Hanna Shcharbakova

Investigating the Detection of ChatGPT-Generated Texts across Radiology Reports

Introduction

ChatGPT, Claude, and similar generative language models have garnered immense popularity, raising concerns about their potential to blur the lines between human and AI-generated content. The extensive training dataset sourced from the internet endows these models with the capacity to grasp the nuances of human language (Liao et al., 2023), which enables them to generate human-like texts (Dai et al., 2023; Herbold et al., 2023).

To avoid misuse of such generative language models, the research on AI-generated texts has received a lot of attention in the past years. Typically, this problem is regarded as a classification task, discerning between AI-generated text and human-written one (Wu et al., 2023; Liao et al., 2023).

While many techniques have been suggested for detecting AI-generated text, it's important to note that when considering a specific AI-text generator like ChatGPT, and training a detector on data from a single domain while evaluating it on a test set from that same domain (in-domain evaluation), the in-domain detection task becomes relatively straightforward and can yield extremely high accuracy rates, sometimes reaching a near-perfect score of 100% (Wang et al., 2024). The reason behind this is that text generators trained on a particular domain learn specific patterns and characteristics inherent to that domain. A detector trained on the same data can easily recognize these patterns and distinguish generated text from human-written text within the same domain. For example, (Wang et al., 2024) finetuned RoBERTA (Liu et al., 2019) base model on the Wikipedia dataset and achieved an F1 score of 99,7%.

However, the challenge arises when attempting to generalize the detector to unseen domains (cross-domain evaluation). In these cases, the detection accuracy drops significantly because the detector cannot recognize the new patterns and features specific to other domains. According to the same study by Wang et al. (2024), when a RoBERTA base model was trained on the Wikipedia dataset and then evaluated on the ArXiv dataset from a different domain, its performance suffered a significant drop, achieving an F1 score of only 20.4%.

To achieve effective out-of-domain generalization for detecting AI-generated text, it's crucial to train the detector on diverse data sources like Reddit and Wikipedia. These platforms cover a wide range of topics and writing styles, enabling the detector to learn patterns across different domains and enhancing its cross-domain performance.

Despite advancements, automatic distinguishing between AI-generated and human-written texts remains a challenge. Current research on the detection of AI-generated content falls behind the rapid evolution of LLMs (Yang et al., 2023), and even humans struggle to distinguish machine-generated text from text written by real people (Wu et al., 2023; Gehrmann et al., 2019; Dugan et al., 2022). OpenAI attempted to address the issue of distinguishing between AI-generated and human-written texts by introducing a classifier to differentiate text written by humans from text written by various AI providers. However, on July 20, 2023, they acknowledged that their classifier is no longer available due to its low accuracy rate. When evaluated on English texts, the classifier correctly identified only 26% of AI-written text as "likely AI-written" (true positives), while incorrectly labeling 9% of human-written text as AI-written (false positives) (OpenAI, 2023).

However, studies have shown that there are differences between AI-generated and human-written texts. Recent research highlights differences such as enhanced organization, logic, and formality in AI-generated text. AI-text is typically characterized by a narrower vocabulary, different distribution of part of speech tags, and conveys less emotional intensity and clearer presentation, possibly due to inherent positive bias (Wu et al., 2023; Liao et al., 2023).

Most of the research on AI-generated text detection is focused on English (Guo et al., 2023; Tulchinskii et al., 2023; Liu et al., 2023; Yang et al., 2023; Pegoraro et al., 2023), a limited number of studies explored other languages such as Arabic (Wang et. al, 2024), Bulgarian (Wang et. al, 2024), Chinese (Guo et al., 2023), French (Tulchinskii et al., 2023), German (Yang et al., 2023), Indonesian (Wang et. al, 2024), Japanese (Tulchinskii et al., 2023), Russian (Shamardina et al., 2022) and some others.[1]

Various studies evaluated AI-generated text detection tools, from basic logistic regression classifiers (Alameh et al., 2023; Theocharopoulos et al., 2023) to advanced pre-trained language models (Guo et al., 2023; He et al., 2024). Hans et al., 2024 discuss that AI-generated text detectors can be categorized into two groups: trained detection models and statistical signatures that are characteristic of AI-generated text. Trained detection models involve finetuning pre-trained language models for binary classification of detection (Zellers et al., 2019; Yu et al., 2023). Statistical signature approaches include detectors based on perplexity (Vasilatos et al., 2023), perplexity curvature (Mitchell et al., 2023), log-rank (Su et al., 2023), intrinsic dimensionality of generated text (Tulchinskii et al., 2023), and n-gram analysis (Yang et al., 2023).[2]

Among large language models, ChatGPT is one of the most popular ones, and it has shown impressive performance in various domains, including critical fields like medicine (Gilson et al., 2023; Holmes et al., 2023; Jung et al., 2023) and law (Choi et al., 2023; Katz et al., 2023). Although there has been significant research on AI-generated text detection, most studies focus on detecting generated text in general domains. Only a few studies have investigated the detection of AI-generated texts in specific domains like finance (Guo et al., 2023) and medicine (Liao et al., 2023).

However, the progress in text generation by LLMs poses significant risks, particularly in critical domains like healthcare, where misinformation can have severe consequences. Medical information requires careful validation, as false information can lead to misjudgment of disease trends, treatment delays, or negatively impact patient health (Bickmore et al., 2018). ChatGPT cannot fully replace human medical writers, as they ensure accuracy, completeness, and compliance with ethical and regulatory guidelines (Biswas, 2023). Additionally, applying ChatGPT to generate medical texts raises ethical concerns regarding data bias, privacy leakage, and the lack of a legal framework for accountability in case of mistakes (Liao et al., 2023).

In this project, we aim to analyze the differences between texts authored by humans and those generated by ChatGPT in the under-studied domain of medicine. Additionally, we test several language models to assess their ability to distinguish between medical texts generated by ChatGPT and those written by humans.

We aim to answer the following questions:

1) What are the differences in vocabulary features, lexical diversity, part-of-speech distributions, and dependency relations between human-written and ChatGPT-generated

---

[1] While it may seem that a wide array of languages has been examined, it is crucial to consider that these languages were explored within the scope of only a few studies, rather than numerous. Consequently, the actual depth of research conducted on these languages is relatively limited.

[2] As was suggested in (Hans et al., 2024), for additional details please refer to the surveys (Ghosal et al., 2023), (Tang et al., 2023), (Dhaini et al., 2023), (Guo et al., 2023).

radiology reports, and what do these differences suggest about the writing styles and language patterns?

2) Does the word frequency distribution in human-written and AI-generated radiology reports follow Zipf's law, and what are the implications of the observed differences?

3) How effective are fine-tuned transformer-based models, specifically RoBERTa and ELECTRA, in detecting ChatGPT-generated radiology reports compared to other machine learning algorithms?

Dataset

We utilized the dataset proposed in (Liao et al., 2023). It consists of 2200 samples from a radiology report dataset MiMic (Johnson et al., 2016) as medical texts written by humans, and 2200 samples generated by ChatGPT using the text continuation with demonstration and in-context learning method. The prompt used to obtain AI-generated texts are shown in Figure 1.

Prompt of radiology report:
{example_head_text}
Please continue to write a radiology report with about {example_len} words:
{example_tail_text}

{head_text}
Please continue to write a radiology report with about {text_len} words:

Figure 1. The prompt used to obtain AI-generated samples.

Methods

To identify AI-generated texts, we conducted a 2-level study: first, we employed several descriptive methods, that in previous studies have indicated that they can distinguish AI-generated texts to some extent; then, we utilized deep learning classification models to detect AI-generated content.

Descriptive Analysis

In this chapter, we summarize the approaches we used for descriptive analysis.

Vocabulary Features

As a starting point, we conducted a basic statistical analysis on human-written and AI-generated texts. We analyzed the vocabulary of the two types of texts to explore to understand if we could make assumptions about their distinctive characteristics at this level.

We calculated the total number of words and the number of unique words. The number of unique stemmed was obtained using PorterStemmer[3]. We also calculated the average number of sentences and words per text across all texts.

Additionally, we extracted the top 10 most frequently used words in human-written and AI-generated texts after preprocessing the texts and filtering out stop words using the default stop word list from NLTK[4].

Alongside these statistics, we calculated the lexical diversity of the two text types using the Type-Token Ratio (TTR) metric, which is the ratio of unique tokens to the total number of tokens.

[3] https://www.nltk.org/howto/stem.html
[4] https://www.nltk.org/search.html?q=stopwords

Part-of-Speech and Dependency Analysis

Using spaCy[5], which is known for its high speed and efficiency, we performed part-of-speech tagging and dependency parsing after preprocessing the texts. We compared the distributions of POS tags and dependency relations between human and ChatGPT data.

Zipf's Law Statistics

Furthermore, we analyzed and compared the word frequency distributions of human-written and AI-generated texts according to Zipf's Law (Zipf, 1949), a statistical phenomenon depicting the frequency of a word as inversely proportional to its rank. This law provides insights into the distribution of word frequencies in human language. The hypothesis to be tested is whether there are significant differences in the word frequency distribution patterns between human and AI-generated texts. To assess the similarity or dissimilarity of the word frequency distributions, we used the Kolmogorov-Smirnov test (Kolmogorov-Smirnov, 1933), which compares the cumulative distribution functions of two samples to determine if they come from the same distribution or not. The statistical test and visualisations can potentially reveal differences in the distributions, which could be useful for distinguishing between the two types of texts.

Cosine similarity

We calculated the cosine similarity between the two types of texts as a measure of semantic similarity using the Universal Sentence Encoder (Cer et al., 2018).[6] The USE is a pre-trained model that encodes text into high-dimensional vectors, allowing for semantic similarity comparisons. Thus, based on the cosine similarity, we can make assumptions about how semantically similar human-written and AI-generated texts are.

Cosine similarity is a metric used to determine the cosine of the angle between two non-zero vectors in a multi-dimensional space. The Universal Sentence Encoder employs a deep neural network that has been pre-trained on a large corpus of texts, enabling it to understand and encode the meaning of sentences in a way that captures semantic similarities and relationships between them.

Deep Learning Classification Models

To detect AI-generated texts, we performed a binary classification task. We tested two deep classifiers with different architectures: RoBERTa-base and ELECTRA-base-discriminator.

RoBERTa is a transformer-based model pre-trained on a large corpus of English data in a self-supervised manner, specifically with the masked language modelling objective. This model is primarily to be fine-tuned for tasks that use the whole sentence, potentially with some parts masked, to make decisions, such as sequence classification, token classification, etc. Finetuning RoBERTa model has already shown good results in previous studies on AI-generated text detection (Shu et al., 2021; Wang et al., 2023; Gaggar et al., 2023).

Additionally, we fine-tuned the ELECTRA model. ELECTRA follows a unique pre-training approach involving two transformer models: a generator and a discriminator. The generator's role is to replace tokens in a sequence, being trained as a masked language model. The discriminator tries to identify which tokens were replaced by the generator. This model is more suitable for our task than RoBERTa because it was pre-trained to predict whether a token in a corrupted input was plausibly replaced by the generator. Although ELECTRA has

[5] https://spacy.io/

[6] We use a TensorFlow implementation of the Universal Sentence Encoder.

shown slightly lower performance than RoBERTa for in-domain evaluation (Wang et al., 2023), we still included it in our analysis.

We fine-tuned both RoBERTa and ELECTRA models for sequence classification on a binary classification task to distinguish AI-generated and human-written text content. For fine-tuning the models, we divided the dataset into training, validation, and test sets, with a 70/10/20 split, respectively. The AdamW optimizer was used for fine-tuning the model parameters, with a learning rate set to 5^{e}-5. We fine-tuned both models for 5 epochs, and the models were trained in Google Colab using TPU.

Results and Discussion

In this section, we first provide the findings from our descriptive analysis, and then discuss the evaluation results of experiments with classification models RoBERTa and ELECTRA.

Vocabulary Features

In Table 1, we report the vocabulary features and the values of a lexical diversity measure TTR for the studied corpus. The top 10 most used words in both human-written and ChatGPT-generated texts, along with their respective word clouds, can be found in Appendices A and B.

	Word Count	**Unique Word Count**	**Unique Stemmed Word Count**	**Sentence Count**	**Average Sentence Length**	**Average Text Length**	**TTR**
Human	245066	11600	10863	**12.73**	8.78	111.39	**0.047**
ChatGPT	**993511**	**13478**	**12582**	12.52	**9.02**	**112.9**	0.014

Table 1. Vocabulary features and TTR values of human-written and ChatGPT-generated texts.

From the obtained results, we can draw the conclusion that the ChatGPT-generated texts have a significantly higher total word count (993511) compared to human-written texts (245066). Although AI texts have a higher number of unique words (13478) compared to human-written texts (11600), the generated texts contain very few unique words in proportion to the total number of words. When considering unique stemmed words, the difference in unique word count is smaller, with 12582 for ChatGPT and 10863 for human texts. Therefore, we can't claim that the AI model has a more diverse vocabulary or tends to use a wider range of words in its outputs.

The average text length is slightly higher for AI-generated texts (112.9) than for human-written texts (111.39), supporting the observation that AI-generated texts tend to be longer (Guo et al., 2023).

The average sentence count is similar between human-written texts (12.73) and AI-generated texts (12.52). However, the average sentence length is slightly higher for AI-generated texts (9.02) compared to human-written texts (8.78), suggesting that the AI model may produce slightly longer sentences. It is worth noting that based on the prompt design, the authors specified the desired text length, and the model followed the instructions.

The Type-Token Ratio is higher for human-written texts (0.047) than for AI-generated texts (0.014), which is expected based on the comparison of the total word count and unique word count. This implies that human-written texts exhibit greater lexical diversity, using a more varied vocabulary relative to the total number of words. The lower TTR for AI-generated texts could indicate a tendency to reuse words more frequently. It should be noted that this metric is sentence length sensitive; however, the average sentence lengths in

both text types are almost the same, and the average text lengths are also very similar, so the use of this metric is reasonable.

Part-of-Speech and Dependency Analysis

The top 10 most common POS tags in the two types of texts are presented in Figure 2. Here we see that nouns are the most prevalent part of speech in the texts. ChatGPT uses slightly more nouns and determiners (which is logical, as they often accompany nouns). The frequent use of nouns may indicate that the text is more argumentative, showing information and objectivity. ChatGPT's increased usage of coordinating conjunctions may be explained by its tendency to convey a logical flow of thought to produce coherent texts.

The more prevalent use of coordinating conjunctions by ChatGPT implies that the AI model tends to employ these conjunctions to connect ideas and create a logical progression within the text.

The alignment of these observations with previous research findings (Wu et al., 2023; Guo et al., 2023) further validates the patterns observed in the part-of-speech distributions between human-written and AI-generated texts.

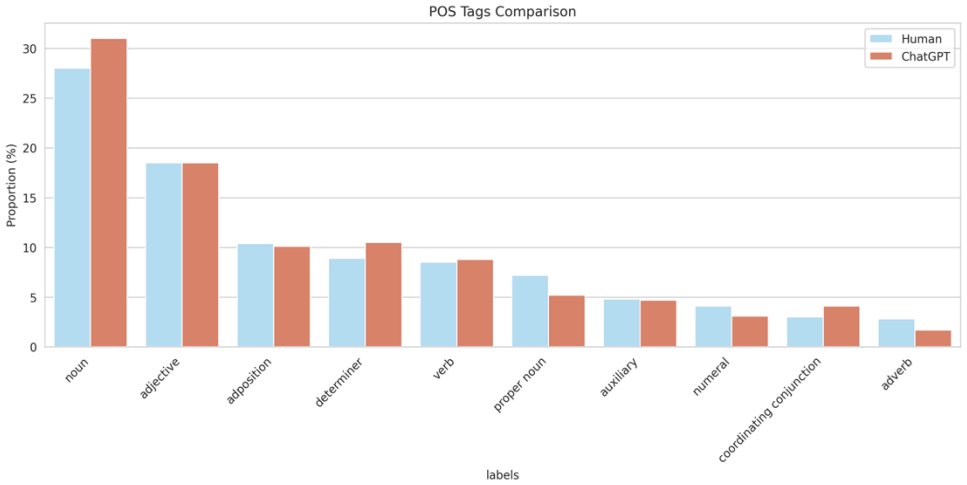

Figure 2. Top 10 most frequent POS tags across human-written and AI-generated texts.

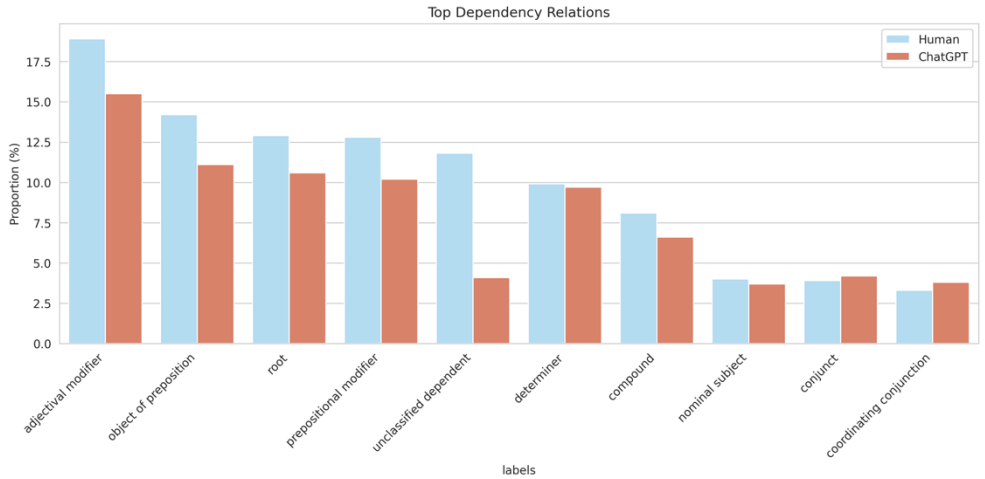

Figure 3. Proportion of frequency of the top 10 most frequent dependency relations across human-written and AI-generated texts.

Figure 3 shows the proportion or frequency of the top 10 most common dependency relations found in human-written and AI-generated texts.

For the most frequent dependency relation, adjectival modifier, human-written texts have a higher proportion (around 18%) compared to ChatGPT-generated texts (around 15%). Similarly, for almost every other dependency relation, human-written texts have a higher proportion, excluding conjuncts and coordinating conjunctions, which is predictable based on the analysis of POS tags above.

The results for the unclassified dependent relation are very remarkable. An unclassified dependent means a dependency when it is impossible to determine a more precise relation. This could be attributed to an unusual grammatical structure or a constraint within conversion or parsing software.[7] We can assume that human writers presented more grammatically interesting and complex constructions that are more difficult to analyze compared to AI.

These results suggest that humans use different grammatical constructions, which could potentially indicate differences in the writing styles and language patterns between human-written and AI-generated texts, reflecting the AI model's tendencies in constructing sentences and using certain grammatical structures more frequently.

Zipf's Law Statistics

In Figures 4 and 5 the curves to visualize Zipf's law for the corpus are shown.

The frequency curve plots indicate that the human text follows a smoother, more gradual decline in word frequencies compared to the AI text. On the log-log plot, the human curve is closer to a straight line, which suggests the human text more closely follows a power law or Zipf's law distribution. The AI curve has a more pronounced curvature, deviating further from the straight line expected under Zipf's law.

The Kolmogorov-Smirnov test was used to compare the frequency distributions of words in human-written vs. AI-generated text. We obtained the test statistic of 0.2622 with a p-value of 0.0000, which indicates a statistically significant difference between the two distributions.

Overall, the statistical test and visualizations reveal significant differences between the human and AI text distributions.

[7] https://universaldependencies.org/

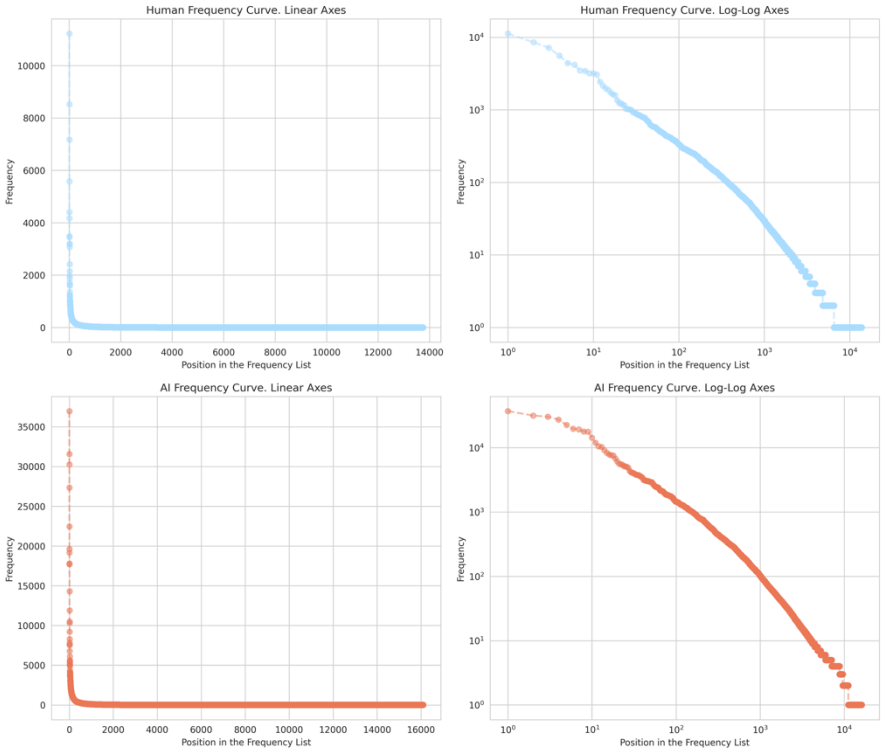

Figure 4. Plots to visualize Zipf's law for the corpus: linear and log-log axes.

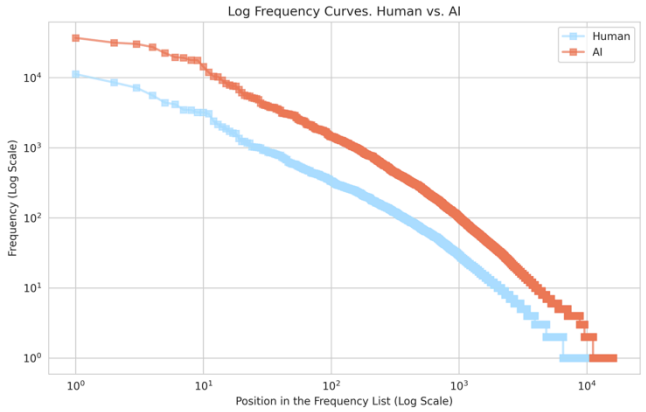

Figure 5. Plot to visualize Zipf's law for the corpus: log-log axes.

Cosine Similarity

The average cosine similarity between human-written and AI-generated radiology reports is 0.558, indicating a moderate level of similarity overall. However, the top 10 most similar pairs (see Appendix C) have much higher similarity scores, ranging from 0.936 to 0.954.

Upon comparing the most similar pairs, several observations can be made. The AI-generated texts mimic the structure, language, and content of the human-written reports. They include similar sections such as indication, comparison, findings, and impression. ChatGPT uses medical terminology and abbreviations accurately, demonstrating an understanding of the domain-specific language. It effectively captures the key findings and impressions from the human-written reports, such as the presence or absence of pneumothorax, etc. In some cases, the AI-generated texts provide additional context or history not explicitly mentioned in the human-written report, possibly drawing from patterns observed in the training data.

However, there are also some differences between the human and AI-generated texts. First, the AI-generated texts sometimes include additional or slightly different clinical information in the indication section, which may not be directly relevant to the radiographic findings. ChatGPT occasionally uses different phrasing or sentence structure compared to human-written reports, although the overall meaning remains similar. In a few cases, the AI-generated texts mention findings or impressions that are not explicitly stated in the corresponding human-written report, which could be due to the AI model's learned patterns from the training data.

Overall, the high similarity scores and the qualitative comparison of the most similar pairs suggest that the AI model has successfully learned to generate radiology reports that closely resemble human-written reports in terms of structure, content, and clinical relevance. This indicates the potential for AI to assist radiologists in report generation, although further validation and refinement may be necessary to ensure the reliability and accuracy of the generated reports in clinical practice.

However, it is important to note that the similarity scores alone do not guarantee the clinical accuracy or completeness of the AI-generated reports. The AI model may still make errors or omissions that could have significant clinical implications. Therefore, human oversight and interpretation remain crucial in the context of radiology reporting.

Classification Models RoBERTa and ELECTRA

	Precision	Recall	F1
CART (Liao et al., 2023)	0.837	0.836	0.836
XGBoost (Liao et al., 2023)	0.929	0.928	0.928
BERT (Liao et al., 2023)	0.982	0.982	0.982
RoBERTa	**0.992**	**0.992**	**0.992**
ELECTRA	0.979	0.978	0.978

Table 2. The performance of different models in detecting ChatGPT-generated texts.

Table 2 presents the performance of various models in detecting ChatGPT-generated text in radiology reports. The models evaluated include CART, XGBoost, and BERT from the study by Liao et al. (2023), as well as RoBERTa and ELECTRA, which we fine-tuned specifically for this task.

The evaluation metrics (precision, recall, and F1 score) demonstrate that all the models perform well in distinguishing between human-written and ChatGPT-generated radiology reports. However, our fine-tuned RoBERTa model stands out, surpassing all other models with impressive precision, recall, and F1 score of 0.992 each. This indicates that RoBERTa is highly accurate in correctly identifying both human-written and ChatGPT-generated reports, with a low rate of false positives and false negatives.

RoBERTa and ELECTRA, both fine-tuned transformer-based models, outperform the traditional machine learning algorithms, CART and XGBoost, emphasizing the effectiveness of these models for the given task.

The exceptional results we obtained can be attributed to several factors, as discussed in the Introduction section. RoBERTa and ELECTRA demonstrate their ability to effectively capture contextual information and learn meaningful representations from text data. The dataset used for fine-tuning our models plays a crucial role in their performance, as it consists of a balanced mix of human-written and ChatGPT-generated samples, ensuring its representativeness.

Furthermore, the in-context learning method employed to generate the ChatGPT samples might introduce certain patterns that the models can effectively learn to distinguish

from human-written text. This characteristic of the generated text may contribute to the models' ability to accurately identify ChatGPT-generated content.

Conclusion

This study investigates the differences between human-written and ChatGPT-generated radiology reports using various linguistic features and machine learning models.

The analysis of vocabulary features revealed that while AI-generated texts had a higher total word count, they exhibited lower lexical diversity compared to human-written texts.

Part-of-speech and dependency relation analyses showed differences in the usage of nouns, determiners, coordinating conjunctions, and unclassified dependents, suggesting variations in writing styles and grammatical constructions between the two text types.

The examination of Zipf's law statistics and cosine similarity further highlighted the differences between human and AI-generated reports. To the best of our knowledge, this study is the first to apply Zipf's law and associated statistical test to the task of distinguishing between human-written and AI-generated texts, providing novel insights into the differences in word frequency distributions between the two types of text.

Fine-tuned transformer-based models, particularly RoBERTa, outperformed traditional machine learning algorithms in detecting ChatGPT-generated texts, achieving high precision, recall, and F1 scores.

The effectiveness of these models can be attributed to their ability to capture contextual information from the balanced dataset used for fine-tuning, demonstrating superior performance and adaptability compared to conventional training-based methods.

As language models continue to advance, the distinction between machine-generated and human-written text is becoming increasingly blurred, making the detection of AI-generated content a growing challenge (Tang et al., 2023; Sadasivan et al., 2023). This trend suggests that in the future, it may be extremely difficult, if not impossible, to reliably identify machine-generated text. Moreover, Liang et al. (2023) have highlighted a potential bias in GPT detectors, which may unfairly flag the writing of non-native English speakers as machine-generated. These developments cast doubt on the long-term viability of black-box detection methods and underscore the importance of ongoing research and innovation to address these issues and develop more robust and equitable solutions for identifying AI-generated text.

Limitations

The study has several limitations that should be acknowledged.

First, it is focused on a specific domain (radiology reports) and a single AI model (one version of ChatGPT), which may limit the generalizability of the findings to other domains or AI models. The dataset used for fine-tuning the models was generated using a specific prompt and in-context learning method, which could introduce certain patterns that the models can learn to distinguish from human-written text. As the dataset is prompt-sensitive, the performance of the models on texts generated using different prompts or methods may vary.

Additionally, the study did not explore the potential biases or errors that the AI model may introduce in the generated reports. Moreover, we focused on the analysis of directly generated text by ChatGPT, and the results and conclusions may differ if the generated text undergoes additional processing, such as paraphrasing, before being analyzed.

Furthermore, while the study applied Zipf's law and the Kolmogorov-Smirnov test to analyze the word frequency distributions, conducting additional statistical tests in conjunction

with Zipf's law could provide a more comprehensive understanding of the differences between human-written and AI-generated text.

Future work

To further enhance the understanding of the differences between human-written and AI-generated radiology reports and improve the detection of AI-generated text, several additional steps could be taken within this project.

Expanding the study to include a wider range of AI models and datasets from different domains would allow for an assessment of the generalizability of the findings and the performance of the detection models. Including closed-source models, such as GPTZero[8], in the comparative analysis would provide valuable insights into their effectiveness in detecting AI-generated text and explore the potential benefits of combining different approaches.

Investigating the cross-domain performance of fine-tuned models is crucial to determine their applicability and robustness in various contexts.

Additionally, investigating the impact of different prompts and generation methods on the linguistic features and detectability of AI-generated text is essential.

Exploring the performance of the detection models on AI-generated text that has undergone additional processing, such as paraphrasing, would assess their robustness and adaptability. Additionally, incorporating stylometric features and text complexity analysis could provide valuable insights into the differences between human-written and AI-generated text, potentially leading to the development of more sophisticated detection methods.

We intended to conduct an analysis of the perplexity distributions for AI-generated and human-written texts using pre-trained ELECTRA and RoBERTa models. Calculating perplexity using pre-trained language models before fine-tuning them for detecting AI-generated texts helps to identify differences in language patterns, guide model selection, and assess the effectiveness of the fine-tuning process. However, directly comparing the raw perplexity values between the two models is not feasible since they output scores are on very different scales. We attempted to normalize the perplexity scores by using log-perplexity but failed. The perplexity scores we obtained can be found in Appendix D.

[8] https://gptzero.me/

Experience with LLM

This report was written using ChatGPT (a current version without a subscription). The main points of my work with this model are:

1) I primarily used it to translate text I had written from Russian to English (prompt used: ‘to English: *text in Russian*‘).

2) I used ChatGPT for paraphrasing. This was very helpful for me because I tend to write complex, long sentences with multiple clauses, and when I try to break them down into smaller ones, I end up with very short sentences. Therefore, I used the prompt: ‘make text coherent and consistent: *text*’. Also, when I didn't like how I composed a sentence, I asked ChatGPT to make it pretty, but it produced a very high-flown text, so I used the prompt the above prompt.

3) I encountered difficulties when writing the Introduction section, as I couldn't get the model to write the text the way I wanted.

4) When writing the entire chapter, I fed the chapter to ChatGPT and asked it to rearrange the paragraphs so the text would be coherent and consistent.

5) ChatGPT also helped with paraphrasing in the Introduction section when I wanted to retell what was written in an article. The algorithm was as follows: I write in my own words in Russian, translate it into English, and get a sentence. Then I ask ChatGPT to rephrase the original sentence. After that, I compare the two versions and rewrite them into one sentence.

6) In general, it seemed to me that ChatGPT doesn't always adhere to a super strict narrative (with words like “It's worth noting”, etc.), and I would like it to, so I introduced strict language into the text myself.

7) When writing the results for the Vocabulary Features section, there were many similar sentences comparing statistics. ChatGPT helped me with this by grammatically reconstructing the sentences, even though there were many of them.

Overall, it seems that it’s not so easy to write text with ChatGPT, but it's worth noting that my prompts were quite simple.

However, it saves a lot of time. In the past, finding the perfect synonym for a word could be a very challenging task, as the available options often didn’t quite fit the context. Consequently, the process of refining the text, ensuring correct grammar, and searching for the right words could become very challenging.

References

Bickmore, T.W., Trinh, H., Ólafsson, S., O'Leary, T.K., Asadi, R., Rickles, N.M., & Cruz, R. (2018). 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.

Biswas, S. (2023). ChatGPT and the future of medical writing. *Radiology*, 307(2).

Cer, D.M., Yang, Y., Kong, S., Hua, N., Limtiaco, N., John, R.S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Sung, Y., Strope, B., & Kurzweil, R. (2018). *Universal Sentence Encoder.* arXiv preprint arXiv:1803.11175.

Choi, J., Hickman, K., Monahan, A., & Schwarcz, D. (2022). ChatGPT Goes to Law School. *Journal of Legal Education,* 71 (3).

Clark, K., Luong, M., Le, Q., & Manning, C. (2020). ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. *International Conference on Learning Representations*.

Dai, H., Liu, Z., Liao, W., Huang, X., Wu, Z., Zhao, L., Liu, W., Liu, N., Li, S., Zhu, D., Cai, H., Sun, L., Li, Q., Shen, D., Liu, T., & Li., X. (2023). *Chataug: Leveraging chatgpt for text data augmentation.* arXiv preprint arXiv:2302.13007.

Dhaini, M., Poelman, W., & Erdogan, E. (2023). Detecting chatgpt: *A survey of the state of detecting chatgpt-generated text. arXiv preprint arXiv:2309.07689*.

Dugan, L., Ippolito, D., Kirubarajan, A., Shi, S., & Callison-Burch, C. (2022). Real or Fake Text?: Investigating Human Ability to Detect Boundaries Between Human-Written and Machine-Generated Text. AAAI Conference on Artificial Intelligence.

Gaggar, R., Bhagchandani, A., & Oza, H.V. (2023). *Machine-Generated Text Detection using Deep Learning.* arXiv preprint arXiv:2311.15425.

Gehrmann, S., Strobelt, H., & Rush, A.M. (2019). *GLTR: Statistical Detection and Visualization of Generated Text*. Annual Meeting of the Association for Computational Linguistics.

Ghosal, S., Chakraborty, S., Geiping, J., Huang, F., Manocha, D., & Bedi, S. (2023). *Towards possibilities & impossibilities of ai-generated text detection: A survey.* arXiv preprint arXiv:2310.15264.

Gilson, A., Safranek, C., Huang, T., Socrates, V., Chi, L., Taylor, R., & Chartash, D. (2023) 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).

Guo,B., Zhang, X., Wang, Z., Jiang, M., Nie, J., Ding,Y., Yue, J., & Wu, Y. (2023). *How close is chatgpt to human experts? Comparison corpus, evaluation, and detection*. arXiv preprint arXiv:2301.07597.

He, X., Shen, X., Chen, Z., Backes, M., & Zhang, Y. (2023). *Mgtbench: Benchmarking machine-generated text detection.* arXiv preprint arXiv:2303.14822.

Herbold, S., Hautli-Janisz, A., Heuer, U., Kikteva, Z., & Trautsch, A.(2023). A large-scale comparison of human-written versus ChatGPT-generated essays. *Sci Rep* 13 (18617).

Holmes, J., Liu, Z., Zhang, L., Ding, Y., Sio, T., McGee, L., Ashman, J., Li, X., Liu,T., Shen, J., & Liu., W. (2023). Evaluating large language models on a highly-specialized topic, radiation oncology physics. *Radiation Oncology* 13.

Johnson, A., Pollard, T., Shen, L., Lehman, L. W. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L., & Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. *Scientific data*, *3*(1), 1-9.

Jung, L., Gudera, J., Wiegand, T., Allmendinger, S., Dimitriadis, K., & Koerte, I. (2023). ChatGPT passes German state examination in medicine with picture questions omitted. *Dtsch Arztebl Int* 120 (21).

Katz, D., Bommarito, M., Gao, S., & Arredondo, P. (2023). *GPT-4 Passes the Bar Exam.*

Kirchner, J., Ahmad, L., Aaronson, S., & Leike, J. (2023). New AI classifier for indicating AI-written text.

Kolmogorov-Smirnov, A. (1933). Sulla determinazione empírica di uma legge di distribuzione. *Giornale dell'Instituto Italiano degli Attuari* 4.

Liao, W., Liu, Z., Dai, H., Xu, S., Wu, Z., Zhang, Y., Huang, X., Zhu, D., Cai, H., Liu, T., & Li, X. (2023). Differentiating ChatGPT-Generated and Human-Written Medical Texts: Quantitative Study. *JMIR Medical Education*, 9.

Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J.Y. (2023). GPT detectors are biased against non-native English writers. *Patterns* 4.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach.* arXiv preprint arXiv:1907.11692.

Liu, Y., Zhang, Z., Zhang, W., Yue, S., Zhao, X., Cheng, X., Zhang, Y., & Hu, H. (2023). *ArguGPT: evaluating, understanding and identifying argumentative essays generated by GPT models*. arXiv preprint arXiv:2304.07666.

Mitchell, E., Lee, Y., Khazatsky, A., Manning, C.D., & Finn, C. (2023). DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature. *International Conference on Machine Learning.*

Mitrović, S., Andreoletti, D., & Ayoub, O. (2023). *Chatgpt or human? detect and explain. explaining decisions of machine learning model for detecting short chatgpt-generated text.* arXiv preprint arXiv:2301.13852.

OpenAI Blog. Link: https://openai.com/blog/new-ai-classifier-for-indicating-ai-written-text.

Pegoraro, A., Kumari, K., Fereidooni, H., & Sadeghi, A. R. (2023). *To ChatGPT, or not to ChatGPT: That is the question!.* arXiv preprint arXiv:2304.01487.

Sadasivan, V.S., Kumar, A., Balasubramanian, S., Wang, W., & Feizi, S. (2023). *Can AI-Generated Text be Reliably Detected?* arXiv preprint arXiv: 2303.11156.

Shamardina, T., Mikhailov, V., Chernianskii, D., Fenogenova, A., Saidov, M., Valeeva, A.E., Shavrina, T., Smurov, I., Tutubalina, E., & Artemova, E. (2022). Findings of the The RuATD Shared Task 2022 on Artificial Text Detection in Russian. *Computational Linguistics and Intellectual Technologies.*

Shu, K., Li, Y., Ding, K., & Liu, H. (2021). Fact-Enhanced Synthetic News Generation. *Proceedings of the AAAI Conference on Artificial Intelligence* 35(15).

Su, J., Zhuo, T.Y., Wang, D., & Nakov, P. (2023). DetectLLM: Leveraging Log Rank Information for Zero-Shot Detection of Machine-Generated Text. *Conference on Empirical Methods in Natural Language Processing.*

Tang, R., Chuang, Y., & Hu, X. (2023). *The Science of Detecting LLM-Generated Texts.* arXiv preprint arXiv:2303.07205.

Theocharopoulos, P. C., Anagnostou, P., Tsoukala, A., Georgakopoulos, S. V., Tasoulis, S. K., & Plagianakos, V. P. (2023). Detection of fake generated scientific abstracts. *IEEE Ninth International Conference on Big Data Computing Service and Applications.*

Vasilatos, C., Alam, M., Rahwan, T., Zaki, Y., & Maniatakos, M. (2023). *HowkGPT: Investigating the Detection of ChatGPT-generated University Student Homework through Context-Aware Perplexity Analysis.* arXiv preprint arXiv:2305.18226.

Wang, Z., Cheng, J., Cui, C., & Yu, C. (2023). *Implementing BERT and fine-tuned RobertA to detect AI generated news by ChatGPT.* arXiv preprint arXiv:2306.07401.

Wang, Y., Mansurov, J., Ivanov, P., Su, J., Shelmanov, A., Tsvigun, A., Whitehouse, C., Afzal, O.M., Mahmoud, T., Aji, A., & Nakov, P. (2023). *M4: Multi-generator, Multi-domain, and Multi-lingual Black-Box Machine-Generated Text Detection.* arXiv preprint arXiv:2305.14902.

Wu, J., Yang, S., Zhan, R., Yuan, Y., Wong, D. F., & Chao, L. S. (2023). *A survey on llm-gernerated text detection: Necessity, methods, and future directions.* arXiv preprint arXiv:2310.14724.

Yang, X., Cheng, W., Petzold, L., Wang, W. Y., & Chen, H. (2023). *Dna-gpt: Divergent n-gram analysis for training-free detection of gpt-generated text. arXiv preprint arXiv:2305.17359*.

Yu, P., Chen, J., Feng, X., & Xia, Z. (2023). *CHEAT: A large-scale dataset for detecting ChatGPT-writtEn AbsTracts.* arXiv preprint arXiv:2304.12008.

Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F., & Choi, Y. (2019). Defending against neural fake news. *Advances in neural information processing systems* 32.

Zipf, G.K. (1949). *Human Behaviour and the Principle of Least Effort: an Introduction to Human Ecology.*

Appendix A

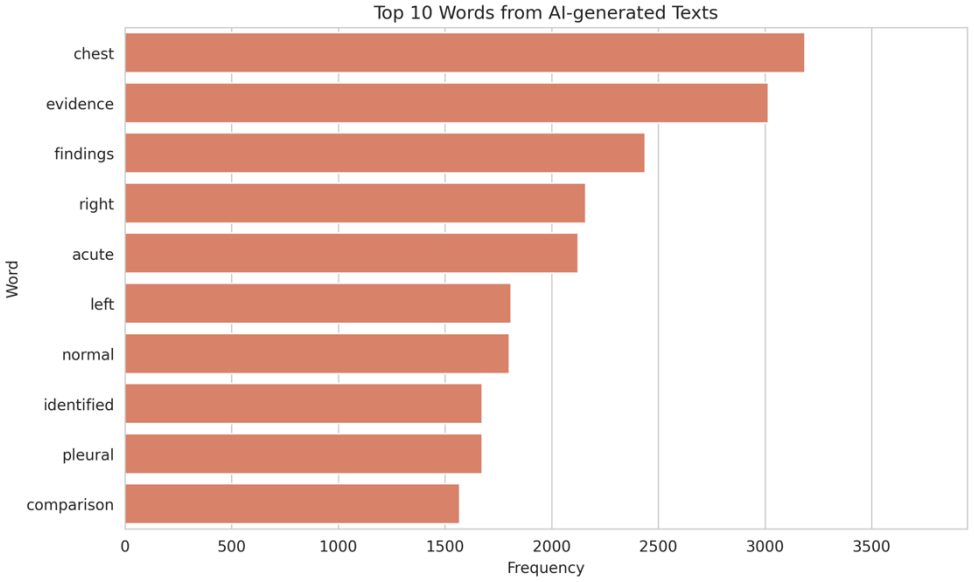

Figure 6. Top 10 most common words in AI-generated texts.

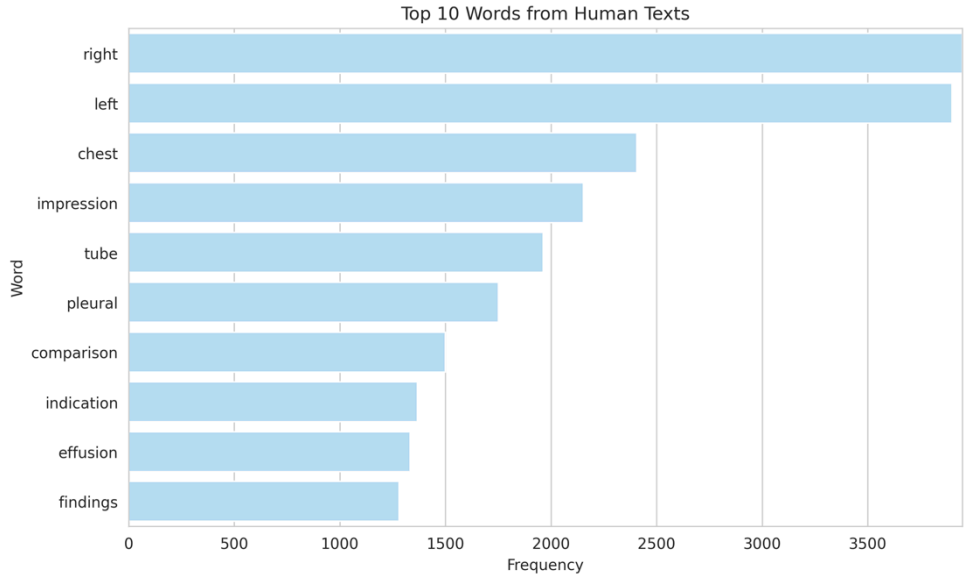

Figure 7. Top 10 most common words in human-written texts.

Appendix B

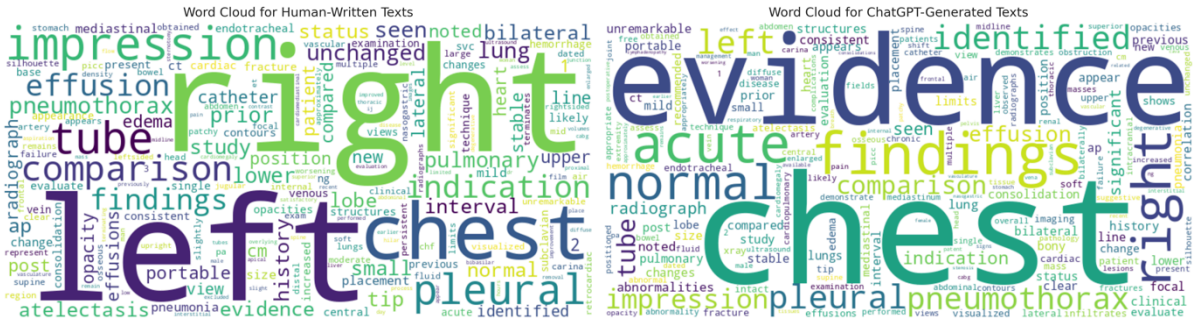

Figure 8. Word clouds for human-written and AI-generated texts.

Appendix C

Similarity: 0.954
Human text: PA AND LATERAL CHEST: [**2155-3-13**].
INDICATION: Cough and hypernatremia.
Comparison: CXR [**2155-3-12**].
The heart is mildly enlarged. The pulmonary vasculature is normal. The thoracic aorta is diffusely calcified and unfolded. There is minimal bibasilar atelectasis. The posterior CP angles are blunted representing either tiny bilateral pleural effusions or pleural scarring. The lungs are otherwise clear. There is minimal right apical pleural scarring. the pulmonary vasculature is within normal limits. The mediastinal and hilar contours are normal. Again seen, is a compression fracture of the 8th vertebra of uncertain chronicity. There is minimal loss of height of the T4 vertebra which may represent some mild compression as well.
IMPRESSION: No evidence of pneumonia.
AI-generated text: PROCEDURE: Chest PA and lateral on [**2172-5-12**].
COMPARISON: [**2172-5-12**] at 10:08. The heart size is normal. The pulmonary vasculature appears unremarkable. There are no abnormal findings within the thoracic aorta. Minimal bibasilar atelectasis is noted, and there are slight blunting of the posterior CP angles, which may suggest either tiny bilateral pleural effusions or pleural scarring. The lungs are clear with the exception of minimal right apical pleural scarring. The pulmonary vasculature is within normal limits. The mediastinal and hilar contours are normal. Additionally, the previously identified compression fracture of the 8th vertebra is again seen, with minimal loss of height of the T4 vertebra, potentially indicating mild compression. Overall, there is no evidence of pneumonia.

Similarity: 0.953
Human text: CHEST RADIOGRAPH PERFORMED ON [**2190-9-28**]
COMPARISON: None.
CLINICAL HISTORY: Acute dyspnea.
FINDINGS: Portable supine AP chest radiograph is obtained. Endotracheal tube is seen with its tip located approximately 5 cm above the carina. NG tube courses into the left upper abdomen. Diffuse bilateral ground-glass opacities, which are most compatible with pulmonary edema. A right pleural effusion is noted, moderate in volume. There may be a tiny left pleural effusion. No pneumothorax is

seen. Heart size difficult to evaluate though is at least top normal. Widening of the mediastinum is likely due to supine portable technique. The imaged osseous structures are intact.
IMPRESSION: Endotracheal and nasogastric tubes appropriately positioned. Pulmonary edema with moderate right and tiny left pleural effusion. Top normal heart size.
SESHa
AI-generated text: PORTABLE CHEST: 20:56
INDICATION: Liver failure, now with hypoxia. PA catheter placed. Portable supine AP chest radiograph obtained at 20:56. Endotracheal tube is positioned approximately 5 cm above the carina. NG tube courses into the left upper abdomen. Diffuse bilateral ground-glass opacities consistent with pulmonary edema are observed. A moderate-sized right pleural effusion is present, and there may be a small left pleural effusion. No pneumothorax is seen. Heart size is difficult to evaluate but appears top normal. Mediastinal widening is likely due to the supine portable technique. Osseous structures are intact. Impression: Appropriate positioning of endotracheal and nasogastric tubes. Pulmonary edema with moderate right and small left pleural effusion. Top normal heart size.

Similarity: 0.946
Human text: INDICATION: 70-year-old female with 1 week of dyspnea now with hypotension.
COMPARISONS: Comparison is made to AP chest radiograph from [**2139-7-15**].
TECHNIQUE/FINDINGS: A single portable AP view of the chest was reviewed. Cardiomegaly is stable. Interval increase in pulmonary vasculature prominence with associated bilateral pleural effusions is concerning for worsening congestive heart failure. No pneumothorax is identified. Stable retrocardiac consolidation is seen and may reflect atelectasis and/or pneumonia. There is no interval change in a right lower lobe ill-defined patchy opacity, which may reflect a developing pneumonia.
IMPRESSION: 1. Interval increase in pulmonary vascular congestion with associated bilateral pleural effusions, indicating worsening CHF. 2. Stable right lower lobe opacity, which could be developing pneumonia. 3. Stable moderate cardiomegaly. 4. Unchanged left lower lobe consolidation.
AI-generated text: AP PORTABLE CHEST, [**2104-5-19**] AT 1652 HOURS.
HISTORY: Pneumothorax with worsening hypoxia. A single AP portable chest radiograph was obtained in a 70-year-old female with a 1-week history of dyspnea and now presenting with hypotension. The cardiomegaly appears stable. There is an interval increase in pulmonary vascular congestion with associated bilateral pleural effusions, indicating worsening congestive heart failure. No pneumothorax is identified. There is a stable retrocardiac consolidation, suggestive of atelectasis and/or pneumonia. Additionally, there is a right lower lobe ill-defined patchy opacity, which remains unchanged and may represent a developing pneumonia. The overall impression includes: 1) interval increase in pulmonary vascular congestion and bilateral pleural effusions consistent with worsening CHF, 2) stable right lower lobe opacity suggestive of a developing pneumonia, 3) stable moderate cardiomegaly, and 4) unchanged left lower lobe consolidation.

Similarity: 0.946
Human text: THREE VIEWS, FOUR RADIOGRAPHS OF THE LEFT KNEE

INDICATION: Evaluate left knee, prior radiograph indicates history of patellar fracture.
FINDINGS:
There is evidence of prior surgery with surgical clips visualized in the prepatellar soft tissues. Soft tissue swelling is also seen in the tissues anterior to the patella and patellar tendon. Irregularity of the inferior aspect of the patella, compatible with prior surgery is evident. Appearance is similar to prior study. Medial and lateral joint spaces are preserved. There is a small joint effusion. No aggressive bone lesion or new fracture is seen.
IMPRESSION: Postoperative findings with evidence of cortical irregularity at the inferior aspect of the patella, presumably representing resected fracture fragment.
DR. [**First Name (STitle) 1388**] [**Initials (NamePattern4) **] [**Last Name (NamePattern4) 1389**]
AI-generated text: INDICATION: CIRRHOSIS AND ENCEPHALOPATHY. SHORTNESS OF BREATH. There are multiple findings on the radiographs of the left knee. Surgical clips are visualized in the prepatellar soft tissues, indicating prior surgery. There is also soft tissue swelling anterior to the patella and patellar tendon. The inferior aspect of the patella is irregular, consistent with previous surgery. The medial and lateral joint spaces appear normal, but there is a small joint effusion present. No aggressive bone lesion or new fracture is seen. Overall, the findings suggest postoperative changes with evidence of cortical irregularity at the inferior aspect of the patella, likely representing a resected fracture fragment.
DR. [**First Name (STitle) 1388**] [**Initials (NamePattern4) **] [**Last Name (NamePattern4) 1389**]

Similarity: 0.944
Human text: TWO VIEW CHEST OF [**2162-12-6**]
CLINICAL INDICATION: Renal failure. Please assess for volume overload.
Comparison is made to previous study of 3 days earlier.
The heart is upper limits of normal in size. The aorta is tortuous. The pulmonary vascularity is normal. There are persistent bilateral small pleural effusions, slightly increased in size in the interval. Skeletal structures reveal diffuse demineralization as well as several compression deformities of indeterminate age.
Tortuosity of the thoracic aorta is noted, with possible component of dilation in the ascending portion.
IMPRESSION: 1) Slight increase in size of small pleural effusions, left greater than right.
2) No evidence of pulmonary edema.
3) Compression deformities in thoracic spine of indeterminate age without older films for comparison.
AI-generated text: AP PORTABLE CHEST [**2169-2-3**] at 10:14 a. m. Comparison is made to previous study of [**2162-12-6**]. The heart is within normal limits in size. The aorta is tortuous. The pulmonary vascularity is normal. There are persistent bilateral small pleural effusions, with the left larger than the right. Skeletal structures show diffuse demineralization as well as several compression deformities of indeterminate age. Tortuosity of the thoracic aorta is noted, with a possible component of dilation in the ascending portion.

IMPRESSION:
1) Slight increase in size of small pleural effusions, left greater than right.

2) No evidence of pulmonary edema.
3) Compression deformities in thoracic spine of indeterminate age without older films for comparison.

Similarity: 0.942
Human text: INDICATION: CABG. Evaluate for pneumothorax after chest tubes have been discontinued.
VIEWS: Upright AP view compared with erect AP view from [**2162-10-6**].
FINDINGS: The right internal jugular central venous line remains in stable and satisfactory position. The mediastinal drains and left-sided chest tubes have been removed. There is no evidence of pneumothorax. There has been interval improvement in the aeration of the left lower lung with residual linear atelectasis present at the left base. The right lung remains grossly clear. The cardiac and mediastinal contours are stable. There is no evidence of overt heart failure. No definite pleural effusions are identified.
IMPRESSION: 1) Improved aeration at the left lung base with residual linear atelectasis s/p chest tube removal. No evidence of heart failure.
AI-generated text: HISTORY: 46-year-old woman status post valve replacement, now with right pneumothorax and chest tube repositioning. TECHNIQUE: AP chest compared to earlier in the day on [**2121-10-31**]. FINDINGS: The right internal jugular central venous line remains in a stable position. The mediastinal drains and left-sided chest tubes have been removed. There is no evidence of pneumothorax. The left lower lung shows improved aeration with residual linear atelectasis at the left base, indicating improvement after chest tube removal. The right lung appears clear. The cardiac and mediastinal contours are unchanged, and there is no sign of heart failure. No definite pleural effusions are observed.

IMPRESSION: The patient shows improved aeration at the left lung base with residual linear atelectasis after chest tube removal. No evidence of heart failure is found.

Similarity: 0.941
Human text: EXAM: Chest frontal and lateral views.
CLINICAL INFORMATION: 65-year-old man with questionable history of chest compression this a.m.
COMPARISON: None.
FINDINGS: Frontal and lateral views of the chest were obtained. There are relatively low lung volumes with mildly decreased expansion of the right lung as compared to the left. Cardiac silhouette is borderline enlarged. The aorta is mildly tortuous. No focal consolidation, pleural effusion or evidence of pneumothorax is seen. There is minimal left basilar atelectasis. Right upper quadrant surgical clips are noted. Mild degenerative changes are seen along the thoracic spine.
IMPRESSION: 1. Low lung volumes with mildly decreased expansion of the right lung as compared to the left. Minimal left basilar atelectasis. No focal consolidation. 2. Borderline-to-mildly enlarged cardiac silhouette.
AI-generated text: INDICATION: Severe pancreatitis, left subclavian line placement. COMPARISON: None.

FINDINGS: Chest frontal and lateral views show no evidence of pneumonia, pneumothorax, or pleural effusion. There is mild decreased expansion of the right lung compared to the left, with relatively low

lung volumes. The cardiac silhouette appears borderline enlarged. Mild tortuosity of the aorta is noted. Minimal left basilar atelectasis is seen. Right upper quadrant surgical clips are evident. Degenerative changes are observed along the thoracic spine.

IMPRESSION: 1. Low lung volumes with mildly decreased expansion of the right lung compared to the left, without focal consolidation or pleural effusion. Mild left basilar atelectasis. 2. Borderline-to-mildly enlarged cardiac silhouette.

Similarity: 0.940
Human text: INDICATION: Status post Roux-en-Y hepaticojejunostomy for benign stricture now intubated, follow up.
COMPARISON: [**2188-1-23**].
TECHNIQUE: Single AP portable supine chest.
FINDINGS: Endotracheal tube terminates 5.2 cm from the carina, at the level of the thoracic inlet. This may be advanced slightly for optimal positioning. Left subclavian venous access catheter terminates in upper SVC in unchanged position. Heart size and mediastinal contours appear unchanged. Since the previous examination, there is increase in bilateral hazy opacity suggesting increased pleural effusions and pulmonary edema. No pneumothorax. Right clavicular pseudoarthrosis again noted. Nasogastric tube terminates in the stomach. A catheter projecting over the left lower quadrant likely represents change in position of the previously seen intraperitoneal drainage catheter.
IMPRESSION: Worsening moderate-to-severe pulmonary edema and bilateral pleural effusions.
AI-generated text: INDICATION: abdominal mass with increased lower extremity edema and possible ascites. Evaluate for DVT and IVC obstruction. FINDINGS: Single AP portable supine chest radiograph demonstrates bilateral pleural effusions and pulmonary edema, which have worsened since the previous examination. The endotracheal tube is positioned 5.2 cm from the carina, at the level of the thoracic inlet, and may be advanced slightly for better placement. The left subclavian venous access catheter remains in an unchanged position, terminating in the upper superior vena cava. The heart size and mediastinal contours are unchanged. There is no evidence of pneumothorax. Additionally, a catheter projecting over the left lower quadrant likely represents a change in position of a previously seen intraperitoneal drainage catheter.

IMPRESSION: Moderate-to-severe pulmonary edema and bilateral pleural effusions have worsened since the previous examination.

Similarity: 0.936
Human text: HISTORY: Status post medial aortic repair check CVL position.
COMPARISON: [**2147-3-1**].
AP SUPINE PORTABLE CHEST: A right IJ central venous catheter is present with its tip in the upper SVC. NG tube courses off the film below the level of the hemidiaphragm. An endotracheal tube is 5.4 cm above the carina. The lungs are hyperinflated. The right lung is grossly clear. There is dense retrocardiac opacity on the left which may represent atelectasis or pneumonia. There does appear to be leftward shift of the heart favoring atelectasis. The left CP angle is excluded. There is no pneumothorax.
IMPRESSION: No pneumothorax. Left lower lobe atelectasis, though superimposed pneumonia is not excluded.

AI-generated text: HISTORY: Pancreatitis. New right IJ placement. AP supine portable chest: A right IJ central venous catheter is noted with its tip at the level of the upper SVC. An NG tube courses off the film below the hemidiaphragm. The endotracheal tube is positioned 5.4 cm above the carina. The lungs are hyperinflated. The right lung appears clear. There is a dense opacity in the retrocardiac region, which could be suggestive of atelectasis or pneumonia. The heart appears shifted to the left, favoring the possibility of atelectasis. The left CP angle is not visualized. No pneumothorax is seen.

IMPRESSION: No evidence of pneumothorax. Findings are consistent with left lower lobe atelectasis, although superimposed pneumonia cannot be excluded.

Similarity: 0.936
Human text: REASON FOR EXAMINATION: Current wound infection, elevated sputum production.
Portable AP chest radiograph compared to [**2157-7-1**].
The tracheostomy tip is 6.5 cm above the carina. The right subclavian line tip, the pacemaker leads, and the direction of the feeding tube are in standard positions.
The heart is moderately enlarged but stable. The bilateral pleural effusion, right larger than left are unchanged with most probably large mediastinal pleural component on the right. The moderate pulmonary edema is also unchanged. Mediastinal widening seen on the last two chest x-rays in comparison to [**6-29**] is most probably due to a enlarged mediastinal veins accompaning pulmonary edema.
IMPRESSION:
1. Unchanged bilateral moderate pleural effusion and pulmonary edema.
2. Slightly widened mediastinum most probably due to enlarged pulmonary veins.
AI-generated text: STUDY: AP chest, [**2121-1-12**].
HISTORY: 57-year-old man with decreased O2 saturations. The tracheostomy tip is 6.5 cm above the carina, and the right subclavian line tip, pacemaker leads, and feeding tube are in standard positions. The heart is moderately enlarged but stable. The bilateral pleural effusion, with the right side larger than the left, remains unchanged, possibly with a large mediastinal pleural component on the right. The moderate pulmonary edema also remains unchanged. The slightly widened mediastinum seen on the previous chest x-rays may be due to enlarged pulmonary veins accompanying the pulmonary edema.
IMPRESSION:
1. Unchanged bilateral moderate pleural effusion and pulmonary edema.
2. Slightly widened mediastinum, possibly due to enlarged pulmonary veins.

Appendix D

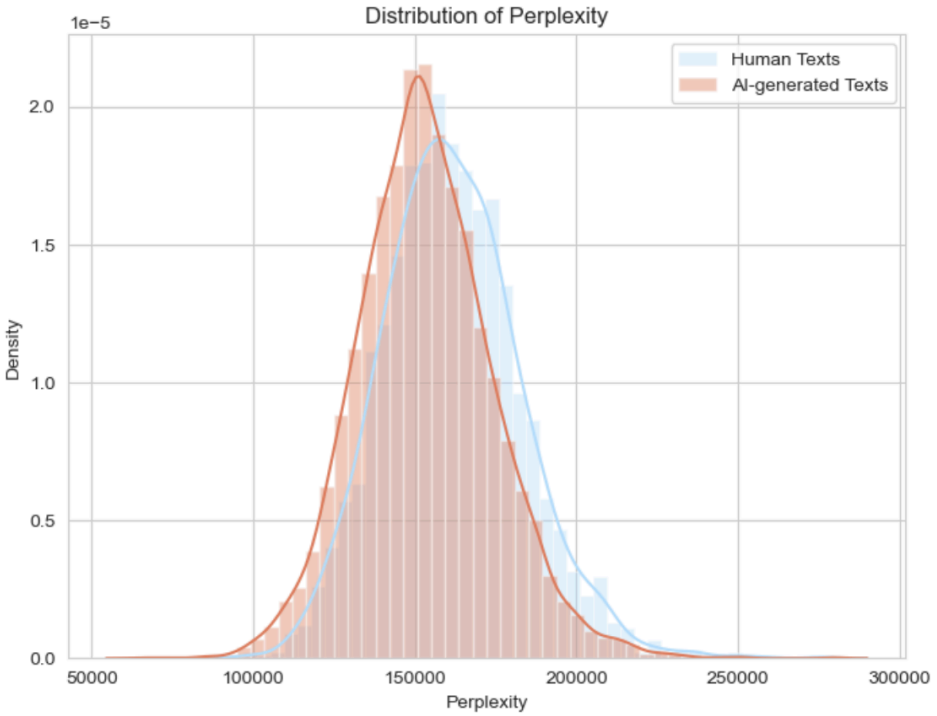

Figure 9. Distributions of perplexity values for pre-trained ELECTRA model.

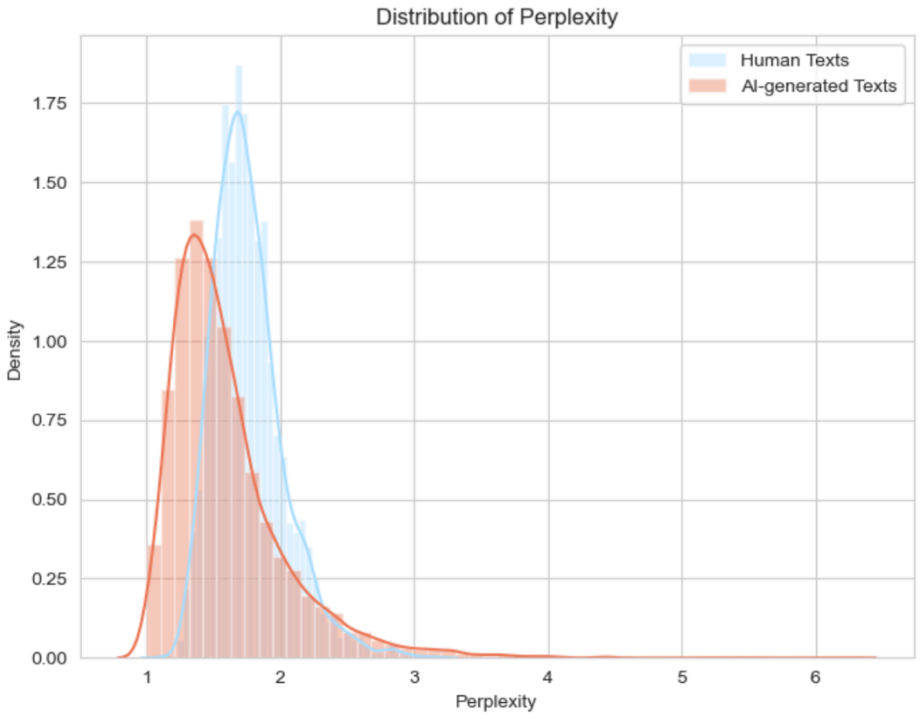

Figure 10. Distributions of perplexity values for pre-trained RoBERTa model.