Detection of AI-generated text

Hemanth Kumar Jayakumar
Master of Computer Science
Rice University
Houston, Texas
hj51@rice.edu

Harshit Khanna
Master of Computer Science
Rice University
Houston, Texas
hk57@rice.edu

***Abstract*—With the rise of Large Language models, the ethical aspects of such text have been at the peak of concern among the Deep Learning community, one among them being the usage of such outputs claiming it to be an individual's effort. Our goal is to solve this problem by providing means to detect if a particular language text(English text in our scenario) is generated by ChatGPT or not.**

I. Introduction

The process of recognizing and differentiating text created by AI language models, like Chat GPT(Generative Pretrained Transformer), from text produced by humans, is referred to as the detection of ChatGPT-generated text. OpenAI's Chat GPT is a popular large-scale language model that generates human-like text answers to questions using deep learning and statistical techniques. It has found widespread use in a variety of applications, including chatbots, virtual assistants, and language translation services.

The problem can be widely looked at in 2 different perspectives: White box methods and Black Box methods, where a black box method denies access to the model itself, removing the presence of probabilities of scores. This provides us with a unique problem, which can be adapted towards various other LLMs as a general methodology to perform analysis of data generated by that model, either to help create a better model or to improve the societal impacts of the model itself, providing a safety barrier due to these abnormalities.

II. Problem Statement

Given a text, this project aims to detect if that particular text was generated by ChatGPT or not.

This problem can be narrowed down by defining the ChatGPT itself as a black-box model, restricting access to the model, and only providing an oracle that provides an output answer given a question. This work touches upon rigorous analysis aiming to providing insights into the types of text that is being generated by ChatGPT and therefore inducing the ability to classify a text to either a human answer or an answer generated by ChatGPT with ease. Further details of the analysis has been explained in the Hypothesis section.

III. Literature Review

The literature review in the paper "The End of Online Exam Integrity" covers a range of studies and research papers that have explored the use of technology in cheating on exams, including previous studies on chatbots used for exam cheating. The authors also review the literature on natural language processing (NLP) techniques used for detecting cheating in written text. The review provides a comprehensive overview of the current state-of-the-art methods for detecting cheating in written text and highlights the potential of GPT-based models in cheating. Tang et al.[5], from the D2K lab at Rice University, discussed several techniques that can be used to do this particular task, ranging from statistical analysis to BeRT-based approaches.

Guo et al.[4] provide statistical input on how ChatGPT answers differ from the usual human text. For example, ChatGPT answers in an objective fashion whereas humans tend to provide a subjective answer. But these methods do not provide a measurable metric to identify these patterns and may change with upcoming Large Language Models. Mitchell et al[3]'s DetectGPT provide an oracle-based approach where the text is perturbed in different ways and the probability is analyzed to detect if the given text is from the model or not.

Several pipelines have begun to emerge to dodge these existing ChatGPT detectors as well. One prominent approach is discussed in GPTZzzs which discusses how replacing words with synonyms, interchanging passive and active voice of a sentence, etc convinces the detector to make the ChatGPT generated text to appear as human written text. This can be however handled by deep learning approaches since the embedding is trained to identify the semantics of the text, through the different perturbations since the embedding layer is intuitively a summation over the vector representing words, eliminating most automated changes from script incapable of bypassing such a detector.

IV. Hypothesis/Proposal

We propose 2 experiments to solve this problem,

1) To analyze large datasets and identify patterns that exist within the AI text which is not present in Human written text, such as TF-IDF, and N-gram histogram comparisons, to use this pattern for detection of such text.
2) To design a transformer-based neural network to build a basic binary classifier over ChatGPT-generated text and Human-generated text to detect which dataset a text is more probable to be related to.

Our work begins with the dataset released by HelloSimpleAI [4] consisting of various questions mined from forums, containing human answers and fed into ChatGPT API to obtain its outputs across a variety of temperature parameters, providing multiple answers from humans and ChatGPT for each question. The sources of these questions range from Reddit, medicine, finance, etc. providing a diverse set of vocabulary and content allowing strong analysis of the text.

We further aim to attempt the generalization of these models to ensure that as newer and bigger LLMs are released, these models can easily adapt by a bit of finetuning to detect these as well.

V. Experimental Settings

We conduct a total of 3 unique experiments, neural networks and statistics combined, to allow us a better perspective towards finding the biases existing between human and ChatGPT-generated texts and plot our results for the same. Each of these experiments was designed and conducted to work in a resource-constrained environment and therefore does not cover the utilization of higher computational power for better results from them, leaving it for the future scope of this paper.

TF-IDF; In the first experiment, we used the ChatGPT vs human text dataset to develop a model that can distinguish between ChatGPT-generated and human-generated text. We cleaned and transformed the data, added classes for Chat-GPT vs Human text, split it into training, and test sets, and used TF-IDF for feature extraction. We then used three approaches to build text classification model:

1) Multinomial Naive Bayes (Multi-NB)
2) Random Forest
3) XG-Boost

First, we created a pipeline that consists of two steps. The first step is to transform the text data into numerical features using TF-IDF. We then trained the model using the transformed data.

Next, we fit the model using the training data so that the model can learn the relationships between the features and the target labels. Then, we use the trained model to predict the labels of the test data and store the predicted labels.

Finally, we evaluate the performance of the model by comparing the predicted labels with the actual labels. The generated report includes precision, recall, F1-score, and support for each class in the classification problem, as well as an overall accuracy score. Code

Deep Learning approach; We use data scraped on questions gathered from Reddit, released by HelloSimpleAI[4] containing multiple human and ChatGPT answers for a question. We finetune a BERT model(bert-base-uncased) sampling an answer at random per epoch over a classification head. This approach provides a consistent 68% accuracy.

BERT was chosen for this due to its intrinsic ability to understand human text, which was used to model the data from ChatGPT. The results of this approach being low are mainly due to the absence of quality data providing biases towards the 2 types of data. We also believe that the model trained on human text may not be capable to identify text generated by an AI since it was entirely trained over human text.

We also mention that while we were not able to utilize the different samples of answers present, we were not able to leverage this under any of our methods and remains future scope in this domain.

The code for these experiments are as follows: Code

N-Gram Analysis; We calculated the frequency of n-grams, from 1 through 6, and plotted histograms to show our findings for the same. The data was initially filtered to remove redundant spaces and punctuations while maintaining word contractions which provides crucial information to our analysis. Since NLTK does not completely support the handling of word contractions(apostrophes in sentences), even with the help of GloVe's tokenizer, we create our own filter to handle these issues.

Histograms were plotted for the top 10 occurring n-grams for each length and compared with its counterpart distribution(top 10 human texts with ChatGPT text and vice-versa). The results are jotted in the Results section with possible analysis interpreted from these plots.

Additional histograms were plotted to showcase the identified "emphasis words" which provide a clear difference from the n-gram analysis over the 2 types of text, allowing identification of the source of text based on their presence.

All these experiments are pipelined such that it is possible to swap out the data containing ChatGPT texts to any other AI-generated dataset to explore

several other upcoming and existing LLMs and their biases in the English text.

The code for these experiments are as follows: Code

VI. Results and Conclusion

1) The Chat-GPT vs Human text classification task can be challenging due to the subtle differences between machine-generated and human-generated text. However, by using a combination of TFIDF and machine learning methods, we were able to achieve decent results. In our experiments, we evaluated the performance of three different methods on this task: MultinomialNB, Random Forest and XG-Boost. The results showed that all three methods achieved good Accuracy, Precision, Recall, F1 Score and Support values on this task which are listed in Table 1. Overall, these results suggest that using a combination of TFIDF and machine learning methods can be effective for the Chat-GPT vs Human text classification task. However, further experimentation and fine-tuning may be required to achieve even better results.
2) The Deep Learning approach provided 68% accuracy with limited training, finetuned over less than 10 epochs, thus showing that a black box approach is definitely a fruitful method for this classification task. Moreover, Bert was able to easily transfer semantics of the english language learned from human text to ChatGPT-generated text showing that both these texts are similarly interpreted by deep learning models. This also implies that better accuracy can definitely be achieved given a large number of parameters and a dataset consisting of ChatGPT-generated text.
3) The analysis with n-gram distribution provided numerous results which are portrayed below, and can be visualized in Figures 3 and 4 at the end of this paper.
    a) The 1-gram model shows different 1-gram words occurring in both datasets and comparing them. As most of these bars primarily show minute differences in their use, they do not provide any concrete evidence for detection. This is due to the size of the corpora used for analysis, as well as the random sampling done by the ChatGPT model allowing a wide range of words to be used.
    b) The 2-gram shows similar charts, not providing any evidence of bias towards the choice of words, etc.
    c) The 3-gram and 4-gram plots show interesting biases towards phrases that ChatGPT uses much more than seen in the human-text dataset. The phrase ”it is important to note”, ”a variety of”, ”a good idea to”, and ”on the other hand” is commonly seen within these plots showing bias towards the recurrent use of this phrase throughout the corpora when answering these questions. Whereas certain phrases are evidently seen within the human text that is not visible throughout the ChatGPT corpora. For example, ”Let me know if ...”, ”hope I have answered...”, ”it has to do with” are several phrases commonly used by texts written on forums to imply a chain of messages is plausible and encouraged whereas ChatGPT fails to adapt these to its repository.
4) Outside these n-gram plots, we identify ”emphasis words” that are widely used by humans which ChatGPT fails to capture. One of these words include the set of words called ”contractions”, as shown in Figure 1. This analysis shows that human text tend to contain more of these contractions as compared to ChatGPT's texts. Another one of these include subjectiveness in the text, pointing out facts from one's perspective, which usually starts with one of the words in Figure 2. These words show a slight bias on human text containing higher frequency compared to ChatGPT texts.

VII. Challenges Encountered

There were several challenges faced during this experiment and their supporting works, noted below:

1) The lack of literature provides a huge stopping point towards designing the definitive architecture to detect from an AI text. Therefore, most of these experiments are based on limited literature, with more logical reasoning or borrowed from other literature.
2) Deep Learning approaches, defined in various literature provide a strong base for this problem. However, training Transformer-based networks take a heavy toll on computational resources, thus limiting our experiments in this field of work.
3) Lack of data provides a certain bias towards one particular dataset used throughout the experiments. Although API wrappers exist to mine data from the models, they do not scale enough and have restrictions due to network traffic controls providing limited space in working with large datasets. However, the dataset used in this paper is large enough to provide strong support for the results calculated over a good variance of data.
4) Experiments were conducted to finding an upper bound towards parameter requirement for Deep Learning based approaches. But they were put in a pause due to lack of resources for even finetuning large transformer

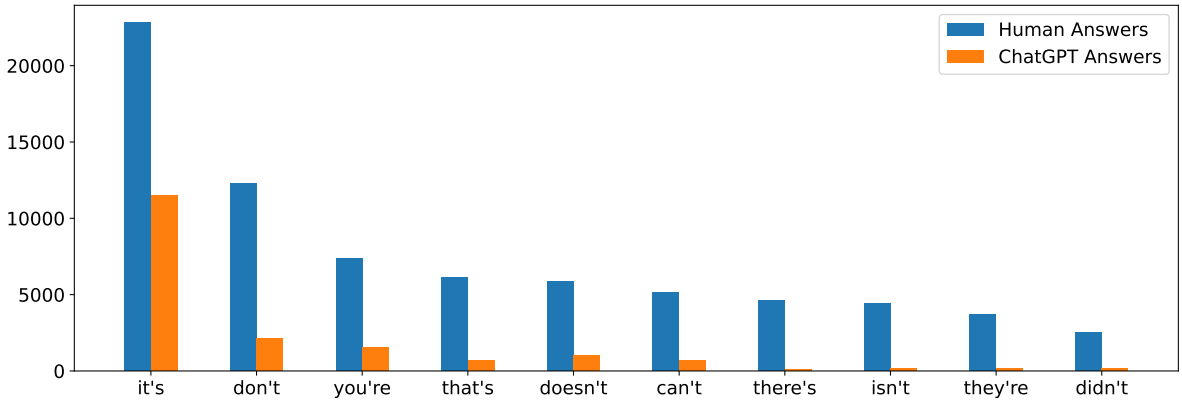

Fig. 1. Histogram of emphasis words chosen from contractions

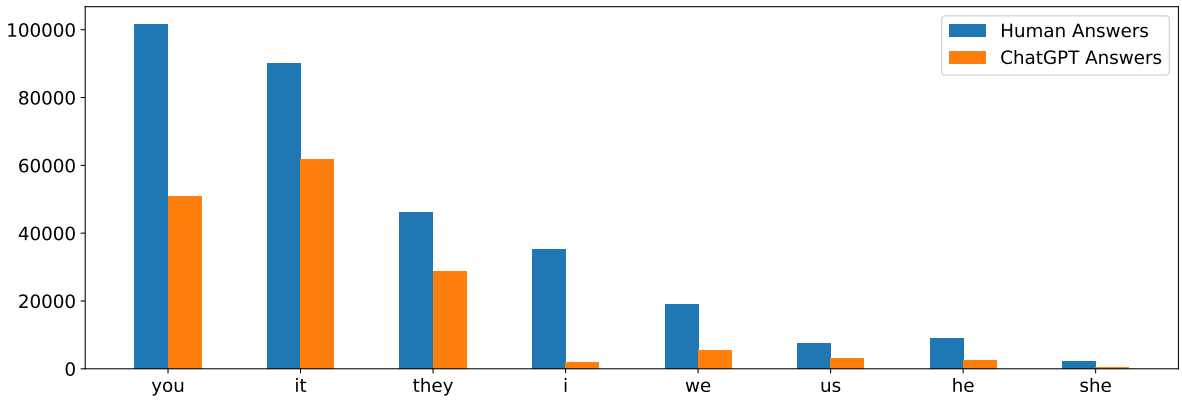

Fig. 2. Histogram of emphasis words showing subjectiveness

Model	Accuracy	Precision	Recall	F1score	Support
MultinomialNB	88%	92%	82%	87%	4880
Random Forest	94%	95%	94%	94%	4880
XGBoost	91%	90%	93%	91%	4880

TABLE I
RESULTS FOR TF-IDF EXPERIMENT

models towards these approaches. This would be a fruitful question to be posed towards this problem field providing a strong guarantee of existence of deep learning models capable of solving this problem.

VIII. IMPACT

The possibility of using ChatGPT-generated text to cheat or get around exam security safeguards is becoming more and more of a worry as remote proctoring and online examinations become more common. By detecting such text, test administrators can more effectively monitor exam replies and guarantee their validity by identifying Chat GPT-generated text. Text created by Chat GPT may also be used to fabricate stories, pose as people or organizations, or even produce malevolent information. Given that it might look to have been produced by a person and that it can be shared online, this kind of material can be difficult to identify. Finding such language can assist stop the spread of dangerous content and false information, which can affect public opinion, create fear, and even affect election results

REFERENCES

[1] Susnjak, T. (2022). ChatGPT: The End of Online Exam Integrity? arXiv, 2212.09292. Retrieved from https://arxiv.org/abs/2212.09292v1

[2] Hiriyannaiah, S., Srinivas, A. M. D., Shetty, G. K., G. M., S., & Srinivasa, K. G. (2020). A computationally intelligent agent for detecting fake news using generative adversarial networks. Hybrid Computational Intelligence. Academic Press. doi: 10.1016/B978-0-12-818699-2.00004-4

[3] Mitchell, E., Lee, Y., Khazatsky, A., Manning, C. D., & Finn, C. (2023). DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature. arXiv, 2301.11305. Retrieved from https://arxiv.org/abs/2301.11305v1

[4] Guo, B., Zhang, X., Wang, Z., Jiang, M., Nie, J., Ding, Y., ...Wu, Y. (2023). How Close is ChatGPT to Human Experts? Comparison Corpus, Evaluation, and Detection. arXiv, 2301.07597. Retrieved from https://arxiv.org/abs/2301.07597v1

[5] Tang, R., Chuang, Y.-N., & Hu, X. (2023). The Science of Detecting LLM-Generated Texts. ResearchGate. doi: 10.48550/arXiv.2303.07205

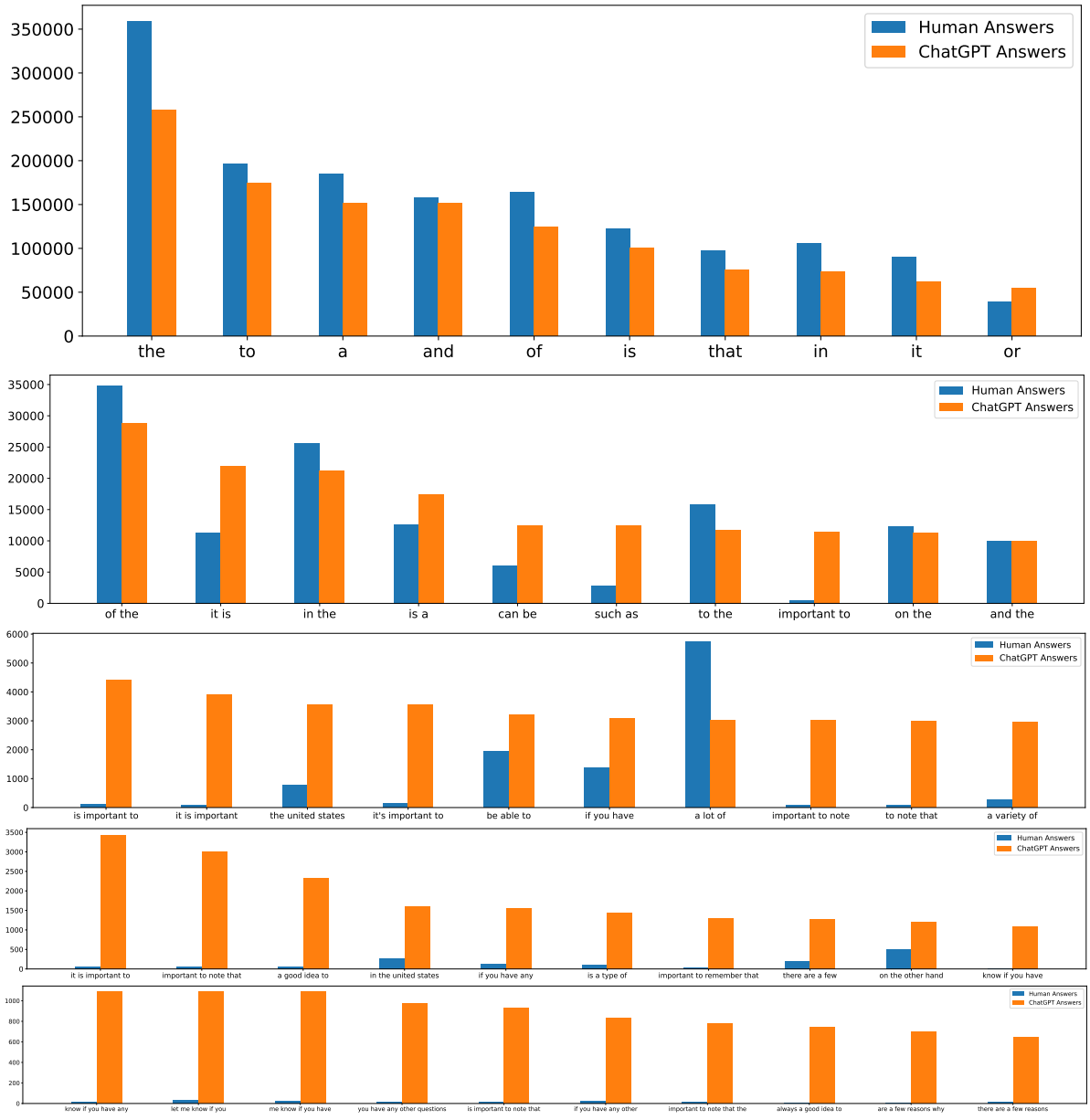

Fig. 3. Human vs ChatGPT answers on highest ChatGPT n-gram frequencies

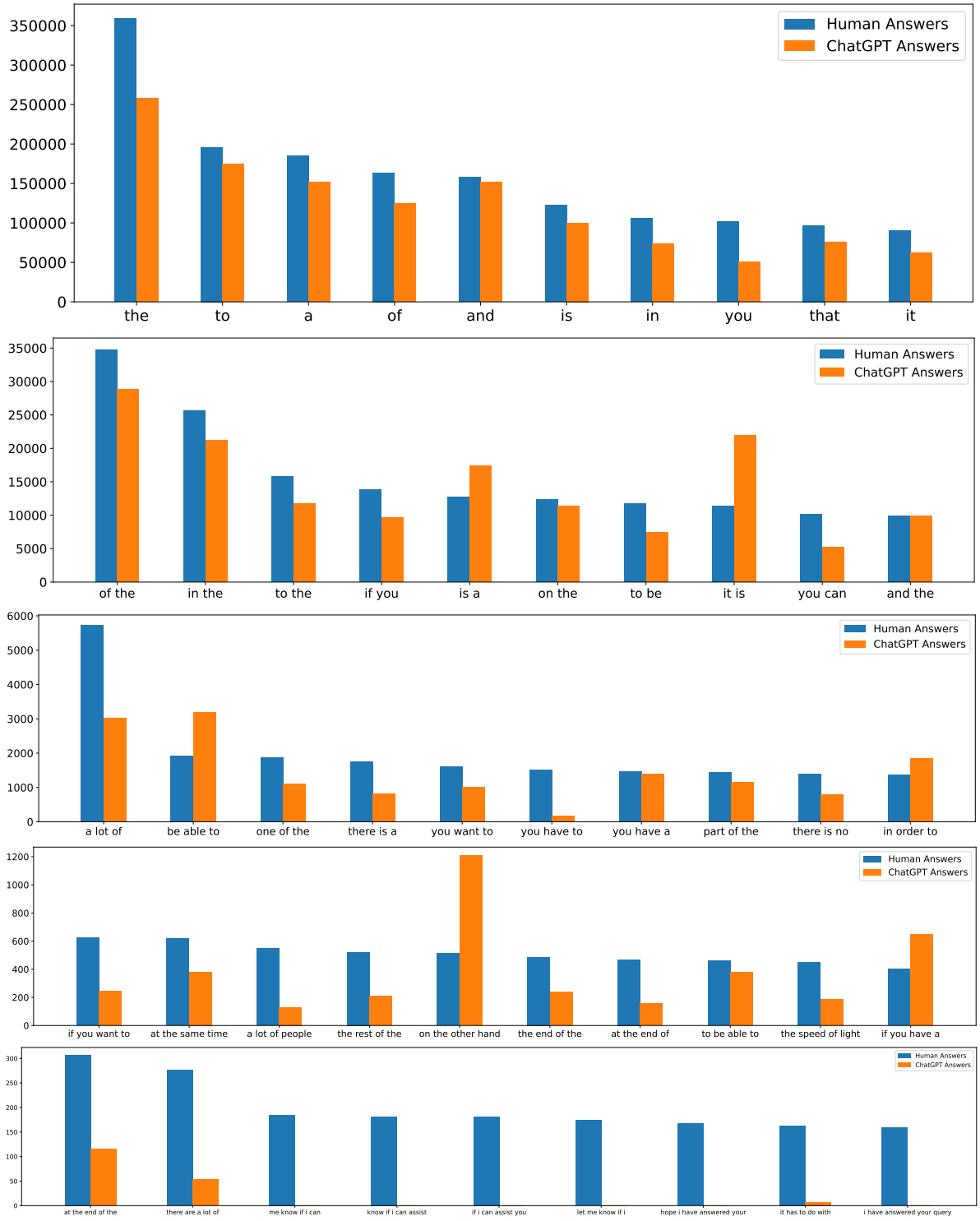

Fig. 4. Human vs ChatGPT answers on highest Human n-gram frequencies