Written by Human or ChatGPT – Authorship Forensics in the era of Generative AI

Robert SCHMIDT
Faculty of Science and Technology
Athabasca University
Athabasca, Canada
rob.schmidt.cst@gmail.com

Greg FREDIN
Faculty of Science and Technology
Athabasca University
Athabasca, Canada
gregfredin@gmail.com

Kevin HAGHIGHAT
Faculty of Science and Technology
Athabasca University
Athabasca, Canada
kevin.haghighat@me.com

Rita KUO
Department of Computer Science
Utah Valley University
Athabasca, Canada
rita.mcsl@gmail.com

Maiga CHANG
Faculty of Science and Technology
Athabasca University
Athabasca, Canada
maiga.chang@gmail.com
*corresponding author

***Abstract*—In this evolving landscape of text generation, distinguishing between human-written and ChatGPT-generated content has become increasingly important. This paper presents a novel approach to authorship attribution, leveraging both Statistical Natural Language Processing (SNLP) and Convolutional Neural Networks (CNN) techniques to differentiate between documents written by humans and ChatGPTs. The research uses 212 abstracts of academic papers written and published by the research group as the human-written set and asks both ChatGPT 3.5 and 4 to generate corresponding abstracts based on paper titles as AI-written set. Models are trained on a laptop to classify human and AI-written abstract texts in 2-class (i.e., human and ChatGPT) and 3-class (i.e., human, ChatGPT 3.5, and ChatGPT 4) based on their part-of-speech tag frequency distribution patterns. The 2-class model is well-trained in less than ONE minute (i.e., 56.82 seconds) and the 3-class model is well-trained in 7 minutes and 26.076 seconds. The results demonstrate a significant ability of the models to distinguish between human and AI-written text, with precision 0.9682 ($F_{0.5}$ score 0.95) for the 2-class (human and ChatGPT) testing subset and precision 0.9806 ($F_{0.5}$ score 0.96) in the 3-class (human, ChatGPT 3.5, and ChatGPT 4) testing subset. The proposed 3-stage Authorship Forensics approach has been implemented as an open access web application to allow teachers and users to either train their own models or use the existing trained model to get some advice on how the model considers a piece of given text written by human or AI.**

Keywords—Natural Language Processing, Statistical NLP, Neural NLP, Convolutional Neural Networks, Part of Speech, ChatGPT

I. INTRODUCTION

Generative Artificial Intelligence (GenAI) (e.g., ChatGPT) is now well-known and popular with the public. Kasneci and colleagues (2023) discuss and summarize twelve key challenges and risks related to the use of ChatGPT in education [8]. Cotton and colleagues (2024) not only point out the potential of students who take advantage to generate content that has higher quality for their assignments, but also could consider as cheating since the assignments or answers are not done by students [3]; however, teachers have difficulty to distinguish whether a work is written by human or GenAI. OpenAI, the AI research and development company that releases ChatGPT in November 2020, has launched a classifier that can distinguish text written by AI and human authors in January 2023 [9]. However, because the classifier is not fully reliable – it can identify 26% of AI-written text but misclassify 9% text written by human authors as AI-written ones, which achieves precision 0.74 – OpenAI dismissed the classifier on July 20, 2023.

The research proposes a high precision Authorship Forensics method that adopts statistical natural language processing technique to explore how writers develop distinct language patterns. Each author favours certain combinations of Parts Of Speech (POS), forming these writing patterns. While GenAI like ChatGPT can imitate human writing to an extent, yet its writing can still have its own pattern. Once these author-specific statistical patterns are identified, Convolutional Neural Network (CNN) models can be trained to differentiate them, thus distinguishing AI-written texts from human-written ones.

Section 2 starts with the introduction of chatbots and ChatGPT-written detection are reported. This research proposes the 3-stage Authorship Forensics method and explains its details in Section 3 and presents the preliminary study and its results in Section 4. Section 5 discusses four important findings. Finally, in Section 6 we highlight the potential directions and tasks for future research, focusing on clearing teachers' doubts and increasing their willingness of using the proposed Authorship Forensics.

II. 2. RESEARCH BACKGROUND

Chatbots are software applications that can do conversations in text-based, or voice-based, or both with their users via keyboard or microphone. A systematic literature review on chatbots in education has been conducted by Wollny et al. (2021) [15]. They categorize chatbots into three categories

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

according to their roles: learning chatbots, assisting chatbots, and mentoring chatbots. Learning chatbots help students learning/practicing and assess students' understanding and mastery level subject domain knowledge/skills. Assisting chatbots provide students administrative level services like help to register a course and respond to student's questions on assignment deadlines. Mentoring chatbots help students knowing their misconceptions, offering supplemental and remedial learning materials, and developing self-regulation learning habits.

While ChatGPT can take all three roles in education when proper prompts are designed and supplied, Adeshola & Adepoju (2023) address the concerns of the new type of cheating that use ChatGPT for writing program code and essays [1]. Guo and colleagues (2023) create a dataset called HC3-English (Human ChatGPT Comparison Corpus in English) that contains 24,332 questions and 85,449 responses provided by human experts and ChatGPT from four existing datasets and Crawled Wikipedia [7]. They invite 17 volunteers (8 ChatGPT frequent users and 9 amateurs who never heard of ChatGPT). The single test experiment gives each tester a pair of question-response and asks him or her to identify whether the response is written by ChatGPT. The single test experiment receives 63.53% accuracy (for ChatGPT frequent users, 81% accuracy in average). Similarly, Gao et al. (2023) created a dataset that contains 50 chosen paper abstracts from high-impact journals and 50 abstracts written by ChatGPT based on the paper titles and journal names [6]. They give each human reviewer 25 abstracts to review, and human reviewer's guess has precision 0.8293 ($F_{0.5}$ score 0.7944). Guo et al. (2023) and Gao et al. (2023) confirm the challenge teachers face in differentiating whether submitted coursework was done by students or ChatGPT and other GenAI applications [6][7].

Is it possible to detect or identify if a piece of text is written by GenAI? As the old saying goes: 'whoever hung the bell on the tiger's neck must untie it.' One straightforward idea is to have GenAI detect the text written by itself or other GenAI applications. Allen AI's Grover (Generating aRticles by Only Viewing mEtadata Records) is a GenAI application that can generate text by a given title. Zellers et al. (2019) detect fake news generated by the Grover and achieve 73% accuracy with deep pre-trained language models and 92% accuracy with Grover itself [16]. On the other hand, Newhouse and colleagues (2019) find low accuracy when using Grover classifier to detect GPT-2 generated text [12].

From another direction, a logistic regression detector is developed and trained on unigram and bigram features [14]. The detector has 93% accuracy for the 40-word texts generated by GPT-2 XL that has 1.5 billion parameters. More can be done with linguistic analysis on the texts, including vocabulary feature analysis, POS distribution & dependency analysis, and sentiment analysis [7]. Based on the analysis, one logistic regression model and two RoBERTa classifiers are developed for detecting texts written by ChatGPT 3. The logistic regression model has F_1 score 0.8161 on the raw text and 0.8776 at sentence level while RoBERTa model has F_1 score 0.8853 on the raw text and 0.9878 at sentence level. Solaiman and colleagues (2019) also develop a fine-tuning RoBERTa classifier that can identify texts written by GPT-2 XL with accuracy close to 95% [14].

Desaire et al. (2023a) had a model trained with a chosen 64 articles published in Science and 128 examples generated by ChatGPT 3.5 [4]. While the model can differentiate AI and human writings with higher than 99% accuracy when it was tested with two separate sets (each has 30 chosen articles and 60 ChatGPT-written examples), the model was built and optimized based on 20 features identified by humans according to the chosen documents. Desaire et al. (2023b) further apply the same 20 features to 100 chosen papers from 10 chemical journals (10 papers per journal) and train new model. At document level, the model has precision 0.8929 ($F_{0.5}$ score 0.9124) for differentiating texts written by ChatGPT 3.5 and has precision 0.9259 ($F_{0.5}$ score 0.9398) for differentiating texts written by ChatGPT 4 [5].

III. Authorship Identification Process

In Natural Language Processing, a word (as well as a symbol) can be called as a unigram (or 1-gram). When more than one word is putting together in a sentence like "learning outcome" and "computers in education", they are a bigram (or 2-gram) and trigram (or 3-gram). An n-gram indicates a continuous sequence that contains "n" words. It is understandable that any word can be assigned to a syntactic category (e.g., verb, noun, adjective, etc.) according to the role the word plays in the sentence. These syntactic categories are parts of speech (POS).

To reach the goal of differentiating if a text is written by GenAI, a 3-stage method is designed. The 1st stage is inspired by the pet phrases that we always can hear in our daily conversations with people; for instance, in a conversation you may hear the one saying, "like" and "you know", many times. We are also aware of authors may use their favorite words much often or even overuse in their works; for instance, an author may use "plop" when he or she has a character sitting down in the books. The research takes this to a higher level, to the part of speech (POS) level from the n-gram level.

The research wonders if authors may use some POS combinations more frequent than others. While there are only 36 POS tags at the unigram level [11], the number of possible combinations can exceed one million at 4-gram level. To assess the assumption, a web dashboard[1] is implemented based on Statistical NLP (SNLP) techniques. Fig. 1 proves two documents may have different POS usage distributions, for instances, Document #0 uses DT-NN (determiner-singular/mass noun) most and NNS (plural noun) and NN-IN (singular/mass noun-preposition/subordinating conjunction) followed; and, Document #28 uses NNS (plural noun) most and IN (preposition/subordinating conjunction) and JJ (adjective) followed.

[1] https://ngrampos.vipresearch.ca/ngram_pos/statistic/index.php

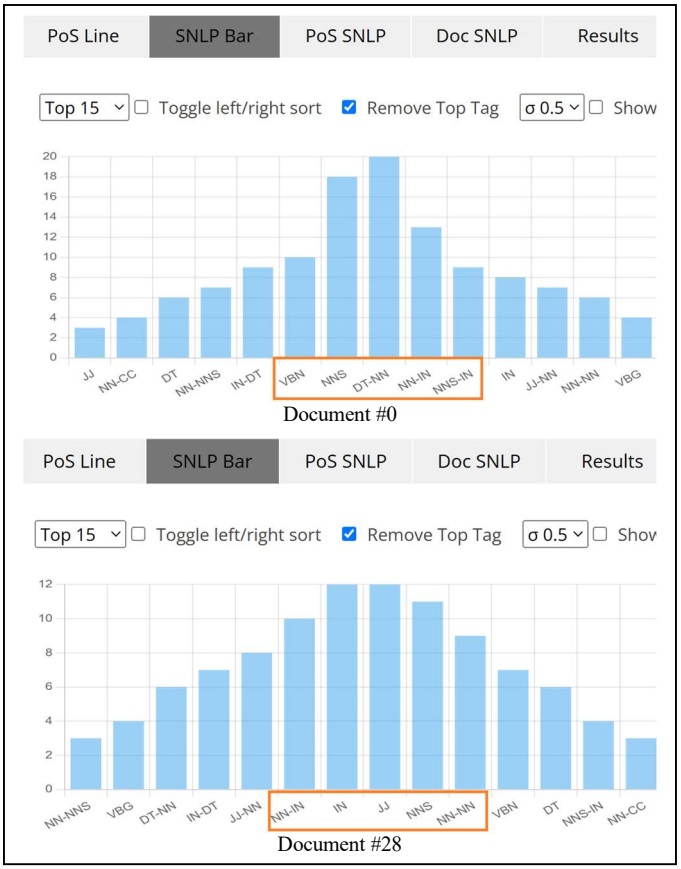


Fig. 1. Top 15 POS tags based on their uses in-between two documents.

At 2nd stage the research considers finding a unified way to present the different POS distributions in a same sequential order so individual document's difference can be told easily. Many of POS combinations are observed from only a few documents. In this research, 22,172 long abstracts from DBpedia, which has content extracted from Wikipedia, are used to filter out the common POS combinations. The SNLP dashboard visualizes and summarizes the most common POS combinations (e.g., top 50). Fig. 2 shows the top 50 POS tags that the 22,172 documents use with standard deviation value 1.5.

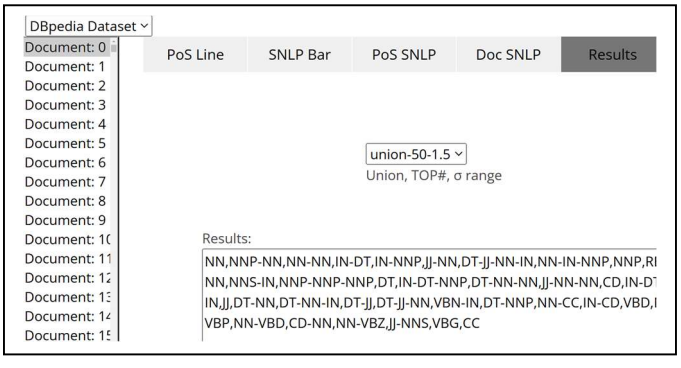


Fig. 2. The common top 50 POS tags filtered from 22,172 documents.

With the top common POS combinations and their sequential order, the proposed method uses line chart to represent a document according to its POS usage distribution. Fig. 3 shows Documents #0 and #28 in line charts again based on the common POS combinations and their sequential order. It is worth to mention that POS labels in the figure are not fully shown on the x-axis due to the space limit. From Fig. 3 we see Document #0 doesn't use POS like IN-JJ (i.e., preposition/subordinating conjunction-adjective) and DT-NN-NN (determiner- singular/mass noun-singular/mass noun) but Document #28 does.

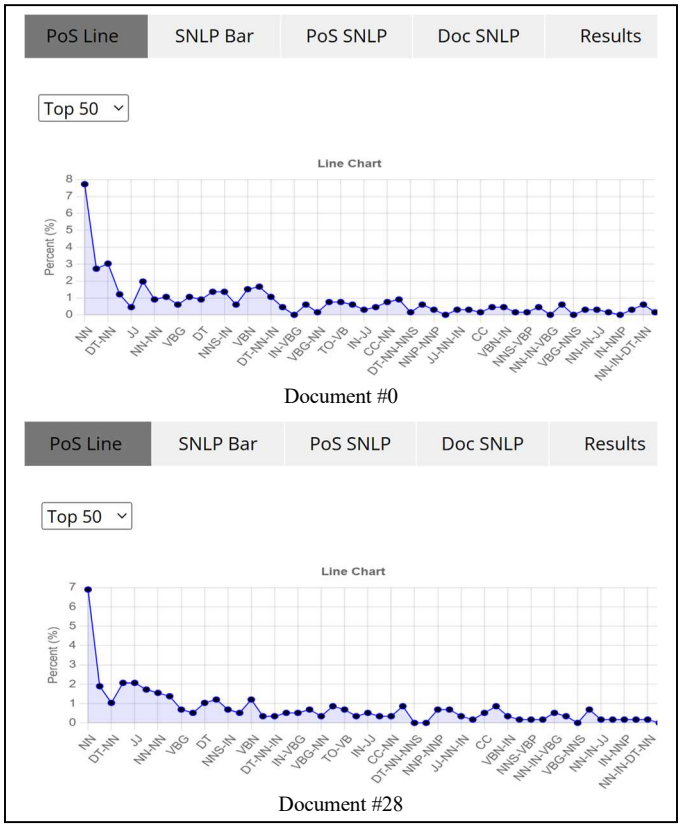


Fig. 3. Two documents' common top 50 POS usage distributions.

The 3rd stage aims to train the computers to recognize and tell the differences among POS usage distributions. The line charts in fact transform the authorship identification problem from a text-based natural language processing and computer linguistic one to an image recognition one. The research believes that it is possible to train a neural network model with the labelled images (i.e., the line charts) so the trained model can recognize a given image's label – that is, human-written or GenAI-written.

Convolutional Neural Network (CNN) is one kind of artificial neural networks, consists of multiple layers, has shown a breakthrough performance on doing scene classification for images, and has been widely and effectively used in pattern recognition applications [13]. The Modified National Institute of Standards and Technology (MNIST) database has 60,000 images of handwritten digits for training purpose and another 10,000 images for testing purpose [10]. Cireşan, Meier, & Schmidhuber (2012) improve and achieve near-human performance with CNN, which is the highest detection rate 99.77% since 1998 [2].

The CNN model designed and used in this research consists of nine layers. To ensure comparability across experiments, the

layers and their configurations are: (1) 1st Conv2D Layer with 32 filters and uses 'relu' for activation function; (2) 1st Pooling Layer for 2x2 pixel regions; (3) 2nd Conv2D Layer with 64 filters and 'relu' function; (4) 2nd Pooling Layer for 2x2 regions; (5) 3rd Conv2D Layer with 128 filters and 'relu' function; (6) 3rd Pooling Layer for 2x2 regions; (7) a Flattening Layer; (8) 1st Dense Layer with 128 units; and (9) 2nd Dense Layer (which is also the Output Layer) with 'softmax' activation function to classify the input image based on the number of authorship classes.

Once a model is trained, the testing subset is used to verify its performance. Precision and recall rates are adopted to measure a trained model's effectiveness, providing insights into the models' accuracy and sensitivity. These are based on the classification outcomes and classified into categories: (1) True Positive (TP), means the model can correctly identify a ChatGPT-written piece of text; (2) False Positive (FP), means the model misidentifying a human-written text as a ChatGPT-written one; (3) False Negative (FN), instead, means the model misidentifying a ChatGPT-written text as a human-written one; and (4) True Negative (TN), means the model can correctly identify a human-written text.

As Fig. 4 shows, Precision measures the proportion of identified ChatGPT-written numbers (i.e., True Positive and False Positive). This research aims to provide teachers advice on the possibility of a course work submitted by their students were written by ChatGPT. In scenarios where ethical considerations are paramount, the emphasis on the precision metric becomes crucial in avoiding false accusations against human authors.

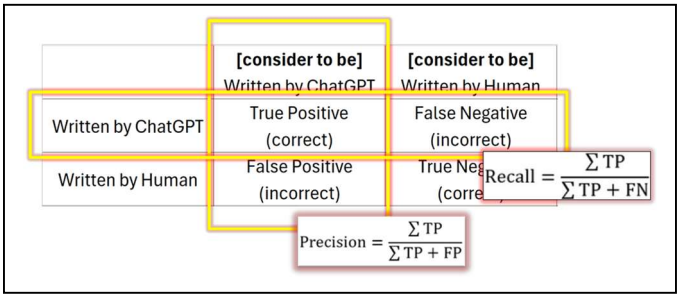


Fig. 4. The distributions of the common top 50 POS tag usage in-between two documents.

As can be told from Fig. 4, if the FP case number is 0, then the precision can be 1 (or says 100%). Maximizing precision in our research is essential to ensure that human-authored texts are accurately recognized, effectively preventing false accusations. The F score is a way that considers both precision and recall together to give people better understanding on a model's performance. The $F_{0.5}$ score gives more weight to precision over the recall while doing the calculations. This research evaluates the proposed 3-stage method in terms of time spent on training models and the performance (i.e., precision, recall, $F_{0.5}$ score) of the trained models.

IV. Preliminary Study

The preliminary pilot study began with a comprehensive preparation of the dataset. The data used in this study consist of 212 human-written abstracts of academic papers published by VIP Research Group[2], alongside 424 AI-generated abstracts, evenly divided between ChatGPT versions 3.5 and 4. The Human-written abstracts cover topics in data analytics, learning technology, educational game, and healthcare. AI-generated abstracts were created using the exact same titles of these human abstracts as prompts, ensuring a direct comparison between the two sets. The common top 50 POS tag usage distribution line chart images were produced for the 636 abstracts.

The research creates two datasets, 2-class has all images labelled with human and ChatGPT and 3-class dataset has images labelled with human, ChatGPT 3.5, and ChatGPT 4. Then both datasets are duplicated, and the images are randomly assigned to training and testing subsets based on 60:40. Training on a laptop the 2-class model to identify text written by Human and ChatGPT required 56.82 seconds and training the 3-class model to identify text written by Human, ChatGPT 3.5, and ChatGPT required seven minutes 26.08 seconds. Both trained 2-class and 3-class models are effective on correctly identifying ChatGPT-written texts and avoiding false accusations against human authors with precision 0.9682 ($F_{0.5}$ score 0.96) and precision 0.9806 ($F_{0.5}$ score 0.96).

V. Findings

The results observed from the preliminary study have revealed four important findings. First, the proposed method is efficient. A model for recognizing and identifying the 2-class (human and ChatGPT) authorship of written text can be trained within couple of minutes. Even the model that can identify the 3-class (human, ChatGPT 3.5, and ChatGPT 4) authorship can be trained within ten minutes. Second, the proposed method has a low requirement for the size of the training data. With 60:40 training and testing data subset distribution, it means the training subset contains around 381 texts in which 343 texts are used for training purpose and 38 texts are used for validation purpose.

Third, the proposed method is effective, and the detections are trustworthy. The trained models have high precisions, 0.9682 for 2-class and 0.9806 for 3-class identification. According to the precision calculation formula, 0.9682 indicates that the trained 2-class models might misidentify up to four human-written texts among a hundred given texts as ChatGPT-written and 0.9806 indicates the trained 3-class models might misidentify up to two human-written texts as written by ChatGPT 3.5 or ChatGPT 4.

The results outperform not only the precision 0.74 that OpenAI's AI classifier has, but also the 93% accuracy that the logistic regression detector has [14], the F_1 score 0.8161 that the logistic regression model achieve at document level [7], the F_1 score 0.8853 that the RoBERTa model has at document level [7], the close to 95% accuracy that the fine-tuning RoBERTa classifier achieve (Solaiman et al., 2019), and precision 0.8929 ($F_{0.5}$ score 0.9124) and precision 0.9259 ($F_{0.5}$ score 0.9398) at document level that the XGBoost classifier has for differentiating texts written by ChatGPT 3.5 and ChatGPT 4 [5].

Last but not least, the proposed 3-stage method this paper presents on the other hand doesn't require human's intervention.

[2] journal and conference papers listed at https://maiga.athabascau.ca/

While the trained model built by [4] has over than 99% accuracy at document level, building and optimizing the model is time and effort consuming and requires human expert's intervention. Moreover, the trained model might not be able to reuse in another context since the 20 features are analyzed and investigated by humans based on the chosen 64 articles published in Science [5]. Human users like teachers only need to prepare a file that contains the text and author label (i.e., human, ChatGPT, etc.) and upload to Authorship Forensics Portal[3]. Coming back after half of hours or shorter, the trained model is ready for them to use.

VI. Future Work

The research team plans to repeat the training with same 636-abstract dataset at least ten times to investigate the time spent and performance in average. Second, the research team is going to collect the perceptions (i.e., predictions or classifications) on the 636 abstracts other free detectors have – for example, Allen AI's Grover classifier[4], GPTZero[5], and GPT-2 Output Detector[6]. We can further compare Authorship Forensics' performance with them. The research team plans to not only train new models but also directly ask the presented trained models to identify the authorship of texts collected from the HC3-English datasets. With such evaluations, we can have clear idea about the reusability of existing trained models and confirm the effectiveness of the proposed 3-stage Authorship Forensics method.

Acknowledgment

The research team wants to thank and acknowledge the support by: NSERC Discovery Grant and NSERC Undergraduate Student Research Awards (USRA). Without their support, this research would not have been possible.

References

[1] I. Adeshola and A. Adepoju, "The opportunities and challenges of ChatGPT in education," *Interactive Learning Environments*, 2023. https://doi.org/10.1080/10494820.2023.2253858

[2] D. Cireşan, U. Meier, and J. Schmidhuber, "Multi-column deep neural networks for image classification," Technical Report No. IDSIA-04-12, 2012. https://doi.org/10.48550/arXiv.1202.2745

[3] D. R. E. Cotton, P. A. Cotton, and J. R. Shipway, "Chatting and cheating: Ensuring academic integrity in the era of ChatGPT," *Innovations in Education and Teaching International*, vol. 61, no. 2, pp. 228–239, 2024. https://doi.org/10.1080/14703297.2023.2190148

[4] H. Desaire, A. E. Chua, M. Isom, R. Jarosova, and D. Hua, "Distinguishing academic science writing from humans or ChatGPT with over 99% accuracy using off-the-shelf machine learning tools," *Cell Reports Physical Science*, vol. 4, p. 101426, 2023. https://doi.org/10.1016/j.xcrp.2023.101426

[5] H. Desaire, A. E. Chua, M.-G. Kim, and D. Hua, "Accurately detecting AI text when ChatGPT is told to write like a chemist," *Cell Reports Physical Science*, vol. 4, p. 101672, 2023. https://doi.org/10.1016/j.xcrp.2023.101672

[6] C. A. Gao, F. M. Howard, N. S. Markov, E. C. Dyer, S. Ramesh, Y. Luo, and A. T. Pearson, "Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers," *npj Digital Medicine*, vol. 6, article 75, 2023. https://doi.org/10.1038/s41746-023-00819-6

[7] B. Guo, X. Zhang, Z. Wang, M. Jiang, J. Nie, Y. Ding, J. Yue, and Y. Wu, "How close is ChatGPT to human experts? Comparison corpus, evaluation, and detection," 2023. https://doi.org/10.48550/arXiv.2301.07597

[8] E. Kasneci, et al., "ChatGPT for good? On opportunities and challenges of large language models for education," *Learning and Individual Differences*, vol. 103, p. 102274, 2023. https://doi.org/10.1016/j.lindif.2023.102274

[9] J. H. Kirchner, L. Ahmad, S. Aaronson, and J. Leike, "New AI classifier for indicating AI-written text," OpenAI, 2023. https://openai.com/index/new-ai-classifier-for-indicating-ai-written-text/

[10] Y. LeCun, C. Cortes, and C. J. C. Burges, "The MNIST database of handwritten digits," n.d. http://yann.lecun.com/exdb/mnist/

[11] M. P. Marcus, M. A. Marcinkiewicz, and B. Santorini, "Building a large annotated corpus of English: the Penn treebank," *Computational Linguistics*, vol. 19, no. 2, pp. 313–330, 1993. https://dl.acm.org/doi/10.5555/972470.972475

[12] A. Newhouse, J. Blazakis, and K. McGuffie, "The industrialization of terrorist propaganda: neural language models and the threat of fake content generation," 2019. https://www.middlebury.edu/institute/academics/centers-initiatives/ctec/ctec-publications/industrialization-terrorist-propaganda-neural

[13] N. Sharma, V. Jain, and A. Mishra, "An analysis of convolutional neural networks for image classification," *Procedia Computer Science*, vol. 132, pp. 377–384, 2018. https://doi.org/10.1016/j.procs.2018.05.198

[14] I. Solaiman, M. Brundage, J. Clark, A. Askell, A. Herbert-Voss, J. Wu, A. Radford, and J. Wang, "Release strategies and the social impacts of language models," 2019. https://doi.org/10.48550/arXiv.1908.09203

[15] S. Wollny, J. Schneider, D. Di Mitri, J. Weidlich, M. Rittberger, and H. Drachsler, "Are we there yet? – A systematic literature review on chatbots in education," *Frontiers in Artificial Intelligence*, vol. 4, p. 654924, 2021. https://doi.org/10.3389/frai.2021.654924

[16] R. Zellers, A. Holtzman, H. Rashkin, Y. Bisk, A. Farhadi, F. Roesner, and Y. Choi, "Defending against neural fake news," in *Proceedings of Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, Vancouver, Canada, Dec. 8–14, 2019. https://doi.org/10.48550/arXiv.1905.12616

[3] https://ngrampos.vipresearch.ca/
[4] https://grover.allenai.org/detect
[5] https://gptzero.me/
[6] https://openai-openai-detector.hf.space/