Sayak Paul and Soumik Rakshit. 2021. arxiv paper abstracts. https://www.kaggle.com/datasets/spsayakpaul/arxiv-paper-abstracts.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refinedweb dataset for falcon llm: Outperforming curated corpora with web data, and web data only.

J. Pu, Z. Sarwar, S. Abdullah, A. Rehman, Y. Kim, P. Bhattacharya, M. Javed, and B. Viswanath. 2023a. Deepfake text detection: Limitations and opportunities. In *2023 IEEE Symposium on Security and Privacy (SP)*, pages 1613–1630, Los Alamitos, CA, USA. IEEE Computer Society.

Xiao Pu, Jingyu Zhang, Xiaochuang Han, Yulia Tsvetkov, and Tianxing He. 2023b. On the zero-shot generalization of machine-generated text detectors.

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Veselin Raychev, Pavol Bielik, and Martin Vechev. 2016. Probabilistic model for code with decision trees. *SIGPLAN Not.*, 51(10):731–747.

Juan Diego Rodriguez, Todd Hay, David Gros, Zain Shamsi, and Ravi Srinivasan. 2022. Cross-domain detection of GPT-2-generated technical text. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1213–1233, Seattle, United States. Association for Computational Linguistics.

Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. 2023. Can ai-generated text be reliably detected?

Areg Mikael Sarvazyan, José Ángel González, Paolo Rosso, and Marc Franco-Salvador. 2023a. Supervised machine-generated text detectors: Family and scale matters. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 14th International Conference of the CLEF Association, CLEF 2023, Thessaloniki, Greece, September 18–21, 2023, Proceedings*, page 121–132, Berlin, Heidelberg. Springer-Verlag.

Areg Mikael Sarvazyan, José Ángel González, Marc Franco-Salvador, Francisco Rangel, Berta Chulvi, and Paolo Rosso. 2023b. Overview of autextification at iberlef 2023: Detection and attribution of machine-generated text in multiple domains.

Dietmar Schabus, Marcin Skowron, and Martin Trapp. 2017. One million posts: A data set of german online discussions. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, page 1241–1244, New York, NY, USA. Association for Computing Machinery.

Tal Schuster, Roei Schuster, Darsh J. Shah, and Regina Barzilay. 2020. The limitations of stylometry for detecting machine-generated fake news. *Computational Linguistics*, 46(2):499–510.

Tatiana Shamardina, Vladislav Mikhailov, Daniil Cherniianskii, Alena Fenogenova, Marat Saidov, Anastasiya Valeeva, Tatiana Shavrina, Ivan Smurov, Elena Tutubalina, and Ekaterina Artemova. 2022. Findings of the the ruatd shared task 2022 on artificial text detection in russian. In *Computational Linguistics and Intellectual Technologies*. RSUH.

Filipo Sharevski, Jennifer Vander Loop, Peter Jachim, Amy Devine, and Emma Pieroni. 2023. Talking abortion (mis) information with chatgpt on tiktok. In *2023 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, pages 594–608. IEEE.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, Miles McCain, Alex Newhouse, Jason Blazakis, Kris McGuffie, and Jasmine Wang. 2019. Release strategies and the social impacts of language models.

Rafael Rivera Soto, Kailin Koch, Aleem Khan, Barry Chen, Marcus Bishop, and Nicholas Andrews. 2024. Few-shot detection of machine-generated text using style representations.

Giovanni Spitale, Nikola Biller-Andorno, and Federico Germani. 2023. Ai model gpt-3 (dis)informs us better than humans. *Science Advances*, 9(26):eadh1850.

Harald Stiff and Fredrik Johansson. 2022. Detecting computer-generated disinformation. *International Journal of Data Science and Analytics*, 13:363–383.

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

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

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, 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, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

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.

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

Michael Völske, Martin Potthast, Shahbaz Syed, and Benno Stein. 2017. TL;DR: Mining Reddit to learn automatic summarization. In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 59–63, Copenhagen, Denmark. Association for Computational Linguistics.

Jian Wang, Shangqing Liu, Xiaofei Xie, and Yi Li. 2023a. Evaluating aigc detectors on code content.

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

Max Weiss. 2019. Deepfake bot submissions to federal public comment websites cannot be distinguished from human submissions. *Technology Science*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing.

Max Wolff. 2020. Attacking neural text detectors.

Kangxi Wu, Liang Pang, Huawei Shen, Xueqi Cheng, and Tat-Seng Chua. 2023. LLMDet: A third party large language models generated text detection tool. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2113–2133, Singapore. Association for Computational Linguistics.

Han Xu, Jie Ren, Pengfei He, Shenglai Zeng, Yingqian Cui, Amy Liu, Hui Liu, and Jiliang Tang. 2023. On the generalization of training-based chatgpt detection methods.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models.

A Experiments on Multilingual and Code Generations (RAID-Extra)

In addition to the main RAID dataset we also release RAID-Extra: a collection of 2.3M generations in three extra challenging domains: Python Code, Czech News, and German News. These extra experiments were not included in the main benchmark as we felt that they were out of scope for most detectors and should not be used as a basis for comparison. Nonetheless, we were still curious to see what sorts of insights they can give us on detector performance.

A.1 Data Generation

Following Macko et al. (2023), multilingual prompts were written in the target language by a native speaker rather than being written in English and explicitly requesting that the model complete the generation in the target language. We found this to be the most effective method to get our generative models to adhere to the target language.

For Python Code generations, we applied an additional post-processing step as, in this domain, generative models had a tendency to write code between sets of triple backtick characters (```) and give natural language explanations of the code outside of the backticks. Thus for this domain and this domain only, we extracted the text between these sets of backticks and discarded all others. This was done to ensure that detectors could not use text descriptions of code for detection and would instead have to rely on the code itself.

A.2 Results

In Table 7 we report the accuracies of our 12 detectors on generations from RAID-extra at 5% FPR. We see an interesting trend, that being the relatively strong performance of metric-based classifiers as compared to neural and commercial detectors. We suspect that metric-based classifiers are particularly well suited for such rare domains as they can be given any generative model to calculate their probabilities.

In Figure 7 we show a heatmap of the performance of our detectors across the extra domains from select models. We see that Binoculars performs decently well when detecting Czech

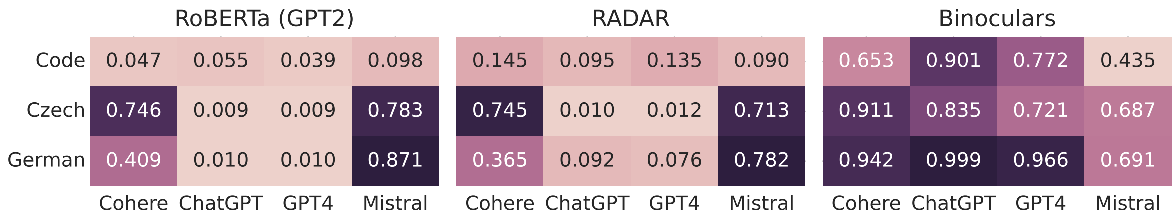

Figure 7: Heatmaps of accuracy for three of our detectors on German News, Python Code, and Czech News generations. We see that metric-based detectors have an edge over neural detectors in their ability to generalize to these unusual domains.

	Code	Czech	German	Total
R-B GPT2	13.4	48.4	39.7	38.2
R-L GPT2	12.7	53.1	48.4	43.5
R-B CGPT	24.0	38.7	51.5	41.1
RADAR	12.9	51.1	53.2	44.7
GLTR	40.7	51.9	68.9	56.7
F-DetectGPT	51.1	55.2	75.5	62.7
LLMDet	17.5	24.0	10.6	17.3
Binoculars	**59.9**	67.0	76.7	**69.6**
GPTZero	33.8	33.6	49.5	39.0
Originality	8.5	69.8	**89.1**	55.8
Winston	24.5	**70.3**	73.8	56.2
ZeroGPT	13.8	49.3	51.7	38.3

Table 7: Accuracy of our 12 detectors at FPR=5% on RAID-extra domains (Python Code, Czech News, and German News). We see that metric based detectors generally perform better than neural detectors.

news articles despite the underlying generative model, Falcon 7B (Almazrouei et al., 2023) being trained with five times as much German data as Czech data (Penedo et al., 2023). This seems to suggest that strong metric-based detectors for low-resource languages can be bootstrapped from highly-multilingual language models. Future work is necessary to understand the optimal setup in such scenarios.

B Fixed FPR Accuracy vs. F1 Score

Throughout our work we report accuracy on machine generated text at a set FPR because we believe it is the most intuitive way of understanding the performance of models in high-risk scenarios (i.e. "What percentage of generations are detected given that we tolerate an x% chance of wrongly accusing someone"). Reporting the more standard F1 score is not only less intuitive but also treats false positives and false negatives as equivalent—which is not the case when dealing with high-risk scenarios or ones where detectors are repeatedly applied to texts from the same author.

In addition, since our dataset has a roughly 40:1 ratio of generated to human-written text, precision scores will artificially favor true positives over false positives as most all examples in the dataset are positive examples. However, in the real world, this ratio is reversed and the majority of texts are human written. Thus precision scores systematically over-represent the capabilities of detectors when used as a metric on a dataset like ours. We hope that our work can help to shed light on this issue and how easy it is to accidentally over-represent the performance of classifiers.

C Per-Domain Threshold Tuning

When tuning the False Positive Rate of classifiers to a specific percentage (in our case 5%), it is important to look not just at total FPR across all human texts, but also at FPR for each individual domain in the dataset. In our work, we ensure that classification thresholds are determined on a per-domain basis, i.e. that the FPRs of every detector on every domain of the data should be 5%. While this undoubtedly adds complexity to the evaluation, it is an important step to ensure that detectors are being evaluated fairly with respect to one another (see Appendix F.2 for details about the threshold searching procedure).

To drive home the importance of this point, in Table 8 we show the FPR of each classifier at a 5% total FPR threshold broken up by domain. As we can see, while the total FPR is consistently 5%, many detectors have particularly acute domain-specific weaknesses: RADAR has a 20.4% FPR on Reviews, GLTR has a 33.4% FPR on Recipes, and Originality has a 13% FPR on Wikipedia.

This asymmetric variation of FPR creates a dampening effect whereby the inclusion of weaker, more obscure domains reduces the accuracy of a classifier on more common domains—ultimately lowering total accuracy in the process.

In order to avoid this issue, we ensure that our thresholds are chosen on a per-domain basis. That