Fighting Fire with Fire: Can ChatGPT Detect AI-generated Text?

Amrita Bhattacharjee
School of Computing and AI
Arizona State University
abhatt43@asu.edu

Huan Liu
School of Computing and AI
Arizona State University
huanliu@asu.edu

ABSTRACT

Large language models (LLMs) such as ChatGPT are increasingly being used for various use cases, including text content generation at scale. Although detection methods for such AI-generated text exist already, we investigate ChatGPT's performance as a detector on such AI-generated text, inspired by works that use ChatGPT as a data labeler or annotator. We evaluate the zero-shot performance of ChatGPT in the task of human-written vs. AI-generated text detection, and perform experiments on publicly available datasets. We empirically investigate if ChatGPT is symmetrically effective in detecting AI-generated or human-written text. Our findings provide insight on how ChatGPT and similar LLMs may be leveraged in automated detection pipelines by simply focusing on solving a specific aspect of the problem and deriving the rest from that solution. All code and data is available at `https://github.com/AmritaBh/ChatGPT-as-Detector`.

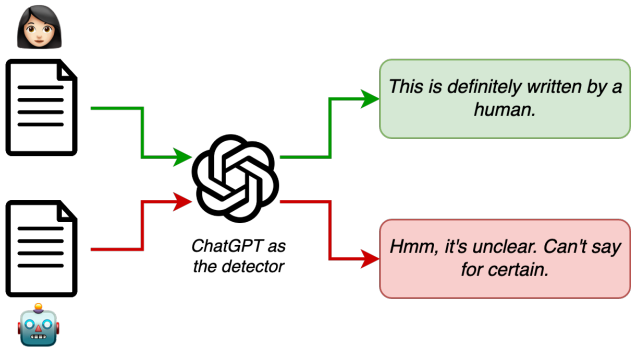

Figure 1: We use OpenAI's ChatGPT as a detector to distinguish between human-written and AI-generated text.

1. INTRODUCTION

Recently there have been incredible advancements in large language models (LLMs) that can generate high quality human-like text, with capabilities of assisting humans on a variety of tasks as well. Larger and more expressive models are released to the public frequently, either as public-facing APIs with no access to the model parameters (such as OpenAI's ChatGPT or GPT3.5 family of models [27; 6]) or often with fully open source access (such as LLaMA [36]). Alongside the numerous ways in which these LLMs can aid a human user, act as an assistant and thereby improve productivity, these models can also be misused by actors with malicious intent. For example, malicious actors may use LLMs to generate misinformation and misleading content at scale and publish such content online [32], create fake websites for ad revenue fraud [30], etc. Apart from these malicious use cases, inexperienced users may overestimate the capabilities of these LLMs. The fact that ChatGPT can confidently spew factually incorrect information, yet be fluent and cohesive in the syntax and grammar [2], can fool newer users and mislead them. Users may use ChatGPT to write essays or reports, expecting factuality but then be penalized when flaws are evident. Especially problematic is when people use these models for high-stakes tasks, without realizing the shortcomings of these models, and eventually face dire consequences [5]. Given the accessibility and ease of use of such models, more and more people are using these models in their daily life, perhaps without realizing the nature of the text that is generated, often mistaking fluency and confidence as truthfulness. Therefore, in this work, we focus on the task of distinguishing between human-written and AI-generated text.

Detection of AI-generated text is very challenging [18], with recent work [31] demonstrating that this challenge is only going to get exacerbated with newer, larger, more capable LLMs. There are existing detection methods, such as feature-based classifiers [20], methods that differentiate human and AI text using statistical measures [14; 26], along with methods that use fine-tuned language models to classify text as human or AI [38; 33]. In this work, however, we want to investigate the capability of ChatGPT (GPT3.5, and the more advanced GPT-4) to differentiate between human and AI-generated text. ChatGPT has been shown to perform well on a variety of NLP and NLU tasks [28], with newer versions of the models being able to use human prompts better than older ones. In order to probe the performance of ChatGPT on detecting AI-generated text, we propose to investigate the following research question:

Can ChatGPT identify AI-generated text from a variety of generators?

The rest of the paper is organized as follows: Section 2 provides a brief overview of LLMs and how they work, Section 3 elaborates on our experiments and experimental settings. Section 4 presents and discusses our results. Section 5 grounds our work in the context of related work, and finally, Section 6 concludes with a discussion of future directions.

2. BACKGROUND: LARGE LANGUAGE MODELS

Large language models (LLMs) are deep neural networks capable of modeling natural language. Most recent LLMs are built using transformer-based architectures, and trained on massive internet-

scale corpora of data. Broadly, LLMs can be of two types: (i) autoregressive (or causal) language models, and (ii) masked language models. Autoregressive LLMs (such as the GPT family of models) are trained to predict the next word or token, given the previous token sequence in a sentence. Formally, at time step t, the model samples a token from the distribution $p(x_t|x_1, ..., x_{t-1})$, based on some pre-defined sampling strategy, and forms the text sequence one token at a time. Sampling strategies may be greedy, top-k, nucleus sampling [19], etc., with the latter two being used commonly in recent LLMs. Masked language models (such as the BERT family of models [11]) are trained on cloze test tasks. Given an input text sequence, $k\%$ of the tokens are masked using a special [MASK] token, and the LLM is trained to predict the tokens in these masked locations in the sequence. Hence these models are bi-directional, unlike autoregressive LLMs. Due to the bi-directional attention mechanisms in masked language models, these are better at natural language understanding tasks, while autoregressive GPT-style models are good at natural language generation [25].

3. CHATGPT AS AI-TEXT DETECTOR

Recent work have used and evaluated ChatGPT for a variety of natural language tasks, annotations of data, few-shot classification settings, and even reasoning and planning (see Section 5). In this work, we want to investigate whether ChatGPT can be used as a detector to identify AI-generated text. ChatGPT was trained on large amounts of text data and we would want to leverage the summarized information or understanding that ChatGPT possesses to try to identify human-written vs. AI-generated text.

Datasets

We use AI-generated text from the publicly available TuringBench dataset [37], which comprises news article style text from 19 different generators. The full list of these 19 generators are: {GPT-1, GPT-2_small, GPT-2_medium, GPT-2_large, GPT-2_xl, GPT-2_PyTorch, GPT-3, GROVER_base, GROVER_large, GROVER_mega, CTRL, XLM, XLNET_base, XLNET_large, FAIR_wmt19, FAIR_wmt20, TRANSFORMER_XL, PPLM_distil, PPLM_gpt2}. Model sizes for each of these 19 generators in provided in Table 1.

For the human-written articles, we use the human articles from the TuringBench dataset. These are news articles from CNN and The Washington Post.

Experimental Setting

We use ChatGPT (gpt-3.5-turbo model endpoint) with version as of June 13, 2023 and GPT-4 with version as of July 12, 2023 as the detector[1]. We experiment with a variety of prompts (as described in the next section) and finally select an appropriate prompt for classifying each news article using ChatGPT. We set the temperature parameter to 0 to ensure minimal variability in the output since we are dealing with a classification task. For each input article, we process the text output produced by ChatGPT or GPT-4 to flag it as one of the three labels: ['human-written', 'AI-generated', 'unclear']. For all ChatGPT experiments, we use the test split of the datasets, which contain around $2,000$ articles. For experiments with GPT-4 as the detector we use the first 500 samples of this test set, due to rate limit constraints on GPT-4.

[1]In this paper, we use the terms 'ChatGPT' to refer to the GPT-3.5 model

Model	# of Parameters
CTRL	1.6B
FAIR_wmt19	656M
FAIR_wmt20	749M
GPT1	117M
GPT2_small	124M
GPT2_medium	355M
GPT2_large	774M
GPT2_xl	1.5B
GPT2_pytorch	344M
GPT3	175B
GROVER_base	124M
GROVER_large	355M
GROVER_mega	1.5B
PPLM_distil	82M
PPLM_gpt2	124M
Transformer_xl	257M
XLM	550M
XLNet_base	110M
XLNet_large	340M

Table 1: Model sizes for the 19 generators in the TuringBench dataset.

Choice of Prompt

Based on some preliminary experiments, we notice that the response from ChatGPT including 'human', 'AI', or 'uncertain' labels for the input text depends significantly on the prompt used. For ease of experimentation and evaluation, we wanted to constrain ChatGPT's responses by using the following prompt for the input text `passage`:

> Task: Identify whether the given passage is generated by an AI or is human-written. Choose your answer from the given answer choices.
> Answer choices: ["generated by AI", "written by human", "unsure"]
> Passage to identify: `<passage>`

But we observe that with this kind of constrained prompt, ChatGPT gets confused. Not only does it fail to generate answers following the given instruction, but it also misclassifies text that it previously classified properly with a simpler prompt. It also provides incorrect labels for inputs that it was previously unsure of.

Hence we revert back to a simpler prompt used in our preliminary experiments:

> 'Is the following generated by an AI or written by a human: `<text>`.'

where `<text>` is the main body text from a human-written or AI-generated news article from our evaluation datasets.

4. RESULTS AND DISCUSSION

In this section we elaborate our main experimental results with ChatGPT, with GPT-4, along with a discussion of the results, followed by some additional experiments.

ChatGPT as Detector

We show the detection performance of ChatGPT (i.e., GPT-3.5) on text from the 19 generators in Figure 2. We see that for the majority of the generators, ChatGPT can identify AI-generated text less than 50% of the samples, and has a very high number of false negatives,

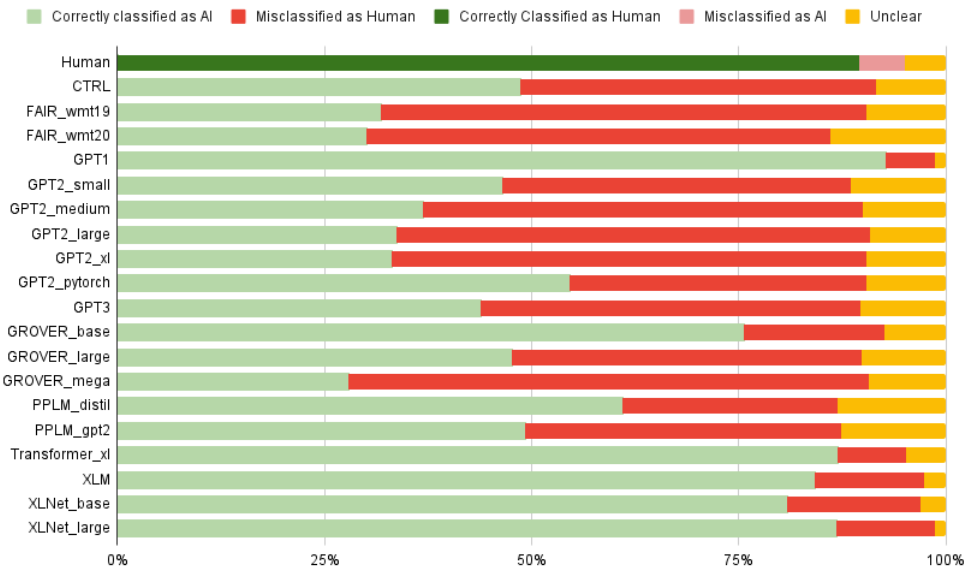

Figure 2: Performance of ChatGPT on texts generated by each of the LLMs in the TuringBench dataset, alongside performance on human-written text from TuringBench (bar at the top).

i.e., AI-generated articles misclassified as human-written. The only satisfactory performance we see is for GPT-1 with around 90% samples correctly identified, and Transformer_xl, GROVER_base, XLM, and the XLNet models with around 75% articles correctly identified as AI-generated. Therefore, our experiment demonstrates that ChatGPT is unable to identify AI-generated text across a variety of generators. We note that even if we flip the label output by ChatGPT and consider that as the final label, i.e. changing all the misclassified 'human-written' labels to 'AI-generated' and the correctly-classified 'AI-generated' labels to now misclassified 'human-written' ones, we do not gain much in performance either, if at all. This is because the proportion of correctly classified and misclassified samples are similar for most of the generators, along with a significant fraction of samples still in the 'uncertain' category.

Next, we investigate ChatGPT's performance on the human-written articles from TuringBench, and we show the distribution of the output labels in Figure 2. Interestingly, we see that for human-written articles, ChatGPT performs significantly better at identifying that the text is human-written, with very few samples being misclassified or even labeled as 'uncertain'.

For the AI-generated texts, we see ChatGPT misclassifies a large fraction of these as human-written. To dig deeper into this phenomenon, we look into how the fraction of false negatives varies with the model size, within a specific model family. Figure 4 shows the fraction of misclassified samples with respect to the different GPT variants with GPT-1 [29] being the smallest and GPT-3 [6] being the largest. We see the percentage of misclassified samples increases with an increase in model size, except for GPT-3, implying that the generation quality becomes more 'human-like' with an increase in the number of model parameters. The discrepancy with GPT-3 having fewer false negatives, even though it is the largest model in our evaluation, seems to be a dataset issue since we see uncharacteristic performance on GPT-3 data even with a fully-supervised classifier. We see a similar trend for the GROVER language model [38] (Figure 5), across three of its size variants: base, large and mega. Similar to the GPT model family, we posit this behavior is due to the text quality becoming better as the model size increases and the models become more expressive. This is also consistent with the performance of other detection methods on these variants of GROVER [38].

GPT-4 as Detector

Similarly, we show the performance of GPT-4 on the 19 generators in Figure 3. We see that for all of the generators, GPT-4 has very good performance, correctly identifying about $97 - 100\%$ of all AI-generated text from the generators. Almost all the text samples are classified as AI-generated, including the human-written texts from TuringBench (top bar in Figure 3). GPT-4 struggles to identify human-written text, and misclassifies over 95% as AI-generated. This would imply that GPT-4 is unable to differentiate between human-written and AI-generated text, for the TuringBench dataset. Interestingly, there are little to no articles (both human and AI-generated) for which GPT-4 outputs an 'unclear' response.

Comparison and Discussion

For the main experiments involving the benchmark TuringBench dataset, we see varied performance between ChatGPT and GPT-4. While there is more *variability* in ChatGPT's performance, GPT-4 tends to label everything as 'AI-generated'. Furthermore, we see a huge difference in the fraction of samples that ChatGPT labeled as 'unclear' vs. the fraction GPT-4 labeled as 'unclear'. This implies that GPT-4 is somehow more confident, even when its predictions are wrong, as in the human-written articles. Hence, these predictions are highly *unreliable*. The degenerate performance of GPT-4 (i.e., labeling everything as one-class, in this case, 'AI-generated') is somewhat unexpected, given the public perception that GPT-4 is better, and more capable than it's previous counterpart ChatGPT. However, recent work has shown empirical evidence that GPT-4's performance might actually be deteriorating over time [8]. This might be due to OpenAI's updates to the GPT-4 model, in order to prevent harmful generations and from people misusing the model.

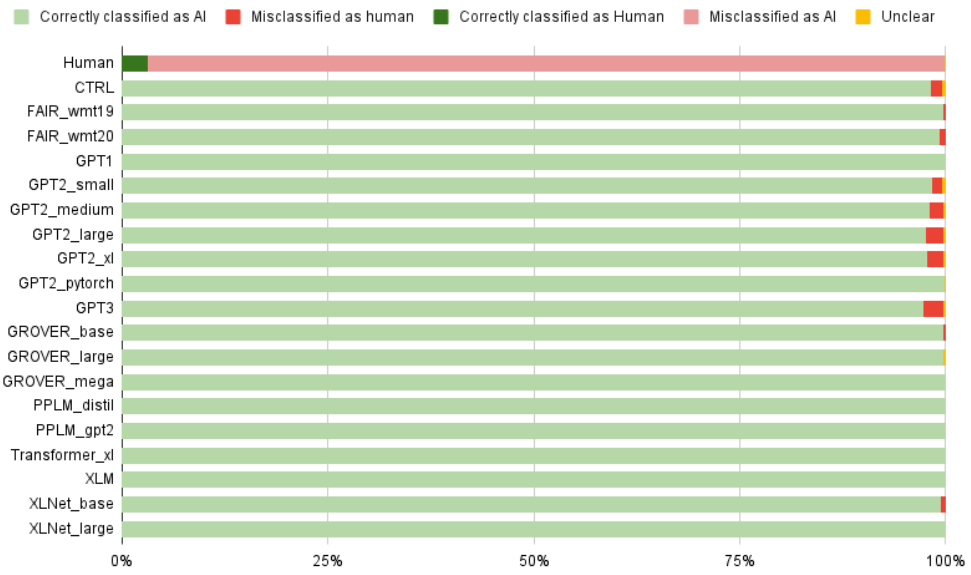

Figure 3: Performance of GPT-4 on texts generated by each of the LLMs in the TuringBench dataset, alongside performance on human-written text from TuringBench (bar at the top).

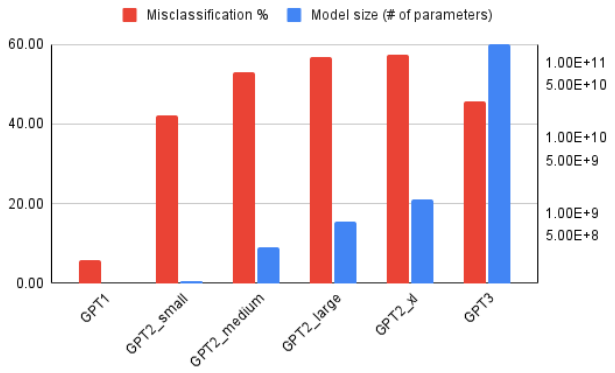

Figure 4: Percentage of AI-generated text misclassified as human-written for different model sizes of GPT. Left-axis: Misclassification %; Right-axis: Model size in log-scale.

Similar to the drift analysis in [8] we used the June 2023 version of ChatGPT and GPT-4 in our experiments, thereby revealing consistent performance degradation of GPT-4, as shown by the authors in [8].

Additional Experiments

Performance on Human-written text from other sources

To test whether ChatGPT and GPT-4's outputs are sensitive to the specific styles, tones, topics, and other features of the human-written text from one specific dataset, we also use the human-written articles from the following datasets:

1. NeuralNews [34]: This dataset consists of human written articles and equivalent articles generated by Grover. For our experiments, we use the human split of the dataset. These are news articles from The New York Times.

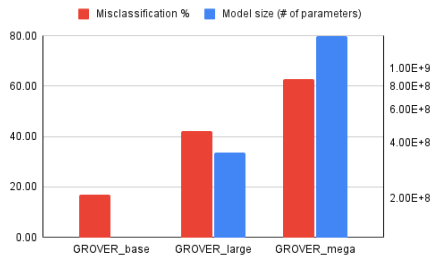

Figure 5: Percentage of AI-generated text misclassified as human-written for different model sizes of GROVER. Left-axis: Misclassification %; Right-axis: Model size in log-scale.

2. IMDb[2]: This dataset comprises of movie reviews, with binary sentiment labels and is originally intended for a sentiment classification task.

3. TweepFake [13]: This dataset comprises Tweets written by humans and also deepfake Tweets. For our purposes, we only use the human-written Tweets.

We show performance of ChatGPT and GPT-4 on these different tyes of human articles in Figure 6 and Figure 7 respectively. For ChatGPT as the detector (Figure 6), majority of human-written text from TuringBench, NeuralNews, IMDb and TweepFake are correctly identified as human-written. There is a significant protion of human-written text labeled as 'unclear' and a much smaller fraction ($\sim 4\%$ and $\sim 6\%$ for TuringBench and IMDb, respectively) misclassified as AI-generated. Hence, we can conclude that the performance of ChatGPT in identifying human-written text is consistent across different sources and styles of human-written text data.

[2]https://huggingface.co/datasets/imdb

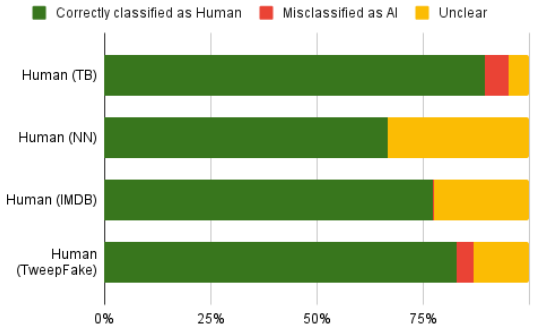

Figure 6: ChatGPT as detector: Distribution of correctly classified and misclassified samples for human-written text from four datasets: {TuringBench, NeuralNews, IMDb, and TweepFake}

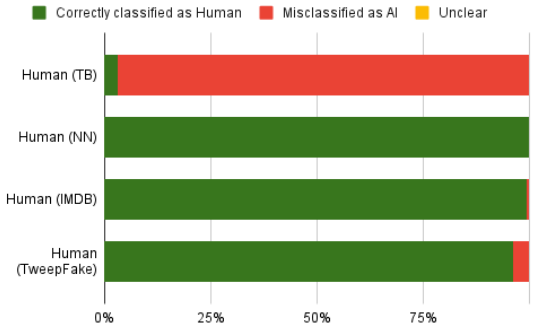

Figure 7: GPT-4 as detector: Distribution of correctly classified and misclassified samples for human-written text from from four datasets: {TuringBench, NeuralNews, IMDb, and TweepFake}

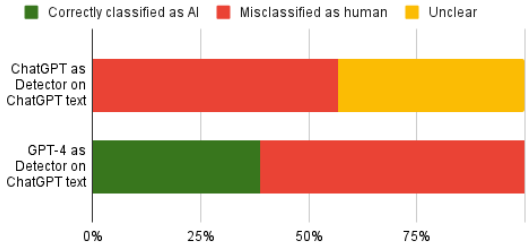

Figure 8: Performance of ChatGPT (top bar) and GPT-4 (bottom bar) on articles generated by ChatGPT.

Interestingly, with GPT-4 as the detector (Figure 7), we see a huge difference in performance across the four human-written datasets. While GPT-4 has almost perfect performance on human texts from NeuralNews, IMDb and very good performance on TweepFake, it misclassifies almost all human text from TuringBench. We think this might be due to dataset specific characteristics: human texts in TuringBench are relatively more 'noisy' than the NeuralNews ones and tend to contain extra characters that might be artifacts from the data collection process. From our experiments, we see that GPT-4 is more sensitive to such dataset artifacts and is therefore unable to classify texts properly. This might imply that GPT-4 cannot be used reliably to identify text written by humans, unlike ChatGPT. This poor performance of recent versions of GPT-4, especially in comparison to ChatGPT, has also been reported in recent work [8], where performance drops of even $\sim 95\%$ have been observed.

Performance on ChatGPT-generated text

Since ChatGPT itself is a language model and hence has the potential of being misused, we are interested in detecting text generated by ChatGPT as well. For this, we use our own ChatGPT-generated data, created in a process similar to the TuringBench dataset [37]. More precisely, we use the same human article sources as in [37], and use the headlines of the articles to generate equivalent ChatGPT articles, using the following prompt:

Generate a news article with the headline '<headline>'.

where <headline> comes from the actual human-written article. For creating this dataset, we use the ChatGPT (GPT-3.5) version as on March 14, 2023. For the rest of this paper, we refer to this dataset as ChatNews, for brevity. We use ChatGPT and GPT-4 as detectors on ChatNews, and report the performance in Figure 8. We see that ChatGPT misclassifies over 50% of ChatNews articles as human-written, and for the remaining, ChatGPT outputs the label 'uncertain'. Only 2 out of 2,000 ChatNews are correctly identified as AI-generated. This poor performance may be due to articles in ChatNews being extremely high quality and human-like, and essentially indistinguishable from actual human-written text. However, when we use GPT-4 as a detector on the same ChatNews dataset, we see more promising results. Interestingly, GPT-4 can correctly identify *some* fraction of ChatNews articles (around 38%) while ChatGPT fails completely. This gives an insight into how newer, larger and perhaps more capable language models may potentially be used to detect text from older language models.

5. RELATED WORK

ChatGPT as a detector or expert

Recent language models ChatGPT and GPT-4 have shown impressive performance on a variety of NLP tasks [28] such as natural language inference, question-answering, sentiment analysis, named entity recognition, etc. There is also empirical evidence of GPT-4 being able to perform discriminative tasks such as identifying PII (personally identifiable information), fact-checking etc. [7]. LLMs have also been evaluated as annotators [12; 17] with recent work showing some LLMs perform at par or even out-perform human crowd workers for text annotation and question-answering [16; 15; 35]. Some recent work also demonstrates how to use an LLM as a controller to use multiple models for AI tasks. Interestingly, there is also evidence [23] that ChatGPT may not perform well for more subjective NLP tasks.

The Landscape of AI-generated text and its detection

The advent of large language models(LLMs) and especially LLM assistants like ChatGPT has normalized the use of AI-generated text for a variety of purposes. Lay persons are also able to use LLMs for work, homework, leisure, or in some cases, even to mislead readers. While such tools can indeed boost productivity and inspire creative thinking, understanding the limitations of these is also important [4; 1; 3]. Given the potential for misuse of LLMs, research on the detection of AI-generated text has gained traction. While several works have shown that humans struggle at identifying AI-generated language [21; 10], a variety of computational methods for detection also exist, including feature-based methods [20; 24], methods exploiting difference in statistical measures across human and AI-generated text [14; 26], more black-box type methods involving fine-tuned language models as the detector backbone [38; 33], etc. With the popularity of ChatGPT and other conversational language models, many commercial AI content detectors have also been released for use, and marketed for use-cases such as plagiarism detection[3][4]. Prominent ones include OpenAI's detector[5], the famous ZeroGPT detector[6], etc. Another recent line of research in the direction of AI-ghenerated text detection is that of watermarking [39; 40; 22; 9] whereby indistinguishable artifacts are embedded into the text, that can be identified by computational or statistical detection methods, but not by a human reader.

6. CONCLUSION & FUTURE WORK

In this work, we investigated the capability of ChatGPT, a large language model, to detect AI-generated text. Our experiments demonstrate an interesting finding that even though ChatGPT struggles to identify AI-generated text, it does perform well on human-written text. This asymmetric performance of ChatGPT can be leveraged to build detectors that focus on identifying human-written text, and thus effectively solve the problem of AI-generated text detection, albeit in an indirect way. A few important takeaways from this empirical analysis would be:

- ChatGPT (GPT3.5) has better, more reliable performance than GPT-4 in identifying AI-generated text vs. human-written text.
- GPT-4 seems to be extremely sensitive to noise and dataset artifacts, such as those that are a result of scraping text from the internet.
- GPT-4's performance is deteriorating over time, and therefore results from using GPT-4 for identifying human-written text may be unreliable and inconclusive.
- The asymmetric performance of ChatGPT may be leveraged in a downstream detection task: ChatGPT (GPT3.5) may be used to specifically identify human-written texts reliably, thereby solving a portion of the AI-generated vs. human-written text detection task.
- Newer, larger generators may be used to detect text from older generators, such as using GPT-4 to identify ChatGPT-generated text.

In future work, we would want to explore *why* this difference in performance exists, and why ChatGPT is much better at identifying human-written text, with significantly less percentage of false negatives (i.e. human-written but misclassified as AI). One hypothesis could be that ChatGPT and these new large language models have been trained on huge corpora of text from the internet. Most of these datasets have data only till 2021, wherein much of the text on the internet was human-written (although with the pervasiveness of ChatGPT and other recent LLMs, the fraction of human to AI text on the internet would possibly change). Therefore, ChatGPT has 'seen' different styles of human writing, how human language flows, and therefore has a better understanding of what human-written text would look like. This subtle capability of ChatGPT may be leveraged to build automated detection pipelines to check the probability of a text being AI-generated. Other interesting future directions for using ChatGPT (or other LLMs) for this task may include few-shot prompting based methods, and ensemble methods leveraging multiple LLMs or feature-based classifiers.

[3]https://docs.thehive.ai/docs/ai-generated-text-detection
[4]https://copyleaks.com/ai-content-detector
[5]https://openai.com/blog/new-ai-classifier-for-indicating-ai-written-text
[6]https://www.zerogpt.com/

Acknowledgments

This work is supported by the DARPA SemaFor project (HR001120-C0123) and by the Office of Naval Research via grant no. N00014-21-1-4002. The views, opinions and/or findings expressed are those of the authors and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government.

7. REFERENCES

[1] H. Alkaissi and S. I. McFarlane. Artificial hallucinations in chatgpt: implications in scientific writing. *Cureus*, 15(2), 2023.

[2] S. Allardice. The confident wrongness of chatgpt, 2023. Accessed on June 29, 2023.

[3] Y. Bang, S. Cahyawijaya, N. Lee, W. Dai, D. Su, B. Wilie, H. Lovenia, Z. Ji, T. Yu, W. Chung, et al. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*, 2023.

[4] N. Bian, X. Han, L. Sun, H. Lin, Y. Lu, and B. He. Chatgpt is a knowledgeable but inexperienced solver: An investigation of commonsense problem in large language models. *arXiv preprint arXiv:2303.16421*, 2023.

[5] M. Bohanno. Lawyer used chatgpt in court—and cited fake cases. a judge is considering sanctions, 2023. Accessed on June 29, 2023.

[6] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[7] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

[8] L. Chen, M. Zaharia, and J. Zou. How is chatgpt's behavior changing over time? *arXiv preprint arXiv:2307.09009*, 2023.

[9] M. Christ, S. Gunn, and O. Zamir. Undetectable watermarks for language models. *arXiv preprint arXiv:2306.09194*, 2023.

[10] E. Clark, T. August, S. Serrano, N. Haduong, S. Gururangan, and N. A. Smith. All that's' human'is not gold: Evaluating human evaluation of generated text. *arXiv preprint arXiv:2107.00061*, 2021.

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[12] A. Efrat and O. Levy. The turking test: Can language models understand instructions? *arXiv preprint arXiv:2010.11982*, 2020.

[13] T. Fagni, F. Falchi, M. Gambini, A. Martella, and M. Tesconi. Tweepfake: About detecting deepfake tweets. *Plos one*, 16(5):e0251415, 2021.

[14] S. Gehrmann, H. Strobelt, and A. M. Rush. Gltr: Statistical detection and visualization of generated text. *arXiv preprint arXiv:1906.04043*, 2019.

[15] F. Gilardi, M. Alizadeh, and M. Kubli. Chatgpt outperforms crowd-workers for text-annotation tasks. *arXiv preprint arXiv:2303.15056*, 2023.

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

[17] X. He, Z. Lin, Y. Gong, A. Jin, H. Zhang, C. Lin, J. Jiao, S. M. Yiu, N. Duan, W. Chen, et al. Annollm: Making large language models to be better crowdsourced annotators. *arXiv preprint arXiv:2303.16854*, 2023.

[18] M. Heikkilä. Why detecting ai-generated text is so difficult (and what to do about it), 2023. Accessed on June 29, 2023.

[19] A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*, 2019.

[20] D. Ippolito, D. Duckworth, C. Callison-Burch, and D. Eck. Automatic detection of generated text is easiest when humans are fooled. *arXiv preprint arXiv:1911.00650*, 2019.

[21] M. Jakesch, J. T. Hancock, and M. Naaman. Human heuristics for ai-generated language are flawed. *Proceedings of the National Academy of Sciences*, 120(11):e2208839120, 2023.

[22] J. Kirchenbauer, J. Geiping, Y. Wen, J. Katz, I. Miers, and T. Goldstein. A watermark for large language models. *arXiv preprint arXiv:2301.10226*, 2023.

[23] J. Kocoń, I. Cichecki, O. Kaszyca, M. Kochanek, D. Szydło, J. Baran, J. Bielaniewicz, M. Gruza, A. Janz, K. Kanclerz, et al. Chatgpt: Jack of all trades, master of none. *arXiv preprint arXiv:2302.10724*, 2023.

[24] T. Kumarage, J. Garland, A. Bhattacharjee, K. Trapeznikov, S. Ruston, and H. Liu. Stylometric detection of ai-generated text in twitter timelines. *arXiv preprint arXiv:2303.03697*, 2023.

[25] Y. Liao, X. Jiang, and Q. Liu. Probabilistically masked language model capable of autoregressive generation in arbitrary word order. *arXiv preprint arXiv:2004.11579*, 2020.

[26] E. Mitchell, Y. Lee, A. Khazatsky, C. D. Manning, and C. Finn. Detectgpt: Zero-shot machine-generated text detection using probability curvature. *arXiv preprint arXiv:2301.11305*, 2023.

[27] R. OpenAI. Gpt-4 technical report. *arXiv*, pages 2303–08774, 2023.

[28] C. Qin, A. Zhang, Z. Zhang, J. Chen, M. Yasunaga, and D. Yang. Is chatgpt a general-purpose natural language processing task solver? *arXiv preprint arXiv:2302.06476*, 2023.

[29] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[30] T. Ryan-Mosley. Junk websites filled with ai-generated text are pulling in money from programmatic ads, 2023. Accessed on July 1, 2023.

[31] V. S. Sadasivan, A. Kumar, S. Balasubramanian, W. Wang, and S. Feizi. Can ai-generated text be reliably detected? *arXiv preprint arXiv:2303.11156*, 2023.

[32] M. Sadeghi and L. Arvanitis. Rise of the newsbots: Ai-generated news websites proliferating online, 2023. Accessed on July 1, 2023.

[33] I. Solaiman, M. Brundage, J. Clark, A. Askell, A. Herbert-Voss, J. Wu, A. Radford, G. Krueger, J. W. Kim, S. Kreps, et al. Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203*, 2019.

[34] R. Tan, B. A. Plummer, and K. Saenko. Detecting cross-modal inconsistency to defend against neural fake news. *arXiv preprint arXiv:2009.07698*, 2020.

[35] P. Törnberg. Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning. *arXiv preprint arXiv:2304.06588*, 2023.

[36] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[37] A. Uchendu, Z. Ma, T. Le, R. Zhang, and D. Lee. Turingbench: A benchmark environment for turing test in the age of neural text generation. *arXiv preprint arXiv:2109.13296*, 2021.

[38] R. Zellers, A. Holtzman, H. Rashkin, Y. Bisk, A. Farhadi, F. Roesner, and Y. Choi. Defending against neural fake news. *Advances in neural information processing systems*, 32, 2019.

[39] X. Zhao, P. Ananth, L. Li, and Y.-X. Wang. Provable robust watermarking for ai-generated text. *arXiv preprint arXiv:2306.17439*, 2023.

[40] X. Zhao, Y.-X. Wang, and L. Li. Protecting language generation models via invisible watermarking. *arXiv preprint arXiv:2302.03162*, 2023.