The present study aims to distinguish between two types of medical texts, namely ChatGPT-generated texts and human-written texts, using appropriate academic writing conventions.

Dr. VIJESH JOE C1,Anuraj Bose2,School of Computer Science and Engineering, VIT, Vellore, India

<anuraj.bose2022@vitstudent.ac.in,>[vijesh.joe@vit.ac.in](mailto:vijesh.joe@vit.ac.in)

*Abstract*—Grammatically flawless and human-like text material can be produced by big language models like ChatGPT, and a significant amount of ChatGPT-generated texts have been published online. However, rigorous validation is required for medical texts like clinical notes and diagnoses, and incorrect medical content produced by ChatGPT could possibly result in  
to misinformation that seriously harms the public and the healthcare, system. One of the earliest inquiries into ethical and responsible artificial intelligence governance and control (AIGC) is represented by the current investigation.  
The topic at hand has to do with the use of artificial intelligence to produce material for the medical industry. The focus of our current work is on closely examining the differences.  
The difference between the medical literature written by human experts and that created by ChatGPT, as well as the development procedure.  
Utilizing machine learning protocols, it is possible to accurately identify and separate medical texts produced by various sources.  
By expanding on the topic, using formal language structures, and deleting colloquialisms, the informal text "ChatGPT" can be updated to fit academic writing.  
This is the first set of datasets produced by ChatGPT that include writings on medicine that were written by human specialists. The linguistic characteristics of these two forms of content are then analyzed to identify differences in vocabulary, part of speech, dependency, sentiment, confusion, etc. Last but not least, we develop and apply machine learning techniques to identify medical text produced by ChatGPT.

Keywords—Medical ethics; language analysis; text classification; chatGPT

# INTRODUCTION

Transformer-based [3] language models have revolutionised and popularised NLP since the introduction of pre-trained language models such as GPT (Generative Pre-trained Transformer) [1] and BERT (Bidirectional Encoder Representations from Transformers) [2] in 2018. Large language models (LLM) [4, 5] have recently proven higher performance on zero-shot and few-shot challenges. Users prefer ChatGPT over other large language models because of its ease of use and ability to deliver grammatically acceptable and human-level answers in a variety of disciplines. Since its release by OpenAI in November 2022, ChatGPT has quickly gained significant attention and has been widely discussed in the natural language processing (NLP) community and other fields.

Researchers employed reinforcement learning to balance the cost and efficiency of data annotation and train a big language model that better corresponds with user intent in a helpful and safe manner.  
  
ChatGPT was created using human feedback (RLHF) [6]. The RLHF trains a reward model using a ranking-based human preference dataset, and using this reward model, ChatGPT may be fine-tuned via proximal policy optimisation (PPO) [7]. As a result, ChatGPT can comprehend the meaning and intent of user inquiries, allowing ChatGPT to react to queries in the most relevant and useful manner. ChatGPT's capacity to handle a range of jobs in multiple domains, in addition to aligning with user intent, is another element that contributes to its popularity.

ChatGPT, on the other hand, is a two-edged sword [12, 13]. Misusing ChatGPT to generate human-like content has the potential to mislead users, leading in incorrect and potentially harmful decisions. Malicious actors, for example, can utilise ChatGPT to generate a huge number of bogus reviews that harm the reputation of high-quality restaurants while artificially enhancing the reputation of low-quality competitors. This is an example that could be harmful to consumers [14].  
ChatGPT has also shown an excellent comprehension of high-stakes fields like medical [15], including subspecialties like radiation oncology. [16]. Medical data is often subject to thorough vetting. Indeed, erroneous medical-related information supplied by ChatGPT can easily lead to misjudgment of illness progression, delay in treatment, or significantly impact patients' lives and health [17].

# Litarature Review

The transformer-based language models have demonstrated a strong language modeling ability.Generally speaking, transformer-based language models are divided into 3 categories: encoder-based models (e.g., BERT [2], Roberta [18], Albert [19] ), decoder-based models (eg: GPT [1], GPT2 [20]), encoder-decoder-based models (e.g. Transformers [3], BART [21], T5 [22]). In order to combine biomedical knowledge with language models, many researchers have added biomedical corpus for training [23, 24, 25, 26, 27]. Alsentzer et al. [28] fine-tuned the publicly release BERT model on the MIMIC dataset [29], and demonstrated good performance on natural language inference and named entity recognition tasks.

BERT was improved by Lee et al. [30] using the PubMed dataset, and it now excels at tasks requiring biomedical named entity identification, biomedical relation extraction, and biomedical question-answering. Luo et al. [31] continue pre-training on the bio-medical dataset based on the core of GPT2 [20] and demonstrate higher performance on six biomedical NLP tasks. AgriBERT [32] for agriculture, ClinicalRadioBERT [33] for radiation oncology, and SciEdBERT [34] for scientific education are a few further cutting-edge uses.  
  
Decoder-based LLM has proven to perform exceptionally well on a number of tasks in recent years [16, 11, 9]. LLM has a lot more trainable 2 parameters than earlier language models, including GPT3, which has 175 billion parameters. GPT-3 is more powerful than earlier models thanks to its larger model size, which also increases its language proficiency to nearly human levels [35]. The ChatGPT is a member of the GPT-3.5 series, which improved its RLHF foundation. According to research [15], ChatGPT successfully completes a medical question-answering test with a score comparable to that of a third-year medical student.

III. Risks associated with utilising ChatGPT

Risks associated with utilising ChatGPTOn the Internet, ChatGPT is creating an increasing amount of content. However, there are some possible risks to keep in mind when using ChatGPT. To start, ChatGPT might restrict human ingenuity. ChatGPT has the capacity to debug code or create college-level essays. It is crucial to take into account if ChatGPT will produce original creative work or just repeat material from their training set. ChatGPT has been outlawed in public schools in New York City.

Second, ChatGPT has a knack to create texts of slightly concerning quality that may trick readers, with the ultimate result being a risky buildup of false information [36]. The usage of ChatGPT-generated content is prohibited on StackOverflow, a well-known forum for programmers and developers. because ChatGPT's average accuracy rate is too low and might adversely impact both the website and the people who depend on it for information.

Thirdly, ChatGPT lacks the competence and understanding required to fully and accurately communicate difficult scientific ideas and information. For instance, ChatGPT cannot yet entirely replace human medical writers since it lacks the same level of understanding and competence in the medical industry [37]. Additionally, human medical writers will be in charge of making sure that the information conveyed is accurate and comprehensive and that it complies with ethical and legal requirements; nonetheless, ChatGPT cannot be held liable.

Additionally, there are some ethical concerns that need to be taken into account when using ChatGPT to create medical texts.  
The trained ChatGPT is biassed because training a big language model necessitates a vast amount of data, the excellent quality of which is impossible to ensure. As an illustration, ChatGPT can produce biassed results and reinforce sexist prejudices [38]. Second, ChatGPT could cause the leakage of confidential data. This could be as a result of the huge language model's memory for the training set's privacy-related data [39]. Thirdly, it concerns the legal system.

In this research, we concentrate on the identification of ChatGPT-generated text for the medical domain in order to stop the improper use of ChatGPT to create medical texts and prevent the potential ethical hazards of employing ChatGPT. Through the OpenAI API, we compile content produced by ChatGPT as well as publicly accessible expert-generated medical knowledge. This research has two objectives: 1) How do human-written and ChatGPT-produced medical material vary from one another? Can ChatGPT or human specialists be distinguished from one other when using machine learning techniques to write medical content?

##### References

.[1] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018. [2] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of NAACL-HLT, pages 4171–4186, 2019. [3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. Advances in neural information processing systems, 30, 2017. [4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. Advances in neural information processing systems, 33:1877–1901, 2020. [5] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. arXiv preprint arXiv:2203.02155, 2022. [6] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. Advances in neural information processing systems, 30, 2017. [7] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347, 2017. 12 [8] Teo Susnjak. Applying bert and chatgpt for sentiment analysis of lyme disease in scientific literature. arXiv preprint arXiv:2302.06474, 2023. [9] Haixing Dai, Zhengliang Liu, Wenxiong Liao, Xiaoke Huang, Zihao Wu, Lin Zhao, Wei Liu, Ninghao Liu, Sheng Li, Dajiang Zhu, et al. Chataug: Leveraging chatgpt for text data augmentation. arXiv preprint arXiv:2302.13007, 2023. [10] Xiang Wei, Xingyu Cui, Ning Cheng, Xiaobin Wang, Xin Zhang, Shen Huang, Pengjun Xie, Jinan Xu, Yufeng Chen, Meishan Zhang, et al. Zero-shot information extraction via chatting with chatgpt. arXiv preprint arXiv:2302.10205, 2023. [11] Zhengliang Liu, Xiaowei Yu, Lu Zhang, Zihao Wu, Chao Cao, Haixing Dai, Lin Zhao, Wei Liu, Dinggang Shen, Quanzheng Li, et al. Deid-gpt: Zero-shot medical text de-identification by gpt-4. arXiv preprint arXiv:2303.11032, 2023. [12] Yiqiu Shen, Laura Heacock, Jonathan Elias, Keith D Hentel, Beatriu Reig, George Shih, and Linda Moy. Chatgpt and other large language models are double-edged swords, 2023. [13] Urfa Khairatun Hisan and Muhammad Miftahul Amri. Chatgpt and medical education: A double-edged sword. Journal of Pedagogy and Education Science, 2(01), 2023. [14] Sandra Mitrović, Davide Andreoletti, and Omran Ayoub. Chatgpt or human? detect and explain. explaining decisions of machine learning model for detecting short chatgpt-generated text. arXiv preprint arXiv:2301.13852, 2023. [15] Aidan Gilson, Conrad W Safranek, Thomas Huang, Vimig Socrates, Ling Chi, Richard Andrew Taylor, David Chartash, et al. How does chatgpt perform on the united states medical licensing examination? the implications of large language models for medical education and knowledge assessment. JMIR Medical Education, 9(1):e45312, 2023. [16] Jason Holmes, Zhengliang Liu, Lian Zhang, Yuzhen Ding, Terence T Sio, Lisa A McGee, Jonathan B Ashman, Xiang Li, Tianming Liu, Jiajian Shen, et al. Evaluating large language models on a highly-specialized topic, radiation oncology physics. arXiv preprint arXiv:2304.01938, 2023. [17] Timothy W Bickmore, Ha Trinh, Stefan Olafsson, Teresa K O’Leary, Reza Asadi, Nathaniel M Rickles, and Ricardo Cruz. Patient and consumer safety risks when using conversational assistants for medical information: an observational study of siri, alexa, and google assistant. Journal of medical Internet research, 20(9):e11510, 2018. [18] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692, 2019. [19] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. arXiv preprint arXiv:1909.11942, 2019. [20] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. OpenAI blog, 1(8):9, 2019. [21] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre13 training for natural language generation, translation, and comprehension. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7871–7880, 2020. [22] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. The Journal of Machine Learning Research, 21(1):5485–5551, 2020. [23] Wenxiong Liao, Zhengliang Liu, Haixing Dai, Zihao Wu, Yiyang Zhang, Xiaoke Huang, Yuzhong Chen, Xi Jiang, Dajiang Zhu, Tianming Liu, et al. Mask-guided bert for few shot text classification. arXiv preprint arXiv:2302.10447, 2023. [24] Homgmin Cai, Wenxiong Liao, Zhengliang Liu, Xiaoke Huang, Yiyang Zhang, Siqi Ding, Sheng Li, Quanzheng Li, Tianming Liu, and Xiang Li. Coarse-to-fine knowledge graph domain adaptation based on distantly-supervised iterative training. arXiv preprint arXiv:2211.02849, 2022. [25] Zhengliang Liu, Mengshen He, Zuowei Jiang, Zihao Wu, Haixing Dai, Lian Zhang, Siyi Luo, Tianle Han, Xiang Li, Xi Jiang, et al. Survey on natural language processing in medical image analysis. Zhong nan da xue xue bao. Yi xue ban= Journal of Central South University. Medical Sciences, 47(8):981–993, 2022. [26] Yiheng Liu, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, Antong Li, Mengshen He, Zhengliang Liu, et al. Summary of chatgpt/gpt-4 research and perspective towards the future of large language models. arXiv preprint arXiv:2304.01852, 2023. [27] Lin Zhao, Lu Zhang, Zihao Wu, Yuzhong Chen, Haixing Dai, Xiaowei Yu, Zhengliang Liu, Tuo Zhang, Xintao Hu, Xi Jiang, et al. When brain-inspired ai meets agi. arXiv preprint arXiv:2303.15935, 2023. [28] Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. Publicly available clinical bert embeddings. arXiv preprint arXiv:1904.03323, 2019. [29] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. Scientific data, 3(1):1–9, 2016. [30] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4):1234–1240, 2020. [31] Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. Biogpt: generative pre-trained transformer for biomedical text generation and mining. Briefings in Bioinformatics, 23(6), 2022. [32] Saed Rezayi, Zhengliang Liu, Zihao Wu, Chandra Dhakal, Bao Ge, Chen Zhen, Tianming Liu, and Sheng Li. Agribert: knowledge-infused agricultural language models for matching food and nutrition. IJCAI, 2022. [33] Saed Rezayi, Haixing Dai, Zhengliang Liu, Zihao Wu, Akarsh Hebbar, Andrew H Burns, Lin Zhao, Dajiang Zhu, Quanzheng Li, Wei Liu, et al. Clinicalradiobert: Knowledge-infused few shot learning for clinical notes named entity recognition. In Machine Learning in Medical Imaging: 13th International Workshop, MLMI 2022, Held in Conjunction with MICCAI 2022, Singapore, September 18, 2022, Proceedings, pages 269–278. Springer, 2022. 14 [34] Zhengliang Liu, Xinyu He, Lei Liu, Tianming Liu, and Xiaoming Zhai. Context matters: A strategy to pre-train language model for science education. arXiv preprint arXiv:2301.12031, 2023.