SynSciPass: detecting appropriate uses of scientific text generation

Domenic Rosati
scite / Brooklyn, NY
dom@scite.ai

Abstract

Approaches to machine generated text detection tend to focus on binary classification of human versus machine written text. In the scientific domain where publishers might use these models to examine manuscripts under submission, misclassification has the potential to cause harm to authors. Additionally, authors may appropriately use text generation models such as with the use of assistive technologies like translation tools. In this setting, a binary classification scheme might be used to flag appropriate uses of assistive text generation technology as simply machine generated which is a cause of concern. In our work, we simulate this scenario by presenting a state-of-the-art detector trained on the DAGPap22 with machine translated passages from Scielo and find that the model performs at random. Given this finding, we develop a framework for dataset development that provides a nuanced approach to detecting machine generated text by having labels for the type of technology used such as for translation or paraphrase resulting in the construction of SynSciPass. By training the same model that performed well on DAGPap22 on SynSciPass, we show that not only is the model more robust to domain shifts but also is able to uncover the type of technology used for machine generated text. Despite this, we conclude that current datasets are neither comprehensive nor realistic enough to understand how these models would perform in the wild where manuscript submissions can come from many unknown or novel distributions, how they would perform on scientific full-texts rather than small passages, and what might happen when there is a mix of appropriate and inappropriate uses of natural language generation.

1 Introduction

While estimated submission rates of machine generated scientific papers are still small (Cabanac and Labbé, 2021), contemporary text generation models can generate highly fluent scientific text (Generative Pretrained Transformer et al., 2022) and manuscripts constructed this way could easily be produced en masse potentially introducing an unprecedented threat to scientific publishing and research integrity. Despite this risk, machine generated text in scientific settings have appropriate uses such as with assistive technology like translation, paraphrasing, and speech-to-text (Li et al., 2022). [1] Scientific manuscripts may increasingly use both appropriate and inappropriate text generation technologies. If appropriate uses of text generation cause a manuscript to be flagged or rejected this could harm populations that might already struggle with manuscript writing and submission. For instance, even if publisher's intention is only to guide editors, misclassified manuscripts can unintentionally bias editors decisions. Inspired by Schuster et al. (2020), we ask whether we can develop a method that could adequately distinguish between appropriate and inappropriate uses of text generation by identifying the category of tool being used such as for translation or paraphrase.

	DAGPap22	SynSciPass	Scielo
DAGPap22	99.6	31.4	52.0
SynSciPass	81.3	98.6	65.6
SciBERT	98.3		
TF-IDF	82.0		

Table 1: F1 scores on the DAGPap22, SynSciPass, and Scielo datasets including baselines for DAGPap22 (see Appendix B for model details)

Alarmingly, our study finds that a DeBERTa v3 (He et al., 2021) detector that achieves state-of-the-art performance when finetuned on a dataset designed for detecting generated academic text (DAGPap22 kag (2022)) does poorly on flagging

[1] This is not to say that other malicious applications of these technologies such as disguising plagarism do not exist or that use of poor quality text generation technologies don't introduce problems such as nonsensical phrases (see Cabanac et al. (2021))

Proceedings of the Third Workshop on Scholarly Document Processing, pages 214–222
October 12–17, 2022.

machine generated text under realistic scenarios of appropriate text generation (see Table 1 which shows SciBERT and logistic regression with TF-IDF baselines trained on DAGPap22 as well as DeBERTa v3 trained on DAGPap22 and SynSciPass). Since misflagging a manuscript as machine generated is harmful to the submitting author, we reframe the problem as detecting the type of tool used for generating text so that authors and publishers can have a more nuanced and neutral approach to understanding flagged texts and guiding editorial decisions. We develop a framework to generate academic texts including labels of the type of technology being used resulting in our dataset of synthetic scientific passages (SynSciPass). Section 2 explores how this dataset was constructed and how it could be extended to further improve robustness under domain shifts. In section 3, we show training on SynSciPass results in being able to distinguish the type of technology and how our reframed task helps us move beyond brittle attribution tasks that rely on having access to particular models or the less informative and potentially misleading binary detection task. Finally in section 4, we show that while models trained on our dataset are able to improve robustness under domain shifts for machine generated scientific texts, models for detecting machine generated scientific text are far from ready for safe use by publishers. We provide a roadmap for how to close the gap by focusing on realistic dataset construction that is designed to test detectors ability to robustly generalize across domain shifts.

2 A framework for robust and granular detection datasets

Previous work on detecting machine generated text has focused on attribution of text to particular models (Uchendu et al., 2020; Munir et al., 2021). These approaches have shown the utility of having knowledge of the underlying models for text generation since by having access to those models synthetic corpora can be built for the detection of synthetic text (Liyanage et al., 2022). However those approaches are limited to attributions on specific models trained on particular datasets and do not present a realistic or comprehensive scenario where models may be trained on different datasets or models might be unknown. Our framework improves upon model attribution methods by creating corpora from a variety of distributions with a hierarchy of labels including parent labels based on the type of tool used such as for paraphrasing, translation, or novel text generation. By having access to the type of tool used, we are able to make more sophisticated judgments about machine generated text such as allowing translation and paraphrase as appropriate uses of text generation while requiring more scrutiny for fully generated passages. Our framework consists of (1) proposing a taxonomy of approaches, model families, and models with a variety of pretraining or finetuning datasets that might be used for text generation and (2) sampling machine generated text from each model in the taxonomy so that each text can be labeled with a granular labeling scheme according to (1). By doing so, we hope to be able to attribute generated text not only to specific models but also model families and types of technology. With these more generic labels we are able to determine if models generalize detection across model families or across approaches used like if an unseen model for translation were to be introduced.

2.1 SynSciPass

In order to address these issues we constructed SynSciPass. For our dataset, we theorized three potential sources of machine generated text (1) free-form text generation using generative models like GPT-2 (2) paraphrase models and (3) translation models. While other approaches like speech-to-text or summarization are also likely used in practice, we restricted to the previously mentioned three. We also did not consider the use of multitask models like GPT-3 that are able to use in-context learning to also do paraphrase and translation (Brown et al., 2020) which future work should follow up on to understand if different uses of the same model can be properly distinguished. For each approach, we selected a variety of models from different model families in order to try to synthesize a distribution of text generations that might be found in manuscripts (as have been identified by Cabanac et al. (2021) and Cabanac and Labbé (2021)). These included common services a user might have access to like GPT-2, Spinbot, SCIgen (cf. Cabanac et al. (2021)) and Google translate as well state of the art approaches for each source such as BLOOM for text generation (BigScience, 2022). For each technology type, we also included at least one model that was trained on a distribution of scientific text. The final dataset consisted of

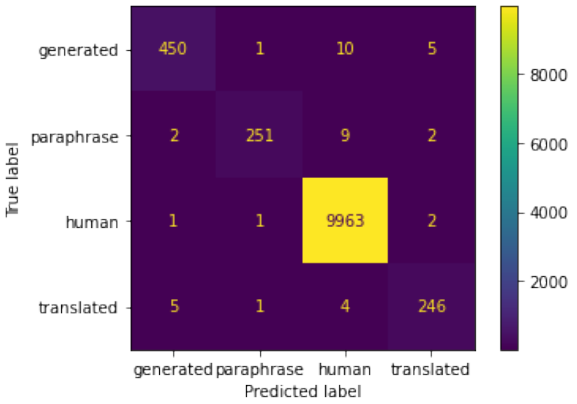

Figure 1: Confusion matrix for multi-class prediction on SciSynPass test set

110,474 passages of which 99,989 (90.5%) were not synthetic to introduce more realistic class imbalance given the estimation that only a few papers per million are machine generated (Cabanac and Labbé, 2021). Please reference Appendix A and B for construction details and Table 4 for full details on dataset construction including the models used to generate data and which model family and technology type they belong to.

3 Reframing synthetic text detection as multi-class classification for understanding appropriate use

Beyond making models more robust through producing a more comprehensive dataset, our framework reframes binary synthetic text detection as multi-class classification that asks not "Is this passage of text synthetic?" but asks "If this passage is synthetic, how was it created?". Given that there are several legitimate uses of text generation tools in the scientific writing process such as using assistive technology, in practice this reframing could allow journal editors to make a more nuanced assessment of potentially synthetic text. For example, if a passage of text was detected as using a translation tool, an editor or submitting author can assess if the translation tool was adequate in conveying meaning or if professional translation services should be employed during revisions. If a paraphrase tool was used, editors can assess whether it might have been used to disguise plagiarism or be the result of a poor quality tool such as Spinbot which is known to introduce non-idiomatic phrases (Cabanac et al., 2021).

Using this approach we trained a multi-class classifier resulting in a micro-averaged F1 score of 99.6% (97.4%, 96.9%, 99.8%, and 96.2% per class F1 for generation, paraphrase, human written, and translation classes respectively) on our held-out test set. To illustrate the models performance we present the confusion matrix in Figure 1 showing that our model does quite well across classes even with the large class imbalance. Additionally as seen in Table 1, the model achieves a F1 score of 81.3% on DAGPap22 which is quite good considering the different domains and notably is about the same as logistic regression with TF-IDF *despite not being trained on the DAGPap22 dataset*. However, this might simply be that DAGPap22 contains a similar underlying distribution. Unfortunately the distribution of DAGPap22 was not known at the time of writing preventing us from providing a nuanced picture of the differences and overlap between DAGPap22 and SynSciPass.

In order to see how our multi-class model might generalize across families of text generation models, we performed an ablation study (Table 2) measuring the performance of DeBERTa v3 trained on SynSciPass as a whole, DeBERTa v3 trained on SynSciPass with texts generated by gpt2-arxiv removed and finally DeBERTa v3 trained on SynSciPass with texts generated by BLOOM removed. F1 scores were reported on model performance on each text generate dataset (see Appendix A and B for details on dataset and model names). In Table 2 we see that removing gpt2-arxiv samples results in a small drop in average performance from the model trained on SynSciPass as a whole (96.5 F1 down from 97.0 F1) indicating that *when we test against a seen model trained on a new dataset detectors may still be effective at detecting the type of technology used*. Interestingly removing gpt2-arxiv samples causes the model to do better on gpt2 than SynSciPass as a whole (94.4 F1 up from 90.7 F1). This indicates that having access to the model on a generic domain might be more important than having access to a model pretrained on a specific distribution as has been studied in Rodriguez et al. (2022). Along these lines we see that removing BLOOM drops performance dramatically on BLOOM from 96.3 to 28.0 F1 score further indicating that having access to underlying models are particularly important and that unseen models may cause detectors to fail. Future work should try to analyze detection models trained to generalize across tools with a wider variety of models including more shifts in underlying pretraining

	BLOOM	distilgpt2	gpt2-arxiv	gpt2	SCIgen	average
SynSciPass	**96.3**	**100.0**	**97.9**	90.7	**100.0**	**97.0**
-gpt2-arxiv	93.5	96.6	**97.9**	**94.4**	**100.0**	96.5
-BLOOM	28.0	97.8	96.8	91.7	**100.0**	82.9

Table 2: Ablation study reporting F1 scores on each text generation subset using DeBERTa v3 trained on SynSciPass as a whole, SynSciPass without the samples generated by gpt2-arxiv and SynSciPass without samples generated by BLOOM. See Appendix A and B for more details.

distribution, a variety of model sizes, different sampling procedures, and introducing unseen models.

4 Out-of-domain Synthetic Passage Detection

Following poor performance of models trained only on DAGPap22 on SynSciPass and vice versa (See Table 1). We wanted to investigate additional domain shifts to understand how robust these models could be in realistic scenarios like seeing new subject domains or new models as this might give us a better picture of how these models might perform in practice. To test robustness over domain shift, we created an additional dataset by sampling human written English passages from Scielo (using Soares et al. (2019)) aligned with human written Spanish passages that were translated back into English. This was done to (1) simulate detection where manuscripts might have used translation and (2) simulate where the underlying distribution from Scielo represents a potential stylistic and disciplinary shift from the Pubmed and arXiv domains which have been seen in SynSciPass.

We sampled 1,000 bilingual English-Spanish human written passages from the Scielo bilingual scientific texts dataset (Soares et al., 2019). We kept the human written English passages labeled as human generated. Then we translated the aligned human written Spanish passages into English using Google translate and labeled these as machine generated. To get a sense of the resulting lexical overlap between the human and machine translated passages, the BLEU score was 40.9 where the overlap between the English passages with themselves is a BLEU score of 100.0. We tested the resulting dataset of 2,000 passages using (1) DeBERTa v3 trained on DAGPap22 only (DAGPap22) (2) DeBERTa v3 pretrained on the Pubmed split of scientific papers and pretrained on the test and train texts from DAGPap22 and then finetuned on DAGPap22 only (DAPT-TAPT) (3) DeBERTa v3 trained on SynSciPass only (SynSciPass) (4) DeBERTa v3 trained on only translations from SynSciPass (SynSciPass (Translation)) (5) DeBERTa v3 trained with potential confounding factors removed (passages generated by google translate and passages generated by a model finetuned on scielo) SynSciPass (SynSciPass (Removed)) (6) DeBERTa v3 trained on both SynSciPass and DAGPap22 (SynSciPass+DAGPap22) (See Appendix B for full training details). In order to compare the results fairly, we should be clear that SynSciPass uses 1 translation model that was finetuned on the same Scielo dataset (Soares et al., 2019) to back translate between English and Spanish as well as Google translate to back translate between English and Chinese so there may be some confounding effect of having samples produced by these models. SynSciPass does not contain any samples from Scielo itself. In order to address this potential confounding factor readers should reference the results from SynSciPass (Removed) where both of those sample sets are removed.

In Table 3, we see that DAGPap22 does quite poorly with an F1 score of 52%, mostly due to poor recall indicating that a state-of-the-art model trained on DAGPap22 would perform as if it's randomly assigning a human generated or machine generated label on translated material mixed in with human written passages from the Scielo domain. Even though this is somewhat expected given that DAGPap22 does not contain information about translations, it is alarming that this is what performance would look like in real life manuscript flagging systems if manuscripts used translators.

A standard approach to improving robustness is pretraining on in-domain and expected task datasets (Gururangan et al., 2020), when utilizing this (DAPT-TAPT) the model does not do too much better (57% F1) than the one trained on DAGPap22 only. Models trained on SynSciPass do improve (up to 66.5 F1 for SynSciPass (Translation)) but do not perform well enough to be considered safe. These results indicate that common approaches for

	AURC ↓	F1 ↑	Precision ↑	Recall ↑
DAGPap22	47.9	52.0	49.6	54.6
DAPT-TAPT	49.3	57.0	50.4	65.7
SynSciPass	51.3	65.6	50.2	94.5
SynSciPass (Translation)	41.3	66.5	50.0	99.1
SynSciPass (Removed)	49.6	66.5	50.1	99.1
SynSciPass+DagPap22	45.6	65.6	50.4	93.9

Table 3: Performance of models presented in Section 4 on an out of distribution translation dataset (Scielo Translations) showing a more than 13 point increase over a model trained on DAGPap22 when using SynSciPass.

machine generated text detection are not robust against shifts in domain and result in dismal performance under a realistic scenario. We also measured the area under risk-coverage (AURC) (El-Yaniv and Wiener, 2010) to present what it might be like if we calibrated our models to only select answers they are most confident in. For AURC, DAGPap22 actually does better than SynSciPass and DAPT-TAPT indicating that it's selective predictions can be made safer. Not surprisingly SynSciPass (Translation) achieves the best AURC of 41.3 indicating that its confidence scores are more meaningful than the others and would perform best at selective prediction, however this requires knowing where the test distribution comes from.

5 Limitations

Given the above results it should be clear that machine generated text detectors in the scientific domain are not very robust to realistic domain shifts. While adding nuance to classifications with the multi-class classifier and providing a more comprehensive dataset enables enhanced robustness, the approach is still sensitive to even small shifts in distribution such as using a known model, google translate in the Scielo case, trained on an unseen dataset, Spanish to English scientific passage translation. The major limitation with our framework is that in order to become more robust we will have to continue to collect more distributions to synthesize from and even as we collect a critical mass of potential distributions of machine generated text, our results are inconclusive as to whether models will continue to be more robust to distributions shifts. Our results with BLOOM removed indicate that generalization to unseen text generation models might not be possible with current approaches. Since machine generated text will continue to approach human-level fluency and new approaches will continue to be developed, it will not be tractable to develop a comprehensive dataset that is representative of the underlying distribution of machine generated text. Additionally, since these models are still sensitive to slight shifts in distribution, we suggest that future work should shift focus to improving robustness of detection on out of domain samples such as with selective prediction or more sample efficient approaches of collecting data to become robust as in Rodriguez et al. (2022). In order to accomplish this, future work should develop a comprehensive suite of tests to evaluate the effects of domain shifts on detectors.

While a multi-class labeling approach might help human evaluators of texts understand why a passage was flagged, this approach should additionally be extended to provide interpretability on why particular passages of texts were flagged. This can be with generating human-like rationales or using methods similar to GLTR (Gehrmann et al., 2019) to assist authors and journal editors in understanding places their manuscript might be improved.

Another limitation of both SynSciPass and DAGPap22 is that they both consist of small passages extracted from scientific texts. Since most manuscripts are submitted as long texts, we are not sure how these results would apply to realistic scientific full-texts, especially when those full-texts include tables, figures, and other non-textual items. While Rodriguez et al. (2022) does provide approaches to address this with passage-based models, future work should still aim to construct datasets that are more realistic and close to the task by providing full-text scientific documents that include layout, figures, and tables. Finally, these datasets should aim to match the extreme class imbalance that has been observed in real world distribution of machine generated texts identified in Cabanac and Labbé (2021).

6 Related Works

Jawahar et al. (2020) outlines many recent approaches to detecting machine generated text in a variety of domains. The closest to our approach is attribution models that attempt to use a stylometric approach for uncovering the authorship of a text where the author is a particular model or particular model using a particular dataset (Jones et al., 2022; Munir et al., 2021). Our approach is unique in that it focuses on attribution of general classes of tools such as translation, paraphrase, and generation rather than specific models.

While we agree in principal with criticisms of the stylometric approaches that seek to center the veracity and coherence of texts (Dou et al., 2022; Schuster et al., 2019), as text generation models improve, the factuality and fluency gap between machine and human generated text will get smaller and smaller and methods that utilize veracity and coherence will no longer work [2]. Additionally, many humans make errors and write poor quality manuscripts so we do not feel like this is a good criterion for detecting machine generated texts but should be clearly separated as an equally important but orthogonal task of understanding the quality of scientific texts. Similarly, we are skeptical of approaches like MAUVE (Pillutla et al., 2021) that rely on distributional artifacts produced by machine generated texts since as text generation models mature the gap between human and machine distributions will also close.

Rodriguez et al. (2022) is the closest to our work in examining the effects of domain shifts in detecting machine generated scientific texts showing that detectors do not generalize well when subject domains shift from physics to biomedicine. While they show that generating even a small number of samples in another domain improves detection, their work is limited to only GPT-2 making their findings reliant on having access to the underlying models. Data augmentation like we used is a common strategy shown to improve the robustness of models in NLP (Wang et al., 2022) and is common for examining text generation model attribution in detecting machine generated text since we have access to the underlying text generation models during analysis (Uchendu et al., 2020). Finally, recent work has examined the robustness of these models (Gagiano et al., 2021; Wolff, 2020) but these methods focus on robustness to adversarial attacks such as homoglyphs and misspellings rather than robustness to domain shifts and generalization to unseen models which is studied in this work and which we understand as area with the most promise for both understanding and improving detectors.

[2] Clark et al. (2021) find that humans already cannot reliably distinguish between human and machine generated text produced by GPT-3)

7 Ethics Statement

The results in this paper should make it clear that at this point machine generated text detectors should not be used in production because they do not perform well on distribution shifts and their performance on realistic full-text scientific manuscripts is currently unknown. Further development is needed on both interpretable and robust detection methods as well as better datasets that are both realistic (such as including full-texts rather than passages) and varied (including comprehensive samples across scientific disciplines). Because erroneously detecting a manuscript as machine generated is a high harm activity, future work should continue more nuanced harm-reduction approaches to synthetic paper detection like the ones introduced in this paper.

8 Data Availability

The final constructed dataset, SynSciPass, source code, and models are available at https://github.com/domenicrosati/synscipass.

9 Conclusion

Given our findings, we envision future work along three lines (1) developing machine generated text detectors that are robust across domain shifts and developing realistic datasets that test this robustness comprehensively (2) developing methods of interpretability that help editorial teams detect and manage the use of both appropriate and inappropriate use of text generation models (3) discussion about the safe and ethical application of these technologies and the potential harm involved in their deployment when use cases such as assistive technology are not considered.

We introduced a framework for collecting datasets to improve the robustness and interpretability of detecting machine generated text in the scientific domain. By developing a comprehensive dataset, SynSciPass, we were able to show that

models trained on it were not only more robust under domain shifts but also that those models were able to detect the generic type of text generation technology such as for translation, paraphrase, or novel generations which could help understand if a passage was generated by appropriate or inappropriate means. Despite these findings, our work has also shown that current models, including our own, do not perform well in realistic scenarios that change the distribution of text seen. Because of this lack of robustness, we suggest that future work concentrate on formulating both datasets and approaches that comprehensively test machine generated text detectors in a wide variety of realistic and unseen scenarios.

References

2022. Detecting generated scientific papers.

BigScience. 2022. Bigscience large open-science open-naccess multilingual language model.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. ArXiv:2005.14165 [cs].

Guillaume Cabanac and C. Labbé. 2021. Prevalence of nonsensical algorithmically generated papers in the scientific literature. *J. Assoc. Inf. Sci. Technol.*

Guillaume Cabanac, Cyril Labbé, and Alexander Magazinov. 2021. Tortured phrases: A dubious writing style emerging in science. Evidence of critical issues affecting established journals. ArXiv:2107.06751 [cs].

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. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7282–7296, Online. Association for Computational Linguistics.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A Discourse-Aware Attention Model for Abstractive Summarization of Long Documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 615–621, New Orleans, Louisiana. Association for Computational Linguistics.

Yao Dou, Maxwell Forbes, Rik Koncel-Kedziorski, Noah A. Smith, and Yejin Choi. 2022. Is GPT-3 Text Indistinguishable from Human Text? Scarecrow: A Framework for Scrutinizing Machine Text. In *ACL*.

Ran El-Yaniv and Yair Wiener. 2010. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(53):1605–1641.

Rinaldo Gagiano, Maria Kim, Xiuzhen Zhang, and J. Biggs. 2021. Robustness Analysis of Grover for Machine-Generated News Detection. In *ALTA*.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander M. Rush. 2019. GLTR: Statistical Detection and Visualization of Generated Text. In *ACL*.

Gpt Generative Pretrained Transformer, Almira Osmanovic Thunström, and Steinn Steingrimsson. 2022. Can GPT-3 write an academic paper on itself, with minimal human input?

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. ArXiv:2111.09543 [cs].

Ganesh Jawahar, Muhammad Abdul-Mageed, and L. Lakshmanan. 2020. Automatic Detection of Machine Generated Text: A Critical Survey. In *COLING*.

Keenan I. Jones, Jason R. C. Nurse, and Shujun Li. 2022. Are You Robert or RoBERTa? Deceiving Online Authorship Attribution Models Using Neural Text Generators. *ArXiv*.

Pengcheng Li, Wei Lu, and Qikai Cheng. 2022. Generating a related work section for scientific papers: an optimized approach with adopting problem and method information. *Scientometrics*, 127(8):4397–4417.

Vijini Liyanage, Davide Buscaldi, and Adeline Nazarenko. 2022. A Benchmark Corpus for the Detection of Automatically Generated Text in Academic Publications. ArXiv:2202.02013 [cs].

Shaoor Munir, Brishna Batool, Zubair Shafiq, P. Srinivasan, and Fareed Zaffar. 2021. Through the Looking Glass: Learning to Attribute Synthetic Text Generated by Language Models. In *EACL*.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, S. Welleck, Yejin Choi, and Z. Harchaoui. 2021. MAUVE: Measuring the Gap Between Neural Text and Human Text using Divergence Frontiers. In *NeurIPS*.

Juan 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.

Tal Schuster, R. Schuster, Darsh J. Shah, and R. Barzilay. 2019. Are We Safe Yet? The Limitations of Distributional Features for Fake News Detection. *undefined*.

Tal Schuster, R. Schuster, Darsh J. Shah, and R. Barzilay. 2020. The Limitations of Stylometry for Detecting Machine-Generated Fake News. *Computational Linguistics*.

Felipe Soares, Viviane Pereira Moreira, and Karin Becker. 2019. A Large Parallel Corpus of Full-Text Scientific Articles. ArXiv:1905.01852 [cs].

Adaku Uchendu, Thai Le, Kai Shu, and Dongwon Lee. 2020. Authorship Attribution for Neural Text Generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8384–8395, Online. Association for Computational Linguistics.

Xuezhi Wang, Haohan Wang, and Diyi Yang. 2022. Measure and Improve Robustness in NLP Models: A Survey. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4569–4586, Seattle, United States. Association for Computational Linguistics.

Max Wolff. 2020. Attacking Neural Text Detectors. *ArXiv*.

A Construction of SynSciPass

SynSciPass was constructed using 100,000 passages that were randomly sampled from the scientific papers dataset (Cohan et al., 2018). Each passage was between 2 and 10 sentences randomly sampled from the full-text of single publication from both the arXiv and Pubmed training splits with a resulting mean token length of 142 tokens roughly matching the 140 token mean of the DAGPap22 dataset. From these passages 1,000 items were randomly sampled (with replacement) for each model found in Table 4. Passages that were constructed using BLOOM and GPT-2 proceeded following the approach of (Liyanage et al., 2022) where the first sentence of the real passage was used as the prompt to construct the synthetic passage, subsequent generations were used to re-prompt the model to sample passages between 2 and 10 sentences. The first sentence from the real passage was then removed. For these models greedy decoding with a temperature of 1.0 was used. For SCIgen, 1,000 papers were generated and then a random passage of between 2 and 10 sentences was extracted from each one. For the paraphrase models, a randomly sampled passage from the human written passages were sent through a paraphrase tool. For the translation models, a human written passage was sent through the translation tool into a target language and then back translated into english. For all models generations, text similarity was measured between the original passage and the synthesized example, if the sample was more than 10% similar it was not kept. This does simplify the problem and make the data less realistic as it removes synthetic passages that have a high lexical overlap with reference passages which might be common with inapporiate uses such as masking plagarism. The final dataset consisted of 110,474 passages of which 99,989 (90.5%) were human written. This was done to try to match the extreme class imbalance that has been observed on synthetic scientific papers in the wild (Cabanac and Labbé, 2021). The final dataset was split by 80%/10%/10% into train, validation, and test sets.

B Training details on models used

For this work, all of our classification models were trained by finetuneing DeBERTa v3 large (He et al., 2021) using the following hyperparameters: adamW optimizer, learning rate of 6e-6, batch size of 8, weight decay of 0.01 with warmup steps of 50. All classification models were trained for 3 epochs. For the domain adaptive pretraining (DAPT) model, we further pretrained using the parameters mentioned above with a masked language modeling objective on the Pubmed train split from the scientific papers dataset (Cohan et al., 2018) using 128 token chunks for 5 epochs. For the task adaptive pretraining (TAPT) model, we used the same approach with 5 epochs on the DAGPap22 dataset. Details of the SciBERT and logistic regression TF-

type	model family	model	passages
generate	bloom	bloom	1073
	gpt2	GPT-2-arxiv_generate	998
		distilgpt2	998
		gpt2-medium	998
	SCIgen	SCIgen	822
paraphrase	pegasus	pegasus-xsum-finetuned-paws*	1000
		pegasus-xsum-finetuned-paws-parasci*	1000
	spinbot	spinbot	990
real	real	real	99064
translate	google_translate	google_translate	901
	opus	opus-es-en	794
		opus-es-en-scielo*	901

Table 4: Approaches used for data augmentation and number of passages generated. Models with an asterisk were trained by the authors. Spinbot, SCIgen, and google translate are the names of the services used available online. The rest of the models are or will be made available on the huggingface repository under those names.

IDF model baselines were not made available at the time of writing this paper.