Examining the Human-Like Proficiency of GPT-2 in Recognizing Self-Generated Texts

Ömer Can Kuşcu, Adem Mert Akkaya
Aydın Science High School
estriiaan@gmail.com; ademmertakkaya@gmail.com

Abstract: In recent years, the widespread use of generative language models has brought opportunities as well as some philosophical and technical questions. GPT-2, a language model with 1.5B parameters, is an open-source language model provided by OpenAI. Our aim in this paper is to utilize the classification capabilities of GPT-2 to create a new perspective on the question of whether language models show some kind of consciousness/self-awareness, in addition to technical questions such as how to detect the misuse of the outputs of language models.

To investigate this phenomenon, GPT-2ForSequenceClassification model was fine-tuned on TuringBench datasets and its performance was examined. In addition, the accuracy achieved by model as a result of training with training sets of different sizes, as well as its performance in human-machine discrimination, were evaluated.

The model exhibits consistent and above-average performance in identifying GPT-2-generated content compared to its classification accuracy in distinguishing other machine-generated text from human writing. This performance of the model in understanding self-generated texts is very similar to people's ability to recognize their own writing, and these results offer an interesting perspective on the self-awareness of artificial intelligence. Additionally, the model showed high accuracy in distinguishing machine generated output from human output, even when trained with very few examples.

Introduction: In recent years, the widespread use of generative language models such as GPT and Llama has opened up a wide range of possibilities, as well as a variety of problems. Some of these problems are of a technical nature, like how to detect the misuse of language models, while others are of a philosophical nature, like whether language models exhibit human-like consciousness and/or self-awareness. For the former of these questions, which requires multi-layered solutions, there have been some commercial solutions published, such as GPTZero, as well as a lot of academic work using various methods [1]. For the latter, the discussions that started long before the appearance of generative language models have led to the publication of research in this area since the 1950s, and numerous metrics like the Turing Test have been developed. Machine learning enables computers to learn without having to be specifically programmed, inspired by the human mind. It is a field of computer science that focuses on enabling algorithms to learn, recognize patterns, predict and make decisions without having to be told exactly what to do [2]. Deep learning, a sub-branch of machine learning, has revolutionized artificial intelligence in recent years. Deep learning is a subset of machine

learning that uses neural networks to learn from data. Neural networks are inspired by the structure and function of the human brain and can learn complex patterns from data that are difficult or impossible for humans to learn manually [3]. Deep learning is widely used in every aspect of our lives today. In addition, it has been observed that it has shown great success in this wide range of areas [4] [5] [6]. Natural language processing (NLP) is a field of artificial intelligence (AI) that deals with the interaction between computers and human (natural) languages [7]. It is concerned with giving computers the ability to understand, interpret, and generate human language. NLP is a broad field that encompasses a wide range of tasks. Deep learning has developed and continues to develop in the field of natural language processing [8]. Large language models (LLMs) have emerged as a powerful tool in the field of natural language processing (NLP), capable of generating human-quality text, translating languages, and answering questions in an informative way [9]. In 2017, the introduction of transformers by Vaswani et al. (2017) marked a turning point in the evolution of LLMs [10]. These models, which rely on an attention mechanism to focus on relevant parts of the input text, demonstrated superior performance in various NLP tasks, including machine translation and language modeling [11] [12]. GPT models, also known as Generative Pre-trained Transformers, are a family of large language models (LLMs) developed by OpenAI. They are trained on massive amounts of text data and code, allowing them to generate humanquality text, translate languages, write different kinds of creative content, and answer your questions in an informative way. If properly fine-tuned, the models can handle various alternative tasks [13]. GPT-2 marked a significant leap forward from GPT-1, boasting 1.5 billion parameters and standing as one of the most expansive language models upon its debut. Trained on an extensive corpus encompassing web content, books, and diverse written materials, it tackled the language modeling task by predicting the subsequent word in a given text sequence based on preceding words [14]. On the other hand, GPT-2 sequence, which this study will be focusing on, refers to the specific sequence of words or tokens that is generated by the GPT-2 model. This sequence is generated by sequentially predicting the next word or token in the sequence, based on the context of the previous words. In other words, GPT-2 is the model, and GPT-2 sequence is the output of the model. Our aim in this study is to provide alternative solutions to these two problems through the GPT-2 model, which is provided by OpenAI, and to contribute to the growing scientific literature on the subject. Generative language models are remarkable for their high number of parameters compared to many pre-trained models. If properly fine-tuned, the models can handle various alternative tasks [13]. Our aim in this study is to provide alternative solutions to these two problems through the GPT-2 model, which is provided by OpenAI, and to contribute to the growing scientific literature on the subject.

Materials and Experiments: We used Python programming language version 3.10 in the study. All experiments were run on Kaggle with the Kaggle P100 accelerator. The source code is accessible at www.kaggle.com/mercankuscu/ GPT-2-classification-source-code We used the TuringBench database containing 20 different datasets in the study. Each dataset contains

original news texts and texts generated by various language models (LM) using the original texts' titles. The targets consist of one of 2 labels: human or the language model from which the text was generated. In the study, the GPT-2ForSequenceClassification model was trained on the datasets consisting of the output of various language models and original news texts, each with 80 examples of training and 20 examples of validation sets, and the results were tested with a test set of 9911 examples for each dataset. The GPT-2ForSequenceClassification model was trained to evaluate the performance of few-shot learning on the GPT-2 Pytorch dataset with random samples ranging from 10 to 100 with an 8-2 ratio between training and validation data, and each dataset was tested on a test set of 9911 samples. To measure human vs. non-human performance, the model was trained with 40 training and 10 validation data and 3 different learning rates on a control group containing machine-generated texts produced by many different models along with original news texts. In addition, the model was trained with a training and validation set of 10 to 50 samples and with the best learning rate to measure the model's performance of few-shot learning on human vs nonhuman task.

Results and Discussion: The use of generative language models has become increasingly common in recent years [15]. ChatGPT is the fastest-growing user application in history [16]. The widespread use of these models has led to practical problems such as how to distinguish the results of the models from human texts, as well as philosophical problems about whether the models show self-awareness and consciousness. The GPT-2 model used in this study was trained in various ways as one solution to these problems. The fact that the average accuracy of the model on the datasets created with the outputs of GPT models is significantly higher than the non-GPT models shows that the model is more successful in recognizing outputs similar to the outputs created by the model itself. Given that recognizing one's own expressions is a measure of human self-awareness [17], the fact that the model achieved the highest accuracy on the dataset generated with the output of the GPT-2 Pytorch model, which is the same as the classification model, indicates that the model shows a human-like behaviour in identifying self-generated texts (Table 1).

Table 1: Accuracy by models trained and tested on different datasets

Model used for dataset creation	Accuracy
GPT2 Pytorch	0.95
GPT2 Small	0.90
GPT2 Medium	0.91
GPT2 Large	0.90
GPT2 XL	0.91

GPT1	0.92
GPT3	0.92
Grover Base	0.73
FAIR WMT19	0.62
FAIR WMT20	0.78
Transformer-XL	0.88
XLNet	0.92
GPT Mean	**0.915**
GPT2 Mean	**0.914**
Non-GPT Mean	**0.786**

When the model was tested on the GPT-2 Pytorch dataset after being trained with varying numbers of examples, it was observed that the models trained on training sets of 50 examples and above achieved accuracy values above 0.90. This success of the model in the classification task with limited examples makes it a viable option for other tasks where few-shot learning is utilized (Figure 1).

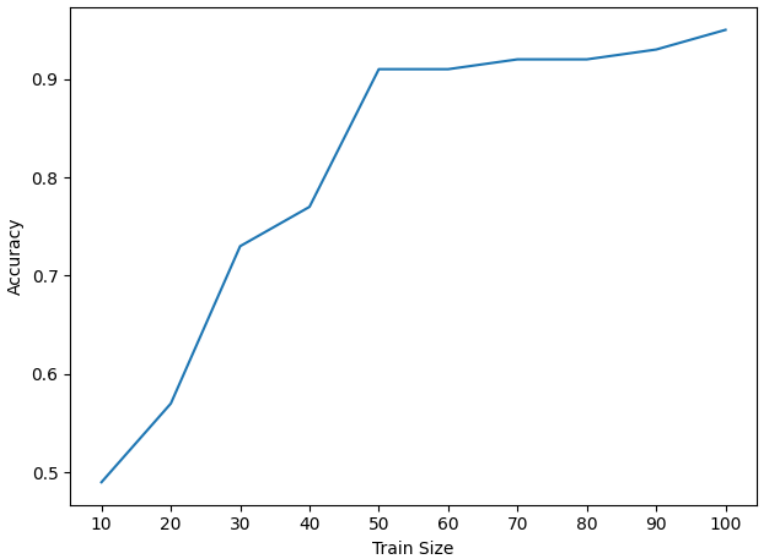

Figure 1: Line graph of few-shot accuracy by training set size on GPT-2 Pytorch dataset

The high accuracy of the model in human vs. nonhuman classification, despite being trained with a set of 40 examples, shows that even with a limited number of examples, the GPT-2 model can be very useful in classifying whether texts are machine output or not, which is a growing problem today [18] (Table 2). The few-shot learning performance of the model increased with a decreasing acceleration up to 50 samples and reached an accuracy metric above 0.90 at 50 samples (Figure 2). The few shot learning performance of the model is consistent with the paper introducing the GPT-2 model [13].

Table 2: Human vs. non-human performance by learning rate

Model Learning Rate	Accuracy
5e-4	0.87
5e-5	0.94
5e-6	0.46

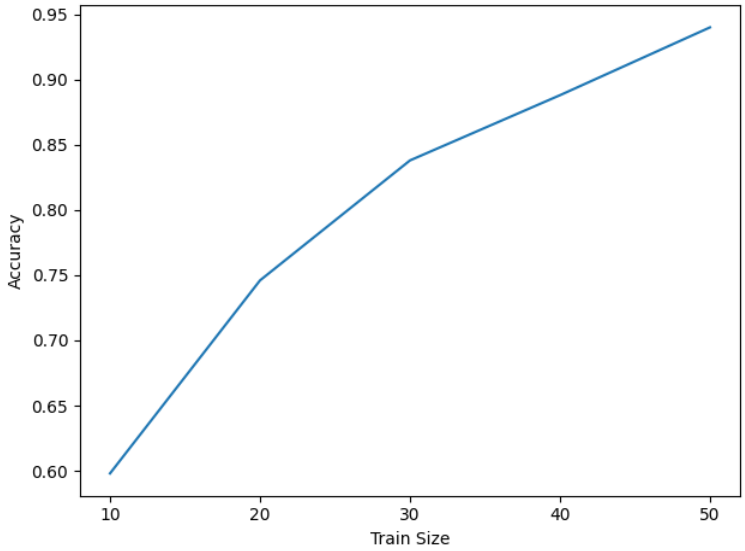

Figure 2: Line graph of few-shot accuracy by training set size on human vs. non-human dataset

Conclusion: The widespread use of generative language models in recent years has created both technical and practical problems. In this study, the GPT-2 model, which is offered as open source by OpenAI, is trained with various datasets from the TuringBench database as a solution to these problems. The increasing use of generative language models has made it important to detect the source of texts. On the human vs nonhuman task, the best model trained with a dataset of 50 examples achieved an accuracy of 0.94, demonstrating that it can succeed in this critical problem with limited data. When the model was trained to classify the outputs produced by GPT-2, it achieved an average accuracy of 0.91, while the average for non-GPT models was

0.78. The significant difference between the two averages may indicate the ability to recognize one's own writing, which is also an important part of human self-awareness. In order to measure the model's few-shot learning performance, the model was trained on GPT-2 Pytorch data with various number of samples and it was observed that it achieved an accuracy above 0.90 for 50 samples and above. These results are important in terms of demonstrating the success of the GPT-2 model in detecting the text source even with limited samples. In conclusion, we believe that the results obtained provide an alternative perspective to the philosophical problems in the developing literature and are important in explaining the success of the GPT-2 model in significant tasks.

Acknowledgments: We acknowledge Tim Gianitsos for his contribution and supervision of the project.

References:

[1] Suchin Gururangan, Margaret Li, Mike Lewis, Weijia Shi, Tim Althoff, Noah A. Smith, and Luke Zettlemoyer. Scaling expert language models with unsupervised domain discovery, 2023.

[2] Kevin P. Murphy. Machine learning - a probabilistic perspective. In Adaptive computation and machine learning series, 2012.

[3] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. Deep Learning. MIT Press, 2016. http://www. deeplearningbook.org.

[4] Andreas Kamilaris and Francesc X Prenafeta-Boldú. Deep learning in agriculture: A survey. Computers and electronics in agriculture, 147:70–90, 2018.

[5] Shervin Minaee, Yuri Boykov, Fatih Porikli, Antonio Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. Image segmentation using deep learning: A survey. IEEE transactions on pattern analysis and machine intelligence, 44(7):3523–3542, 2021.

[6] Andre Esteva, Alexandre Robicquet, Bharath Ramsundar, Volodymyr Kuleshov, Mark DePristo, Katherine Chou, Claire Cui, Greg Corrado, Sebastian Thrun, and Jeff Dean. A guide to deep learning in healthcare. Nature medicine, 25(1):24–29, 2019.

[7] Erik Cambria and Bebo White. Jumping nlp curves: A review of natural language processing research [review article]. IEEE Computational Intelligence Magazine, 9(2):48–57, 2014.

[8] Daniel W Otter, Julian R Medina, and Jugal K Kalita. A survey of the usages of deep learning for natural language processing. IEEE transactions on neural networks and learning systems, 32(2):604–624, 2020.

[9] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. A survey of large language models, 2023.

[10] Tianyang Lin, Yuxin Wang, Xiangyang Liu, and Xipeng Qiu. A survey of transformers. AI Open, 3:111–132, 2022.

[11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

[12] Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. A comprehensive overview of large language models, 2023.

[13] Alec Radford, Jeff Wu, Rewon Child, D. Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners, 2019.

[14] Partha Pratim Ray. Chatgpt: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. Internet of Things and Cyber-Physical Systems, 3:121–154, 2023.

[15] Albert Gatt and Emiel Krahmer. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation, 2018.

[16] Krystal Hu. Chatgpt sets record for fastest-growing user base - analyst note, 2023.

[17] Steven Brown. The “who” system of the human brain: A system for social cognition about the self and others. Front. Hum. Neurosci., 14:224, June 2020.

[18] Evan Crothers, Nathalie Japkowicz, and Herna Viktor. Machine generated text: A comprehensive survey of threat models and detection methods, 2023.