CS 545: AI Generated Research Analysis Proposal

Thanya Nguyen
The University of Tennessee
COSC 545
Knoxville, Tennessee
Thatngu1@vols.utk.edu

Jake Marlow
The University of Tennessee
COSC 345
Knoxville, Tennessee
Amarlow6@vols.utk.edu

Samuel Lavey
The University of Tennessee

COSC 345
Knoxville, Tennessee
slavey@vols.utk.edu

Haoran Chen
The University of Tennessee
COSC 345
Knoxville, Tennessee
Hchen73@vols.utk.edu

***Abstract*—We seek to grade research by analyzing the prevalence and frequency of content generated Artificial Intelligence (AI) in research papers.**

Keywords—Artificial Intelligence, Research, Analysis (key words**)**

I. OBJECTIVE

AI has grown more advanced in recent years and has become widely integrated due to its convenience and productivity. In every sector there has been a dramatic push to replace human effort with more dynamic computer intelligence. It is our objective to analyze to what extent this can be observed with scientific journals and research papers. By using a series of metrics and a wide range of papers across journals and genres, we seek to give a score to each. We can classify the likelihood of AI generation based on this score. A lower score means a higher likelihood of AI generation, and a failing grade would be given to scores below a certain threshold. We aim to report the likelihood based on the year a paper was published was published, the institution it originates from, the journal it appears in, and the type of research it is attempting to conduct. Through these lenses, we can visualize the presence of AI in research and narrow down which classifiers make it either more or less likely to appear.

II. A BRIEF MOTIVATION

A. Media Attention and Academic Disgrace

In recent years, media attention has been focused on the rise of AI generated work in academic papers. In February of 2024, an image was published in a well reputed journal of an obviously generated rat reproductive system. This made its way past editors and was published before being quickly retracted after the error was discovered [1]. Editorial negligence aside, as AI becomes more advanced it may not be as easy to detect AI imagery. The fact that this was published leads to potentially wide-reaching questions about academic integrity and credibility. In general research papers have seen a 72% increase in AI generated work [2]. Computer science as a field was by far the worst offender. Our team is curious to verify the extent to which these findings are correct and if they have increased as time has gone on. We will also attempt to identify patterns in AI generated writing and uncover which metrics are most accurate in detecting AI generated work. It is crucial for academia to retain the trust of the public that we offer transparency and integrity, especially when it comes to generated research.

III. DETAILED DISCUSSION OF THE DATA THAT WILL BE OBTAINED OR USED IN THE PROJECT

To obtain the most relevant data for this project, the Generated Research team expects to use recently written research papers in the field of computer science. The most popular LLM, ChatGPT, was released in November of 2022, which means that we will be looking for databases of research papers published in 2023.

Initially, the team wants to begin the research with around 50 papers and test to see if our hypothesis about AI generation in research is visible enough. Topics such as cybersecurity, computer hardware, algorithm analysis, optimization, and graphics visualization are areas of interest for the team. If our results end up being too sparse, then we may expand the field and sample size of our research paper data.

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

To test this data, the team will gather a separate set of data to measure the probability of AI generation in each research paper. While there is no perfect method to detect or measure AI generation, LLMs are constrained by their training data. This means that they can sometimes use words/phrases repeatedly or change tone unexpectedly. We will take advantage of this known feature of AI generated text to measure levels of AI generation by accumulating target words and phrases known to be used in higher-than-normal frequencies within language generation.

IV. RESPONIBILITIES OF EACH MEMBER

Each member of the project will be finding data for this project. There is probably not Kaggle datasets that have research papers that are published in 2023 or beyond, so we will be extracting our own data. Our focus on this project is computer science related papers. Each member will be finding around 16 papers each, relating to topics like cybersecurity, machine learning, or artificial intelligence.

After each member in the research generated team finds enough papers, one person will find a dataset that includes ChatGPT or any AI's dataset. This ideally should include what their output is so we can evaluate its verbiage and ways it chooses to form sentences. We can also further expand on this idea by analyzing the pictures that are generated by AI that may be included in the research papers.

When coding, each member of the team will be using Git and Github to push and pull our codes, we will be working on this separately but together at the same time. The generated research team will plan to meet after class on Thursday to further work on this project. We will tokenize the words in the papers so we can use Python to work on this analysis.

V. A TIME-LINE OF MILESTONES

The general timeline of this milestone will be throughout the semester and along the class's general structure will. For example, this paper will be done by the due date of next Thursday. Next week, we plan on gathering all our datasets. October we will start coding our project. This is when we will make tweaks, see other different research papers in different fields and their use of AI.

After October, this when be when we will expand on our ideas such as a using logistic regression to determine with AI if the paper is generated with AI or not. We will use factors such as word frequency, word choice, sentence length, punctuation usage, average word count length, etc. to determine if the paper is written by AI We will begin our PowerPoint for our final presentation and final papers during this time. I am spacing out the timeline so there is wiggle room for the team in case there are problems in the code or there's issues with our idea and a dramatic change must occur.

VI. THE EXPECTED OUTCOME

The expected outcome from this project is that we learn more fundamentals of digital archelogy. The team hopes to find more insight on how AI has impacted research papers since the recent surge of ChatGPT. We hope to learn more about time management and teamwork because we are working in a team. The team also hopes to learn more about how data scraping works and how artificial intelligence has impacted the culture of society today.

Regarding the project, the expected outcome of that is to see how AI's influence on research papers. Although some of the team members hope that it has not influenced research papers that much, we are unsure what the results we will find are and will be surprised to what we find out throughout this journey.

VII. REFERENCES

1. https://gizmodo.com/rat-dck-among-gibberish-ai-images-published-in-science-1851260727
2. https://originality.ai/blog/ai-generated-research-papers