings in the same domain. Compounded by the lack of evaluation at different false positive rates, domain-specific detectors for critical issues like fake news and education are particularly at risk of mislabeling human-written text as machine-generated without our awareness.

Finding 5: Detectors perform better on domains and models seen during training. Figure 6 shows the performance of RoBERTa-Large GPT2, GPTZero, and RADAR on a cross-section of models and domains from RAID. We see that RoBERTa GPT2 achieves 95+% accuracy on five domains generated by GPT2, but it rarely achieves beyond 60% accuracy on text of the same domain from different models. This detector is open-source, so we know that it was trained exclusively on GPT2 in an open-domain setting. We observe similar trends with RADAR, as it performs uncharacteristically poorly when detecting movie reviews regardless of generative model.

All detectors known to have constrained training data skew heavily towards test data with similar characteristics, leading us to believe that detectors perform better on domains and models seen during training. Some closed-source models such

	None	Paraphrase	Synonym	Misspelling	Homoglyph	Whitespace	Delete Articles
R-L GPT2	56.7	72.9 (+16.2)	79.4 (+22.7)	39.5 (-17.2)	21.3 (-35.4)	40.1 (-16.6)	33.2 (-23.5)
RADAR	70.9	67.3 (-3.6)	67.5 (-3.4)	69.5 (-1.4)	59.3 (-11.6)	66.1 (-4.8)	67.9 (-3.0)
GLTR	62.6	47.2 (-15.4)	31.2 (-31.4)	59.8 (-2.8)	24.3 (-38.3)	45.8 (-16.8)	52.1 (-10.5)
Binoculars	79.6	80.3 (+0.7)	43.5 (-36.1)	78.0 (-1.6)	37.7 (-41.9)	70.1 (-9.5)	74.3 (-5.3)
GPTZero	66.5	64.0 (-2.5)	61.0 (-5.5)	65.1 (-1.4)	66.2 (-0.3)	66.2 (-0.3)	61.0 (-5.5)
Originality	85.0	96.7 (+11.7)	96.5 (+11.5)	78.6 (-6.4)	9.3 (-75.7)	84.9 (-0.1)	71.4 (-13.6)

Table 6: Accuracy Score at FPR=5% for select detectors across different adversarial attacks. Colors indicate an increase, slight increase, slight decrease, and decrease in performance. We see that not all adversarial attacks affect models equally—with some occasionally even improving performance of detectors instead of harming them.

as GPTZero display similar behavior, allowing us to infer what data was used to train them. These findings demonstrate the need for multi-generator training corpora, especially since many publicly available neural detectors focus on only one or two generative models (Guo et al., 2023).

Finding 6: Different detectors are vulnerable to different types of adversarial attacks In Table 6, we see that Binoculars and other metric-based methods degrade as much as 36.1% when a small portion of words are swapped with synonyms. All detectors were sensitive to homoglyph attacks except for GPTZero which sustained only a 0.3% loss under the homoglyph attack while five others dropped an average of 40.6%. Detectors like RADAR that underwent adversarial training, unsurprisingly, were much more robust to adversarial attacks. These detector-dependent differences in vulnerability suggest that attacking an arbitrary detector without prior knowledge of the detector type or training distribution will be difficult. Adversaries may respond by attempting to discover what the detector was trained on—which our findings have shown could be possible—or attacking detectors with repeated queries.

In addition, we see that detector accuracy sometimes increases after an adversarial attack. RoBERTa GPT2, for example, improved after texts were paraphrased with T5 and after words were replaced with BERT-based synonyms. GPT2, RoBERTa, T5, and BERT are contemporaneous models trained on similar data, leading us to believe that detectors benefit from adversarial attacks that inadvertently modify text to be more similar to their training data. Our previous findings on the influence of training data on performance reinforce our hypothesis.

7 Conclusion

As the generation capabilities of language models have continued to increase, accurately and automatically detecting machine-generated text has become an important priority. Detection efforts have even surpassed the bounds of natural language processing research, spurring discussions by social media companies and governments on possibly mandating labels for machine-generated content. Despite the protective intentions of these mandates, our work shows that such regulations would be difficult to enforce even if they were implemented. Detectors are not yet robust enough for widespread deployment or high-stakes use: many detectors we tested are nearly inoperable at low false positive rates, fail to generalize to alternative decoding strategies or repetition penalties, show clear bias towards certain models and domains, and quickly degrade with simple black-box adversarial attacks.

The bulk of our findings may sound bleak, but we did uncover promising signs of improvement. Binoculars, for example, performed impressively well across models even at extremely low false positive rates, Originality achieved high precision in some constrained scenarios, and GPTZero was unusually robust to adversarial attacks. We believe that openly evaluating detectors on large, diverse, shared resources is critical to accelerating progress—and trust—in detection. Evaluating robustness is particularly important for detection, and it only increases in importance and the scale of public deployment grows.

We also need to remember that detection is just one tool for a larger, even more valuable motivation: preventing harm by the mass distribution of text. Detecting machine-generated text was a useful proxy for identifying harmful text for a long time, but language models have improved to the point that generated text is frequently legitimate and not harmful (Schuster et al., 2020). Therefore,

detecting specific harmful elements—like misinformation, hate speech, and abuse—should take precedence over whether or not the text was authored by a machine. Knowing if a text was machine-generated, however, does still offer insights on the types of errors we can expect or the recency of the facts cited within. We hope that our analyses and the RAID dataset are a step toward a future in which AI detection tools are safely integrated into society as a multi-pronged approach to reducing harm. We encourage future work to build on this by including more models, languages, and generation settings in future shared resources.

Limitations

While we attempt to cover a wide variety of domains, models, decoding strategies and adversarial attacks in our dataset, we recognize that there can never be a truly comprehensive dataset for robustness. In particular, our dataset lacks the inclusion of multilingual text in many diverse domains. We release our RAID-extra data to help begin this process but we acknowledge the limited nature of this approach (only having multilingual text in the news domain). We encourage future work to expand on our foundation and use our tools to create truly robust shared benchmarks in many languages.

Furthermore, as the state-of-the-art in language modeling continues to improve, datasets of generated text will naturally obsolesce and will need to be continually maintained with new generations. This creates issues with shared evaluations as detectors will need to be re-run on any new dataset items and any accuracy metrics will have to be updated. While we believe this dataset will continue to be useful for many years, we do acknowledge this limitation and plan to alleviate this by occasionally releasing new updated versions.

Finally, and most importantly, the concept of a public benchmark for out-of-domain robustness is an inherently limited one. As practitioners seek to improve performance on our benchmark they will undoubtedly specialize to the particular aspects of robustness we cover. This will lead to overfitting, even if detectors are not explicitly trained on examples. Such overfitting will result in the reappearance of exactly the problems we wished to alleviate by creating this dataset, namely that detector accuracies are generally over-reported. We trust that this process will, to some extent, be alleviated by regular releases of new versions and keeping a set of hidden test data private. That being said, it does not nullify the utility of the dataset as a resource for profiling robustness of classifiers.

Ethics Statement

Detecting generated text is often accusatory in nature and can frequently result in disciplinary or punitive action taken against the accused party. This can cause significant harm even when detectors are correct, but especially when they are incorrect. This is especially problematic given recent work by Liang et al. (2023c) showing that detectors are biased against non-native English writers. Our results also support this and suggest that the problem of false positives remains unsolved.

For this reason, we are opposed to the use of detectors in any sort of disciplinary or punitive context and it is our view that poorly calibrated detectors cause more harm than they solve. Therefore, until better evaluation standards are widely adopted in the detection community, the use of detectors in this fashion should be discouraged. We intend for our work to be the start of this conversation and look forward to a future where machine-generated detectors are deployed in safe and responsible ways.

Acknowledgements

The authors would like to thank the members of the lab of Chris Callison-Burch for their testing and detailed feedback on the contents of this paper. In addition, we'd like to thank Professor Dan Roth for his early and enthusiastic support of the project.

This research is supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via the HIATUS Program contract #2022-22072200005. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

Aaditya Bhat. 2023. Gpt-wiki-intro.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru,