Using DetectGPT to distinguish AI generated university problems and solutions - Working title

Smriti Suresh
smritis@bu.edu
Boston University

Shaunak Joshi
ssjoshi@bu.edu
Boston University

Gitika Jha
gitika@bu.edu
Boston University

Abstract

Implications of AI for higher education have seen incremental growth both in terms of content generation as well as course evaluation. The advent of Large Language Models and Neural Networks to solve and generate user defined problems from written texts and research has made a lasting impact on how AI will play a considerable role in automating learning environments. Our project intends to extend upon the existing work Drori et al. [2022] of a neural network built upon program synthesis and few shot learning to solve, explain and generate university math problems. The primary goal of this project is to use DetectGPT Mitchell et al. [2023] to evaluate the results obtained by the interpretable NN model by [Drori et al., 2022] and form better conclusions as to how similar the questions or solutions generated by the AI are to actual human ones. In addition to this, we plan to employ adversarial attacks on DetectGPT employing adjustments or perturbations in the machine-generated text such that it may be classified as non-machine generated. Using the image specific aspects of the existing GPT-3 model with OpenAI's Codex transformer, particularly in machine-generated image detection and interpretable machine-generated images can also be a part of future work.

1 Introduction

Mathematics is a field of knowledge that necessitates in-depth comprehension and problem-solving skills. Earlier, it was widely understood that advanced mathematical problems are a big challenge for artificial intelligence and neural networks [Choi, 2021, Hendrycks et al., 2021]. However, recent advancements as well as varied approaches taken to handle this problem have enabled AI models to solve university-level mathematical problems with high accuracy. Building upon the work done in the paper Drori et al. [2022], where the authors enhanced the performance of the GPT-3 model using OpenAI's Codex Zaremba [2021] transformer to fine-tune on code rather than text to automatically solve these questions with an unprecedented 81% accuracy, a huge improvement from the previous state-of-the-art models with a mere 8.8% accuracy.

We introduce a novel approach to test the robustness of the model in the paper [Drori et al., 2022] and its ability to generate new questions by evaluating it against DetectGPT [Mitchell et al., 2023], GPTZero [gpt, 2022], OpenAI ([AIT]) and other detection models to check if the model can fool these tools and figure out where we can improve detection results by adversarial examples. Our work represents a significant milestone in the field of higher education where we find different use-cases of AI in this field.

2 Related Work

There has been a drastic increase in the number of LLMs created that achieve not only high accuracy in their answers but also the improvement in their performance that generates highly convincing

Preprint. Under review.

answers to the prompts by users [Zellers et al., 2019, OpenAI, 2022]. For zero-shot or few-shot NLP tasks, we have seen the increase in the use of transformers instead of other architectures of neural networks [Brown et al., 2020, Wei et al., 2021, Wang et al., 2022].

In the paper on which we are building on Drori et al. [2022], the model is pre-trained on textual data in addition to fine-tuning on code data increases the performance of the model to solve mathematical problems substantially.

It has become imperative to detect if solutions are generated by an AI model since there has been a growing misuse of the technology in the field of education as [Hacker et al., 2023, Cotton et al., 2023]. To counter this, we plan to check if the mathematical answers generated by the model from Drori et al. [2022] gets detected by the top AI detection models in the field currently like DetectGPT [Mitchell et al., 2023], GPTZero [gpt, 2022], OpenAI ([AIT]) among others. We also plan to dive deeper into exploring how solutions provided by the AI model can be manipulated to pass as a human-generated result for these detection models. Using this, we can see in what scenarios do these detection models fail exactly using the concept of adversarial attacks [Miyato et al., 2016].

3 Evaluation Criteria

Evaluation of the various models (DetectGPT Mitchell et al. [2023], GPTZero, OpenAI (AIT) when tested on the dataset is proposed to be done in the following ways:

1. Quantitative Metrics such as accuracy, F1 score on the labeled human-generated vs. machine-generated data

2. Confusion matrix assessment : Addressing questions like “Were all the AI-generated questions/answers classified correctly as machine-generated?” ; “Were some human-generated texts also falsely classified as AI-generated” and so on, so as to present an in-depth analysis of the performance of the state-of-the-art AI text detection models for the Mathematical dataset of questions+answers fed into it.

3. For Adversarial attacks : evaluate proportion of texts misclassified when perturbed and fed through the model to test robustness of the Detection model.

4. For Images (part of future work) : similarity measures such as Image Similarity API (Douze et al. [2021]) to compare existing image v/s AI-generated image.

4 Datasets Used

The project’s datasets include: XSum Narayan et al. [2018], Wikipedia Paragraphs Wu et al. [2019], SQuAD contexts Rajpurkar et al. [2016], Reddit Writing-Prompts Fan et al. [2018], English and German splits of WMT16 Bojar et al. [2016], and long-form answers from the PubMedQA dataset.

5 Algorithms

DetectGPT is the fundamental basing point for this project. We shall apply detectGPTto the model given by Drori et all and explore also other methods such as GPTZero, etc. We plan on using the model outputs inconjuction with DetectGPT. This means also exploring in detail how the given outputs can be perturbed to deceive the detection models. Using this methodology we also plan on applying adversarial attacks and trying to defeat detection methods.

References

Iddo Drori, Sarah Zhang, Reece Shuttleworth, Leonard Tang, Albert Lu, Elizabeth Ke, Kevin Liu, Linda Chen, Sunny Tran, Newman Cheng, Roman Wang, Nikhil Singh, Taylor L. Patti, Jayson Lynch, Avi Shporer, Nakul Verma, Eugene Wu, and Gilbert Strang. A neural network solves, explains, and generates university math problems by program synthesis and few-shot learning at human level. *Proceedings of the National Academy of Sciences (PNAS)*, 119(32), 2022.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. Detectgpt: Zero-shot machine-generated text detection using probability curvature. *arXiv preprint arXiv:2301.11305*, 2023.

Charles Q Choi. 7 revealing ways ais fail, Sep 2021. URL https://spectrum.ieee.org/ai-failures.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. *CoRR*, abs/2103.03874, 2021. URL https://arxiv.org/abs/2103.03874.

Wojciech Zaremba. Openai codex, Aug 2021. URL https://openai.com/blog/openai-codex/.

2022. URL https://gptzero.me/.

January .

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. Defending against neural fake news. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/3e9f0fc9b2f89e043bc6233994dfcf76-Paper.pdf.

OpenAI. Chatgpt: Optimizing language models for dialogue, Nov 2022. URL https://openai.com/blog/chatgpt/.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *CoRR*, abs/2005.14165, 2020. URL https://arxiv.org/abs/2005.14165.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. Finetuned language models are zero-shot learners. *ArXiv*, abs/2109.01652, 2021.

Thomas Wang, Adam Roberts, Daniel Hesslow, Teven Le Scao, Hyung Won Chung, Iz Beltagy, Julien Launay, and Colin Raffel. What language model architecture and pretraining objective work best for zero-shot generalization? In *International Conference on Machine Learning*, 2022.

Philipp Hacker, Andreas Engel, and Marco Mauer. Regulating chatgpt and other large generative ai models, 2023. URL https://arxiv.org/abs/2302.02337.

Debby Cotton, Peter Cotton, and J. R Shipway. Chatting and cheating. ensuring academic integrity in the era of chatgpt, Jan 2023. URL edarxiv.org/mrz8h.

Takeru Miyato, Andrew M. Dai, and Ian Goodfellow. Adversarial training methods for semi-supervised text classification, 2016. URL https://arxiv.org/abs/1605.07725.

Matthijs Douze, Giorgos Tolias, Ed Pizzi, Zoë Papakipos, Lowik Chanussot, Filip Radenovic, Tomas Jenicek, Maxim Maximov, Laura Leal-Taixé, Ismail Elezi, et al. The 2021 image similarity dataset and challenge. *arXiv preprint arXiv:2106.09672*, 2021.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. Ranking sentences for extractive summarization with reinforcement learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3642–3652, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1207. URL https://www.aclweb.org/anthology/D18-1207.

Yu Wu, Wei Wu, Chenyan Xiong, Zhoujun Li, Richard Socher, and Caiming Xiong. Open domain web retrieval and passage generation for qa: A strong baseline and new dataset. *arXiv preprint arXiv:1911.10470*, 2019.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, 2016.

Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 889–898, 2018.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 131–198, Berlin, Germany, August 2016. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/W16-2340`.