Nuanced Multi-class Detection of Machine-Generated Scientific Text

Shiyuan Zhang[1], Yubin Ge[1], Xiaofeng Liu[2],
[1]University of Illinois Urbana-Champaign, [2]Yale University
{sz54, yubinge2}@illinois.edu
xiaofeng.liu@yale.edu

Abstract

Recent advancements in large language models (LLMs) have demonstrated their capacity to produce coherent scientific text, often indistinguishable from human-authored content. However, this raises significant concerns regarding the potential misuse of such techniques, posing threats to research advancement across various domains. In this study, we focus on nuanced detection of machine-generated scientific text and build a new multi-domain dataset for this task. Instead of treating the detection as binary classification task, as in previous work, we additionally consider the classification of diverse practical usages of LLMs, including paraphrasing, summarization, and title-based generation. Additionally, we introduce a novel baseline model integrating contrastive learning, encouraging the model to discern similar text more effectively. Experimental results underscore the efficacy of our proposed method compared to prior baselines, supplemented by an analysis of domain generalization conducted on our dataset.

1 Introduction

Language models, particularly large language models, have brought significant advancements to various tasks. These models typically undergo pretraining on extensive text corpora, endowing them with unprecedented accuracy in predicting the next token given some context (Ge et al., 2023a). Based on them, LM-powered writing tools have gained widespread adoption and substantial interest. Notably in the scientific domain, advanced LMs exhibit remarkable proficiency in generating scientifically fluent text (Transformer et al., 2022), and have proven useful in various associated tasks such as scientific document summarization (Cachola et al., 2020; Meng et al., 2021), citation text generation (Xing et al., 2020; Ge et al., 2021), keyphrase extraction (Kontoulis et al., 2021; Glazkova and Morozov, 2023), and peer review synthesis (Wang et al., 2020; Yuan et al., 2022). Nonetheless, concerns regarding the misuse of these tools have been raised (Cabanac et al., 2021), underscoring the critical importance of detecting machine-generated scientific text to mitigate the proliferation of counterfeit scientific publications and citations (Else, 2021).

Various endeavors have been undertaken to promote the automatic detection of machine-generated scientific text. Conventionally, prior research has framed this task as binary classification, wherein models are trained to predict whether scientific texts are "fake" (likely generated) or "real," i.e., human-authored (Kashnitsky et al., 2022). Furthermore, previous study demonstrates that distinguishing the specific technologies employed in generating scientific text can enhance robustness against domain shifts, thereby suggesting a promising direction for further research in this domain (Rosati, 2022).

Drawing from the above inspiration, this paper delves into the multi-class classification for the nuanced detection of machine-generated scientific text. Specifically, we construct a new dataset by prompting ChatGPT to generate paper abstracts through various practical usages, covering paraphrasing, summarization, and generation from paper titles. Notably, each paper in our dataset is annotated with a domain label, facilitating exploration into domain generalization or adaptation. Additionally, we introduce a novel baseline model leveraging contrastive learning to encourage discernment between similar paper abstracts with differing labels. Comparative analysis against prior baseline models on our dataset underscores the superiority of our proposed baseline, and we also show performing domain generalization on our dataset.

Our contributions are delineated as follows:

- To the best of our knowledge, we present the

first publicly available dataset[1], spanning diverse fields of study, for nuanced multi-class detection of machine-generated scientific text.

- We introduce a new baseline model based on contrastive learning to encourage the model to distinguish similar scientific texts.
- Through experiments, we empirically demonstrate the superior effectiveness of our approach compared to previous baselines, and show domain generalization on our dataset.

2 Related Work

Most previous studies on understanding machine-generated text have approached it as a binary classification task, where the model must differentiate between text that is entirely human-written and text generated by a machine (Dugan et al., 2023). Despite advancements in detecting machine-generated texts, datasets specifically for scientific literature remain scarce. For instance, a previous study (Kashnitsky et al., 2022) curated a dataset containing summarized, and paraphrased paper abstracts and excerpts, alongside text generated by LLMs like GPT-3 (Brown et al., 2020). However, this dataset is limited in size and lacks coverage across diverse scientific fields. Another research (Liyanage et al., 2022) proposed an alternative strategy, generating papers using GPT-2 (Radford et al., 2019) and Arxiv-NLP4. This dataset, while larger, still focuses mainly on text generation and lacks sufficient annotations for more nuanced tasks. Additionally, another benchmark dataset (Mosca et al., 2023) was compiled, containing both human-written and machine-generated scientific papers from various LLMs including GPT-2 (Radford et al., 2019), GPT-3 (Brown et al., 2020), ChatGPT (OpenAI, 2022), and Galactica (Taylor et al., 2022). However, these datasets are predominantly designed towards binary classification, overlooking the different practical approaches employed in generating scientific texts, such as paraphrasing or summarization. Such a nuanced detection has been shown to enhance detector robustness against domain shifts (Rosati, 2022). Furthermore, the absence of field-of-study labels in these datasets restricts their application in domain generalization research, a critical aspect for robust scientific text detection across various domains.

[1]Our code and dataset are made public at: https://github.com/SeanZh30/ScientificText_Detection.

The detection of automatically generated scientific texts represents an emerging subfield of research with limited extant literature. Traditionally, approaches have relied on hand-crafted features (Amancio, 2015; Williams and Giles, 2015), grammar-based detectors (Cabanac and Labbé, 2021), and nearest neighbor classifiers (Nguyen and Labbé, 2016) to address this challenge. However, with the advent of large language models, recent studies have demonstrated promising outcomes in detection leveraging pre-trained models such as SciBERT (Beltagy et al., 2019) and other variants (Glazkova and Glazkov, 2022; Liyanage et al., 2022; Mosca et al., 2023).

Current research trends indicate that improving the robustness of detection models against domain shifts with diverse data generation techniques and richer annotations is important. Moreover, addressing the limited diversity and scope of existing datasets, particularly in terms of scientific fields and generation techniques, will be vital for advancing the detection of machine-generated scientific texts. Future work should prioritize the creation of well-annotated, cross-disciplinary datasets that encompass a variety of text generation methods to improve model generalization ability and applicability across domains.

3 Dataset

Motivated by prior research, we build our dataset according to the following principles:

- We focus on the *abstracts* of academic papers and employ a widely utilized LLM, i.e., ChatGPT [2], for the generation of scientific text.
- We consider diverse practical usages of the LLM in scientific text generation, categorizing instances into nuanced labels for multi-class classification based on generation methods.
- Each data instance is annotated with a field-of-study label, enabling analysis pertinent to the domains of scientific texts.

3.1 Data Preparation

We first collect scientific papers from Semantic Scholar Open Research Corpus (S2ORC) (Lo et al., 2019), which is currently the largest collection of machine-readable academic text dataset and covers multiple domains. We randomly sample papers

[2]We use gpt-3.5-turbo specifically

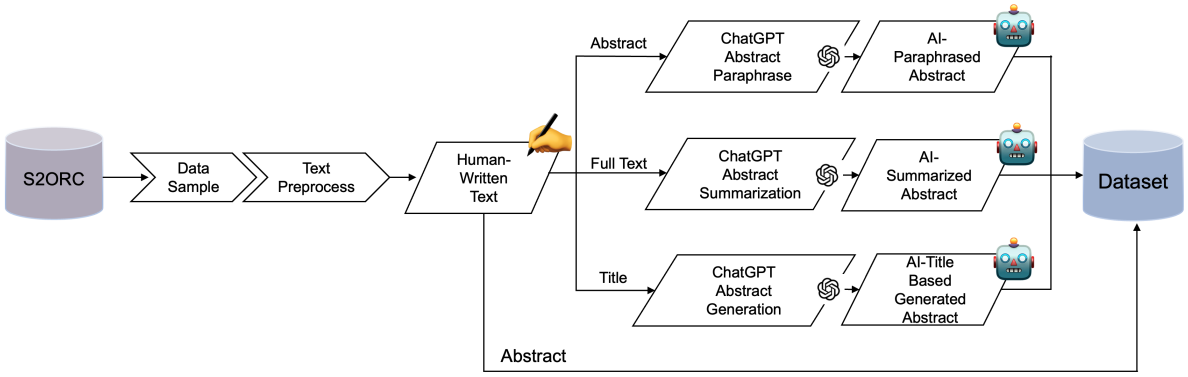

Figure 1: Overview of our dataset construction pipeline.

from the raw S2ORC and further clean the sampled data by removing the noisy samples that satisfy any of the following criteria:

- Data lacking field-of-study labels annotated by S2ORC.
- Data that miss titles, abstracts, or any textual component.
- Abstracts containing fewer than 50 words.
- Data with text encoded rather than in standard text format.
- Non-English data.

Finally, we retain data from the popular fields of *Medicine*, *Computer Science*, *Physics*, *Engineering*, and *Biology*. These data are further be sampled for different generation approaches to obtain machine-generated scientific texts.

3.2 Fake Abstract Generation

Previous research on AI-generated text datasets has often relied on translators, moderately sized language generation models (e.g., distilGPT-2 and GPT-2), or models specifically designed to generate scientific or nonsensical texts (e.g., GPT-2-arxiv, SCIgen) (Rosati, 2022). However, in real-world applications, text generation is increasingly dominated by LLMs used at the application level. Therefore, this article focuses on exploring the use of ChatGPT, a more practical LLM widely employed in real-world applications.

We utilize ChatGPT to generate synthetic abstracts, employing various generation approaches designed to closely simulate real-world scenarios:

- **Abstract Paraphrase** (Kashnitsky et al., 2022): This approach entails providing a human-written abstract as a prompt, prompting the LLM to paraphrase it while preserving the academic style. The resulting paraphrased abstracts are categorized as **paraphrase**.
- **Introduction Summarization** (Cachola et al., 2020; Meng et al., 2021; Ge et al., 2023b): We use the LLM to produce a formal and academic abstract based on the provided introduction section of a scientific paper. Introduction sections exceeding length constraints are truncated. The resultant abstracts are labeled as **summarization**.
- **Title-Based Abstract Generation** (Wang et al., 2019; Mosca et al., 2023): Inspired by prior research leveraging paper titles to generate paper abstracts, we prompt the LLM to generate abstracts based solely on provided paper titles. Correspondingly, the produced abstracts are categorized as **generation**.

3.3 Prompt Design

Prompting is the main tool for interacting with large language models and can be used to inform the model of task instructions (Brown et al., 2020). Meanwhile, it has been widely used in assisting scientific writing, and so we design the prompts based on different practical usages introduced in Section 3.2. Specifically, we take a part of the original human-written texts as partial input and instruct LLM to complete abstract generation. In this section, we present the prompt templates used for querying ChatGPT. Each approach corresponds to a specific method of generating synthetic abstracts based on human-written scientific articles.

Abstract Paraphrase We use a prompt designed to rephrase the original abstract while preserving its core topics and structure. The source document here is the original human-written abstract.

Abstract Paraphrase Prompt

Read the abstract of the research paper provided below. Paraphrase the abstract into a single paragraph, maintaining a formal and academic tone. The abstract is as follows:

{source document}

Introduction Summarization This type of prompt is designed to condense the full article into a shorter abstract, focusing on the essential elements of the research. Since the input source document will be the full text, the input text will be truncated at the maximum input tokens.

Introduction Summarization Prompt

Read the introduction and the full text of this research paper. Summarize the paper and write an abstract in one paragraph and in a formal and academic style. Do not include any prefixes and only keep the text of the abstract. Here is the full text:

{source document}

Title-Based Abstract Generation We use a prompt that generates an abstract based solely on the provided article title, simulating how an abstract might be constructed from key points inferred from the title alone. The source document used for input only contains human-written titles.

Title-Based Abstract Generation Prompt

Write an abstract in one paragraph and in a formal and academic style according to this title. Do not include any prefix and only keep the text of the abstract. Here is the title:

{source document}

3.4 Dataset Construction and Statistics

We combine all machine-generated scientific abstracts with the remaining human-written abstracts to form our dataset. We also perform a processing step for the machine-generated text. The underlying reason is that ChatGPT tends to exhibit specific patterns or flaws when generating text. For instance, even when explicitly instructed in the prompt to exclude prefixes, such as "Do not generate any prefixes in the response, only include the generated abstract," some outputs still contain prefixes like "Abstract:" or "Abstract: \n". We preprocess the input data by removing these prefixes, ensuring the subsequent predictions are closer to real-world scenarios. We finally perform the train-test split and provide the statistics of our dataset in Table 1.

Statistics	Train	Test	Validation
Avg num. of words	169.39	168.58	167.85
Min num. of words	50	50	50
Max num. of words	8574	2107	1425
Avg num. of sentences	5.93	5.88	5.87
Min num. of sentences	2	2	2
Max num. of sentences	387	60	122
Num. of instances	39,706	5,200	5,200

Table 1: Dataset statistics

The domain distribution whose proportion exceeds 0.6%. is shown in 2 and more detailed in Appendix Sec. A. The composition covers a range of scientific disciplines. Notably, the major components such as *Medicine*, *Computer Science*, *Physics*, *Engineering*, and *Biology* each contain more than 8,000 instances. Importantly, we ensure non-overlap among the source papers; for instance, a paper for paraphrasing cannot be chosen for summarization during the generation process.

3.5 Data Generation Example

We show one example generated via abstract paraphrase in Table 2 and provide additional examples generated by other approach in Appendix Sec. B. The input is used as source document in prompt mentioned in Section 3.3.

4 Method

In this section, we describe the methodology in our study for distinguishing between human-written and machine-generated texts. We utilize advanced pre-trained baseline models such as SciBERT (Beltagy et al., 2019), RoBERTa (Liu et al., 2019), and

Input:

Abstract: Towards building intelligent dialogue agents, there has been a growing interest in introducing explicit personas in generation models. However, with limited persona-based dialogue data at hand, it may be difficult to train a dialogue generation model well. We point out that the data challenges of this generation task lie in two aspects: first, it is expensive to scale up current persona-based dialogue datasets; second, each data sample in this task is more complex to learn with than conventional dialogue data. To alleviate the above data issues, we propose a data manipulation method, which is model-agnostic to be packed with any persona-based dialogue generation model to improve its performance. The original training samples will first be distilled and thus expected to be fitted more easily. Next, we show various effective ways that can diversify such easier distilled data. A given base model will then be trained via the constructed data curricula, i.e. first on augmented distilled samples and then on original ones. Experiments illustrate the superiority of our method with two strong base dialogue models (Transformer encoder-decoder and GPT2).

Output:

Abstract: In the pursuit of developing intelligent dialogue agents, a notable trend has emerged emphasizing the incorporation of explicit personas within generation models. However, the scarcity of persona-driven dialogue data poses a significant challenge to effectively training dialogue generation models. This challenge is underscored by two primary factors: the expense associated with expanding current persona-centric dialogue datasets and the heightened complexity of individual data samples compared to conventional dialogue datasets. To address these data-related hurdles, we introduce a model-agnostic data manipulation technique designed to enhance the performance of persona-based dialogue generation models. By initially distilling the original training samples to facilitate easier fitting, followed by implementing diversification strategies to enhance the distilled data, our approach aims to optimize the learning process. Through a structured training regimen involving augmented distilled samples followed by original ones, we demonstrate the effectiveness of our method using established base dialogue models such as the Transformer encoder-decoder and GPT2 in empirical experiments.

Table 2: An example of a generated instance via abstract paraphrase.

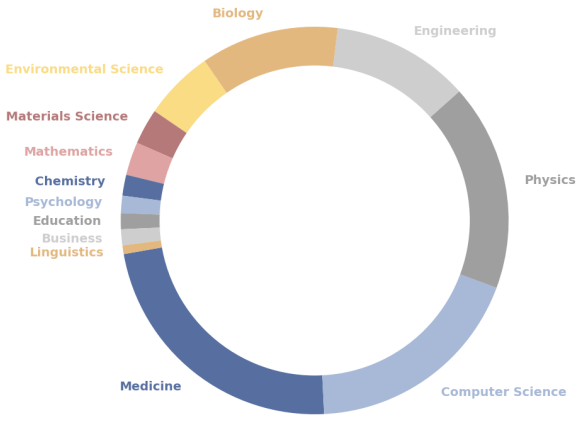

Figure 2: Representative domains of dataset.

DeBERTa (He et al., 2020), known for their efficacy and accuracy in similar classification tasks. Additionally, we incorporate contrastive learning (Radford et al., 2021; Yang et al., 2023; Bo et al., 2024) to enhance our model's performance, focusing on refining representations to better identify textual differences.

4.1 Backbone Models

Prior studies have demonstrated significant success in binary classification for this task through the fine-tuning of various BERT-related pre-trained models (Kashnitsky et al., 2022; Rosati, 2022; Mosca et al., 2023), including SciBERT (Beltagy et al., 2019), RoBERTa (Liu et al., 2019), and DeBERTa (He et al., 2020). Thus, we employ these models as the backbone encoders to encode input texts, denoted as $f_{\text{enc}}(\cdot)$. Subsequently, the final hidden state corresponding to the special token `[CLS]` serves as the aggregated sequence representation for an input text x_i, denoted as $h_i = f_{\text{enc}}(x_i)$. Following standard practice, we augment this representation with

an MLP for multi-class classification, employing the cross-entropy function to compute the loss:

$$\hat{y}_i = \text{softmax}(\text{MLP}(h_i))$$
$$\mathcal{L}_{\text{cls}} = \sum_i \text{CrossEntropy}(\hat{y}_i, y_i),$$

where $\hat{y}_i$ is the prediction and y_i is the target label.

4.2 Contrastive Learning

Given the widespread adoption of contrastive learning across various domains and tasks for proficient representation learning (Yang et al., 2023), we incorporate it into our classifier to enhance the discrimination between human-written and machine-generated text. Our objective is to group similar texts with the same label while segregating those with differing labels. Specifically, for a given text x_i, we identify its positive sample, denoted as x_i^+, as those share the same target label and exhibit similarity to x_i. Conversely, the negative sample x_i^- is recognized as similar texts to x_i but bears a different target label. We calculate text similarity using cosine similarity between the tf-idf representations of texts. Subsequently, drawing from (Chopra et al., 2005), we augment our model with an additional contrastive learning loss, defined as follows:

$$\mathcal{L}_{\text{con}} = \sum_i \|f_{\text{enc}}(x_i) - f_{\text{enc}}(x_i^+)\|_2^2 + \max(0, \epsilon - \|f_{\text{enc}}(x_i) - f_{\text{enc}}(x_i^-)\|_2^2),$$

where ϵ is the margin set to separate negative samples and is set to 0.1.

Finally, the objective function is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \alpha \cdot \mathcal{L}_{\text{con}},$$

where α is the hyperparameter to balance the two losses, and we set it to 0.5.

5 Experiments

To address the key challenges in detecting machine-generated scientific text and explore the quality of our dataset, we bring up three research questions:

Q1: Performance and Contrastive Learning. Does the baseline model show relatively high-quality performance on our dataset and does the integration of contrastive learning enhance the detection capabilities of baseline models?

Q2: Nuanced Dataset Classification. What is the significance of nuanced classification of datasets in identifying real-world scenarios for machine-generated scientific text?

Model	Performance	
	Accuracy (%)	F1 (%)
Baseline Models		
SciBERT_{512}	**96.98**	96.93
SciBERT_{256}	96.31	96.24
SciBERT_{128}	96.03	96.10
RoBERTa	95.78	95.81
DeBERTa	96.97	**97.02**
Contrastive Models		
SciBERT_{512} + contrastive	**97.60**	**97.58**
RoBERTa + contrastive	96.50	97.01
DeBERTa + contrastive	97.01	97.38

Table 3: Comparison of baseline and contrastive models on the dataset.

Q3: Domain Generalization. Can the models have a well performance on generalizing across different scientific domains on our dataset?

For those three research questions, we design our experiments to evaluate the performance of fine-tuned baseline models in detecting machine-generated scientific text and to explore the effect of contrastive learning and the impact of different input lengths on model accuracy.

5.1 Implementation Details

We follow one previous work (Glazkova and Glazkov, 2022) to use pre-trained models from HuggingFace (Wolf et al., 2020) and adopt their configurations. Specifically, we fine-tune SciBERT, RoBERTa, and DeBERTa on our dataset for three epochs. To maintain consistency across experiments, we set the maximum sequence length of input for all models to 512. Additionally, we vary the input length for SciBERT to 128 and 256 for testing purposes. Each model input will automatically read tokens within the length limit. As for the hyperparameters, we set the learning rate at 2e-5, AdamW as the optimizer, and 16 as the batch size. The mode of the three classifications is taken as a final output.

5.2 Q1: Performance and Contrastive Learning

We evaluate model performance using accuracy and Macro F1 as metrics, shown in Table 3. Our findings indicate that integrating contrastive learning enhances the performance of all baseline models, underscoring the effectiveness of our proposed approach in fostering effective discrimination of similar scientific texts. We attribute this improve-

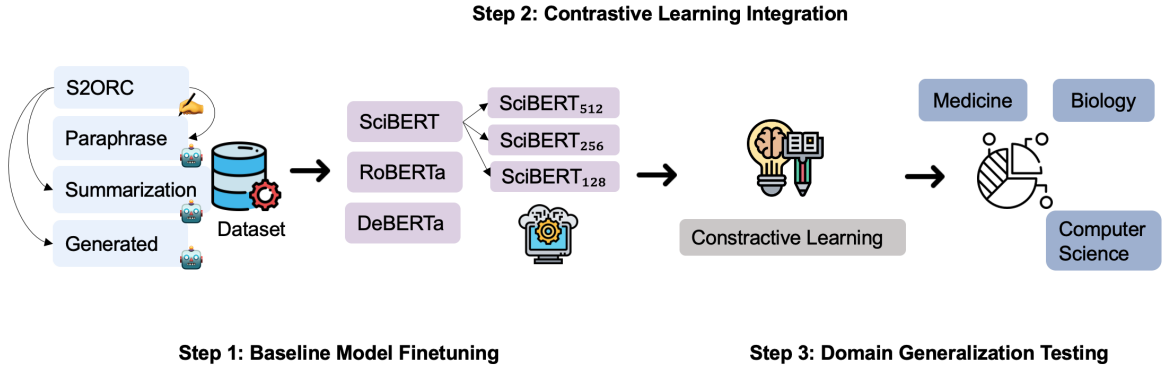

Figure 3: Overview framework of experiment.

ment to contrastive learning's property that can encourage the model to focus on subtle differences between similar texts, which in turn improves its ability to distinguish borderline cases. Moreover, the discrepancies among these pre-trained models are marginal, aligning with previous research findings (Glazkova and Glazkov, 2022). Additionally, we investigate the impact of input length on SciBERT, observing that increasing input length enhances both prediction accuracy and F1 score, demonstrating longer input sequences allow the model to capture more context, which is particularly important for distinguishing between nuanced variations in scientific articles.

5.3 Q2: Nuanced Dataset Classification

Examining the confusion matrix for SciBERT_{512} with contrastive learning, as depicted in Figure 4, we observe high accuracy in discerning between human-written and machine-generated data. This could be attributed to ChatGPT adhering to consistent language patterns when generating synthetic scientific text. These patterns, minimally influenced by variations in the usage of LLM, enable the model to identify the differences that distinguish human-authored articles from machine-generated texts. Nevertheless, some degree of imprecision persists in classifying synthetic article abstracts, indicating potential areas for future improvement. Further analysis of the confusion matrix reveals the importance of nuanced classifications, particularly in some real-world scenarios where the definition of a "fake article" varies. Some may consider machine-assisted writing also as "valid articles". For example, in certain situations, paraphrasing or summarizing articles might be permissible. These differing classifications can affect how well models distinguish between genuine and synthetic scientific content.

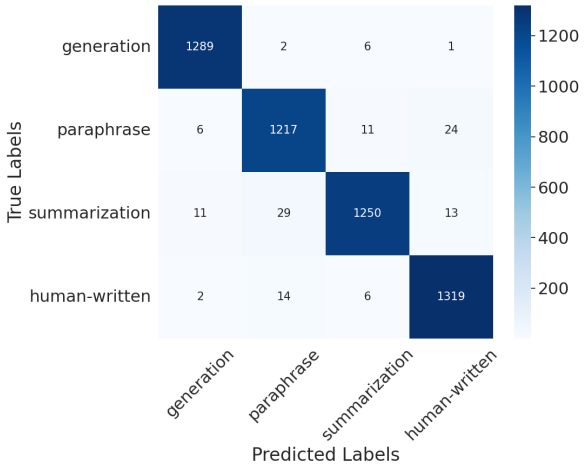

Figure 4: Confusion matrix for SciBERT_{512} with contrastive learning.

5.4 Q3: Domain Generalization

We test domain generalization using SciBERT_{512} by leaving out one domain for testing each time and training the model on the remaining domains.

Training Domain	Test Domain	Accuracy (%)	F1 (%)
Bio, Eng, Med, Phy	CS	96.21	96.19
CS, Eng, Med, Phy	Bio	94.85	94.75
Bio, CS, Eng, Med	Phy	**97.38**	**97.33**
Bio, CS, Med, Phy	Eng	96.49	96.52
Bio, CS, Eng, Phy	Med	93.68	93.73

Table 4: Domain generalization results of SciBERT_{512}

As the results shown in Table 4, it is evident that the highest performance was achieved when training on the dataset excluded *Physics* data and testing on it, achieving an accuracy of 97.38% and an F1 score of 97.33%. Conversely, the performance in the *Medicine* domain was least impressive, with accuracy and F1 values at 93.68% and 93.73%, respectively. These results indicate that academic articles from various disciplines have a significant impact on text detection capabilities, highlighting the importance of including the labels of academic fields for studying domain generalization.

One possible explanation for the robust performance on the *Physics* domain, despite the exclusion of its data during training, could be attributed to the similarities in structural and linguistic between *Physics* and training domains, particularly those in the natural sciences. The model may have learned features which is easier to generalize on the training domain. On the other hand, the lower performance in the *Medicine* domain indicates that the model struggles more with texts that exhibit higher variability in structure and terminology, pointing to domain-specific challenges. This suggests that future work could focus on a deeper analysis across more fields and on enhancing the robustness of machine-generated text detection.

6 Discussion on ethical and societal implications

The objective of our work is to promote a more nuanced classification of machine-generated scientific texts. However, it is worth to mentioning that we do not condemn or oppose the use of machine learning, particularly the utilize of Large Language Models (LLMs) on scientific articles. In contrast, we recognize the immense potential and benefits that these machine learning technologies bring to various fields, including scientific research, communication, and education. One of our concerns and point we against is the potential for these LLMs to produce misleading or fraudulent scientific papers, which can undermine the integrity of academic research (Zhang et al., 2023). However, a more nuanced categorization would enhance the practical meaning of the task. In certain instances, machine-assisted paraphrasing or summarization of non-plagiarized content is legally valid or even practical, provided it does not introduce additional information or alter the semantics.

Addressing the ethical and societal implications of LLMs is a collective responsibility that extends beyond the research community. We believe that our work can contribute to the advancement of text detection methodologies and the development of effective strategies, thereby enhancing the reliability and credibility of scientific papers. Besides, we expect our study can contribute to the responsible advancement of machine learning technologies, ensuring their positive impact on society.

7 Conclusion

This study focuses on nuanced multi-class detection of machine-generated scientific texts, aiming to bridge the gap in current works, which predominantly prioritize binary classification. To achieve this goal, we build a dataset to simulate diverse text generation methods using LLMs, with the field-of-study label for each scientific text. Experimental findings show that the inclusion of contrastive learning improves the model's discriminative capacity, which beats previous baselines. Furthermore, the analysis on domain generalization underscores varying levels of generalization across different scientific domains, signaling a need for future efforts to enhance detection robustness.

8 Limitations

Although this study proposes a more reasonable approach to simulating machine-generated text in real-world scenarios, there are still some limitations. In the real world, the use of large language models to generate scientific articles can be more complex. For instance, when paraphrasing scientific articles, many users may choose only specific sentences instead of the entire scientific paper as a prompt. They might replace some sentences in the original article with the generated sentences. They might also manually adjust the output to ensure consistency with the paper's original title or key arguments. Some users even use more than one large language model to assist their work on writing or revising text.

In addition, there are certain limitations in the

prompt design used in this study. In real-world scenarios, users often customize prompts to fit their specific needs, which can vary greatly depending on the context. This flexibility is crucial in applications where the generated scientific text needs to serve specific functions, especially in some professional settings.

9 Acknowledgement

This work is partially supported by NSF NAIRR240016, NIH R21EB034911, and Google Cloud research credits.

References

Diego Raphael Amancio. 2015. Comparing the topological properties of real and artificially generated scientific manuscripts. *Scientometrics*, 105:1763–1779.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620.

Deyu Bo, Yuan Fang, Yang Liu, and Chuan Shi. 2024. Graph contrastive learning with stable and scalable spectral encoding. *Advances in Neural Information Processing Systems*, 36.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Guillaume Cabanac and Cyril Labbé. 2021. Prevalence of nonsensical algorithmically generated papers in the scientific literature. *Journal of the Association for Information Science and Technology*, 72(12):1461–1476.

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 preprint arXiv:2107.06751*.

Isabel Cachola, Kyle Lo, Arman Cohan, and Daniel S Weld. 2020. Tldr: Extreme summarization of scientific documents. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4766–4777.

Sumit Chopra, Raia Hadsell, and Yann LeCun. 2005. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 539–546. IEEE.

Liam Dugan, Daphne Ippolito, Arun Kirubarajan, Sherry Shi, and Chris Callison-Burch. 2023. Real or fake text?: Investigating human ability to detect boundaries between human-written and machine-generated text. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 12763–12771.

Holly Else. 2021. Tortured phrases' give away fabricated. *Nature*, 596:328–9.

Yubin Ge, Ly Dinh, Xiaofeng Liu, Jinsong Su, Ziyao Lu, Ante Wang, and Jana Diesner. 2021. Baco: A background knowledge-and content-based framework for citing sentence generation. 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 1466–1478.

Yubin Ge, Devamanyu Hazarika, Yang Liu, and Mahdi Namazifar. 2023a. Supervised fine-tuning of large language models on human demonstrations through the lens of memorization. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*.

Yubin Ge, Sullam Jeoung, Ly Dinh, and Jana Diesner. 2023b. Detection and mitigation of the negative impact of dataset extractivity on abstractive summarization. In *Findings of the Association for Computational Linguistics: ACL 2023*, Toronto, Canada. Association for Computational Linguistics.

Anna Glazkova and Maksim Glazkov. 2022. Detecting generated scientific papers using an ensemble of transformer models. In *Proceedings of the Third Workshop on Scholarly Document Processing*, pages 223–228.

AV Glazkova and DA Morozov. 2023. Applying transformer-based text summarization for keyphrase generation. *Lobachevskii Journal of Mathematics*, 44(1):123–136.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Yury Kashnitsky, Drahomira Herrmannova, Anita de Waard, Georgios Tsatsaronis, Catriona Fennell, and Cyril Labbé. 2022. Overview of the dagpap22 shared task on detecting automatically generated scientific papers. In *Third Workshop on Scholarly Document Processing*.

Chrysovalantis Giorgos Kontoulis, Eirini Papagiannopoulou, and Grigorios Tsoumakas. 2021. Keyphrase extraction from scientific articles via extractive summarization. In *Proceedings of the Second Workshop on Scholarly Document Processing*, pages 49–55.

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

Vijini Liyanage, Davide Buscaldi, and Adeline Nazarenko. 2022. A benchmark corpus for the detection of automatically generated text in academic publications. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, Marseille, France. European Language Resources Association.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Dan S Weld. 2019. S2orc: The semantic scholar open research corpus. *arXiv preprint arXiv:1911.02782*.

Rui Meng, Khushboo Thaker, Lei Zhang, Yue Dong, Xingdi Yuan, Tong Wang, and Daqing He. 2021. Bringing structure into summaries: a faceted summarization dataset for long scientific documents. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 1080–1089.

Edoardo Mosca, Mohamed Hesham Ibrahim Abdalla, Paolo Basso, Margherita Musumeci, and Georg Groh. 2023. Distinguishing fact from fiction: A benchmark dataset for identifying machine-generated scientific papers in the llm era. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, pages 190–207.

Minh Tien Nguyen and Cyril Labbé. 2016. Engineering a tool to detect automatically generated papers. In *BIR 2016 Bibliometric-enhanced Information Retrieval*.

OpenAI. 2022. Chatgpt. *OpenAI blog*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Domenic Rosati. 2022. Synscipass: detecting appropriate uses of scientific text generation. In *Proceedings of the Third Workshop on Scholarly Document Processing*, pages 214–222.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*.

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

Qingyun Wang, Lifu Huang, Zhiying Jiang, Kevin Knight, Heng Ji, Mohit Bansal, and Yi Luan. 2019. Paperrobot: Incremental draft generation of scientific ideas. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1980–1991.

Qingyun Wang, Qi Zeng, Lifu Huang, Kevin Knight, Heng Ji, and Nazneen Fatema Rajani. 2020. Reviewrobot: Explainable paper review generation based on knowledge synthesis. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 384–397.

Kyle Williams and C Lee Giles. 2015. On the use of similarity search to detect fake scientific papers. In *Similarity Search and Applications: 8th International Conference, SISAP 2015, Glasgow, UK, October 12-14, 2015, Proceedings 8*, pages 332–338. Springer.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Xinyu Xing, Xiaosheng Fan, and Xiaojun Wan. 2020. Automatic generation of citation texts in scholarly papers: A pilot study. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6181–6190.

Yuhao Yang, Chao Huang, Lianghao Xia, Chunzhen Huang, Da Luo, and Kangyi Lin. 2023. Debiased contrastive learning for sequential recommendation. In *Proceedings of the ACM web conference 2023*, pages 1063–1073.

Weizhe Yuan, Pengfei Liu, and Graham Neubig. 2022. Can we automate scientific reviewing? *Journal of Artificial Intelligence Research*, 75:171–212.

Jianyi Zhang, Xu Ji, Zhangchi Zhao, Xiali Hei, and Kim-Kwang Raymond Choo. 2023. Ethical considerations and policy implications for large language models: Guiding responsible development and deployment. *arXiv preprint arXiv:2308.02678*.

A Dataset Detailed Domain Distribution

The academic disciplines distribution of our dataset is given in Figure 5. The distribution here is roughly similar to the original domain distribution of S2ORC.

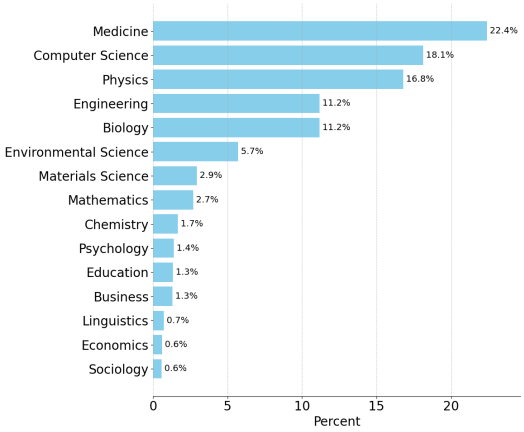

Figure 5: Disciplines distribution.

B Example Data Instances

In the subsequent subsections, we present examples of the fake abstract generation methods described in Section 3.2. The inputs are from the original human-written papers, and the output are the responses generated by ChatGPT using the corresponding prompts.

B.1 Title-Based Abstract Generation

Input:

Title: Influence of NIR Laser and Alternating Magnetic Field on Core-Shell Nanoparticles for Cell Lines Treatment

Output:

Abstract: This study investigates the influence of near-infrared (NIR) laser and alternating magnetic field on core-shell nanoparticles specifically designed for the treatment of cell lines. The interactions between the nanoparticles and the external stimuli hold significant promise for enhancing the efficacy of cell-based therapies. The effects of the combined NIR laser and alternating magnetic field on the core-shell nanoparticles are examined in terms of their potential to improve targeted cellular treatment and therapeutic outcomes. This research contributes to advancing the understanding of utilizing nanotechnology-based approaches in cell line treatments, paving the way for innovative strategies in biomedicine.

B.2 Introduction Summarization

Input:

Full text: Object detection consists of several subtasks such as face recognition, pedestrian detection, skeleton detection, etc., and has use cases such as surveillance systems, autonomous cars, etc. [1] [2]. There are two types of approaches to object detection in images: one based on two-stage detectors and the other based on one-stage detectors. One-step object detection algorithms work by immediately detecting objects on a sample of possible locations such as Fast R-CNN [3], R-CNN [4], Faster R-CNN [5], etc. Two-step object detection algorithms will first propose a set of regions of interest and then rank the relevant regions such as SSD [6], YOLO [7], CenterNet [8], etc.

.....

In future work, we will study the algorithms and models that are effective for the classification of satellite images, and we will try to make an implementation of the most efficient model for the detection and classification of images, especially satellite images. This work provides a contribution to computer scientists and data scientists to help them choose between the different existing models and algorithms, according to their needs and the criteria that matter most to them. The aim of this study is to help the user to make the decision to choose the most efficient model for his project.

Output:

Abstract: This paper presents a comparative study of object detection algorithms in images based on the Weighted Scoring Model (WSM). The study analyzes various criteria such as Average Accuracy (AP), Detection time (FPS), Real-time capability, Number of stages, and Simple network structure to compare popular models like YOLO, Fast R-CNN, Faster R-CNN, R-FCN, Mask R-CNN, CenterNet, CornerNet, RetinaNet, and FCOS, among others. By assigning weights to each criterion and calculating final scores for the models, the study utilizes a spider graph representation to identify the best detection model. The results suggest that YOLO v2 stands out with high accuracy and fast processing speed, making it suitable for real-time applications, compared to other models that may exhibit lower accuracy or slower speed. The study emphasizes the importance of using multi-criteria decision analysis methods like WSM in evaluating object detection algorithms and provides valuable insights for researchers and practitioners in the field.