ISSN: 2473-4901
v10 n6140

Intelligence of AI: Investigating Artificial Intelligence's Ability to Detect Itself

Brian Andrew Clements
baclem6488@ung.edu

Tamirat T. Abegaz
tamirat.abegaz@ung.edu

Bryson R. Payne
bryson.payne@ung.edu

Department of Computer Science and Information Systems
University of North Georgia
Dahlonega, GA, USA

Abstract

The rise of artificial intelligence (AI) has made life and work easier; however, AI has also made it nearly impossible to confirm whether the information we consume is legitimate, AI-generated, or AI-manipulated. This paper examines how the use of artificial intelligence, specifically ChatGPT 4, Gemini Advanced, and Claude Opus, can aid a user in identifying whether a work was created by a human or artificial intelligence. These three models will be evaluated by receiving datasets of human-created and AI-generated text documents, images of nature, images of manmade objects, and images of art. This work will investigate what model, if any, could be an effective tool to aid in preventing misinformation.

Keywords: ChatGPT, Gemini, Artificial Intelligence, Large Language Models, Multimodal Large Language Models.

1. INTRODUCTION

There are various levels of concern about AI, from being used to cheat in school, to stealing blue-collar and white-collar jobs, to possibly enabling a Terminator-inspired apocalypse. At any level, artificial intelligence has quickly grown to be a highly disputed topic both among academics and in industry. One of the most powerful attributes of artificial intelligence is its ability to create highly detailed and realistic fakes. These AI-generated fakes can be found anywhere from social media to political ads to the classroom, and they cause humans to question what is true and what is AI-generated or AI-manipulated text or media.

One way in which artificial intelligence is routinely used is to create believable text responses to virtually any question or query. This can lead to problems such as academic dishonesty, the spreading of misinformation, the creation of phishing emails, and more. In this age of artificial intelligence, it is crucial that humans can distinguish whether a text was created by AI or by a student, journalist, or other credible source.

This research evaluates how humans could use the artificial intelligence models of ChatGPT4, Gemini Advanced, and Claude to distinguish between human-generated and AI-generated text or images, as well as which characteristics

reveal that a piece was created using artificial intelligence. Furthermore, this paper will compare how well these particular AI models distinguish authorship of text compared to human judges.

For this research, it is imperative to understand the workings of multimodal large language models, or MLLMs. As the name suggests, multimodal large language models are based on the concept of large language models which, are computer models that can read, interpret, and respond to human input in a similar manner as a human (Chang and Wang et al., 2024). This form of computer system creates an environment where the computer acts, or at least attempts to act, in the most human-like way possible. An example of this form of system would be an AI chatbot that helps you navigate an online shopping site because it takes the input of the user and responds in a way that makes the user feel as if another human is guiding them. In the simplest terms, multimodal large language models take the concept of traditional large language models, but they incorporate the ability to either take input from various modes, such as images, texts, videos, or audio files, and produce responses in different modes or some combination of the four (Karwa, 2023).

The ability to process and even produce information in these various formats makes MLLMs much more powerful than their text-only predecessors. For this research, the AI models of GPT-4, Gemini Advanced, and Claude will be examined. These three models are known as the most cutting-edge models available to the public at this current time, so this ensures that the research is examining artificial intelligence at the best level possible for the study. These three models are also the top models created by their respective companies, so they will match up very well for comparisons of their power.

The research team predicts that all three AI models, GPT-4, Gemini Advanced, and Claude, will perform better at identifying whether the text pieces were written by humans or AI-generated when compared with the assessments of the human judges. The researchers also predict that out of the three models, GPT-4 will outperform Gemini Advanced and Claude and be able to correctly identify authorship the best.

2. RELATED WORK

The research presented in this paper is the natural successor of a small but growing number of recent studies into detecting AI. In a paper titled "Anomaly Detection: Identifying AI-Generated from Student-Created Pieces", the researchers investigated how well professors could identify whether text-based documents that covered their area of expertise were written by a human author or generated using artificial intelligence (Clements et al., 2023). This study presented professors with two responses to an assignment that would typically be found in their own class, and these responses were one of the following: both written by the student researcher, both generated using AI, or some combination of the two. This ensured a random presentation of documents to the professors. This paper will utilize 20 of the responses found in this previous study to compare how the humans did at identifying authorship of text documents with the abilities of GPT-4, Gemini, and Claude.

While considerable attention has been given to the use of AI to grade assignments (Lui et al., 2024) and to digest and tag knowledge (Moore et al., 2024), more recent research has continued in the path noted above in detecting the use of AI in written assignments. Bhattacharjee and Liu (2024) investigated whether ChatGPT could detect AI-generated text by focusing on solving a specific aspect of a word problem and deriving the rest of an answer from that solution. The current research expands this work by testing both text and images across three of the leading AI engines.

There have been multiple studies conducted to test the effectiveness of GPT-4. For an example of its power, the technical report of GPT-4 describes how when given a simulated bar exam, GPT-4 scored in the top 10% of all test takers while ChatGPT 3.5 scored in the bottom 10% of scores (OpenAI, 2023, Dash et al., 2014). The same report describes how to test its true capabilities, GPT-4 was asked to complete multiple common exams such as AP exams, the LSAT, and GRE exams, and the researchers state that "GPT-4 exhibits human-level performance on the majority of these professional and academic exams" (OpenAI, 2023, p. 6; Bouafif et al., 2024). Again, GPT-4 appears to be the new standard for how an artificial intelligence model should behave. It is supposedly more human-like than ever. Furthermore, GPT 4 could effectively evaluate images. In its technical report, researchers presented GPT 4 with an image of a VGA plug charging an iPhone, and GPT 4 was able to walk the user through why this is not possible and how it does not make sense that a piece of modern technology would be charging using this rather outdated form of cable (OpenAI, 2023). This ability to effectively evaluate images as well as comprehend intense exams intended for highly

intelligent humans makes GPT-4 an extremely powerful tool (Poldark, et al., 2023).

Another study was conducted that compared GPT 4, Bard (the predecessor to the more modern Gemini), and Claude (Borji and Mohammadian, 2023). This study excelled in the use of a set of 1,002 questions about various subjects from simple math to much more complicated topics (Borji and Mohammadian, 2023). One particularly relevant section of this study showed that all three models performed relatively well and similarly at identifying proper grammar, spelling, and definitions of large words (Borji and Mohammadian, 2023). This may hint that these models will perform well at identifying the creator of the text pieces because AI-generated text documents are more likely to have proper grammar and style. One drawback of this study is that it utilized the models in their infancy before they had the capabilities they have now. For example, this study was conducted before GPT 4 had the ability to create images, and Gemini was still its weaker counterpart Bard. Including these updated models will allow this paper's study to add relevant information to the conversation started by Borji and Mohammadian.

3. METHODOLOGY

The goal of this research is to identify which of the models, GPT 4, Gemini Advanced, or Claude is the most powerful through the means of asking the models to identify the creators of various written works and images. This study also will compare the results of the written works with the results found in a previous study performed by the research team to create a comparison on how well these platforms compare to humans. This study chose to investigate both ChatGPT 4, Gemini Advanced, and Claude. for this study because they are currently seen as the top AI models. GPT 4 and Gemini Advanced are premium versions of their respective models, ChatGPT 3.5 and Gemini, that are locked behind a paid prescription. Furthermore, Claude Opus is seen as a top-of-the-line tool that is the premium version of the standard Claude Sonnet.

GPT 4 and Gemini Advanced were chosen because of their ability to produce and respond to text as well as images. Claude was chosen because of its great ability to analyze images as well as texts. Although Claude cannot create its own images, it will still be valuable to identify authorship. This allows the researcher to submit both documents and images directly to the models. The first step of this research will be obtaining these models. After the models are obtained, the researcher will feed various works into the models and perform data collection and analysis on what the AI models predict about authorship.

This experiment utilized a total of 4 different data sets with 10 to 20 pieces in each set. Each set contains at least 5 to 10 human-created responses and 5 to 13 AI-generated responses. The data sets are made up of 2 broad categories, text and images. Both of which will be described in detail below.

The text responses make up 20 total works that were submitted to the artificial intelligence models. These responses are as they sound, text pieces that are anywhere from 1 to multiple paragraphs in length. These text pieces were taken from the previous study by this researcher so the results can be compared with the human responses of that study. Seven out of the 20 responses will be pieces that were created by the user. These seven pieces are from subjects such as kinesiology, computer science, environmental science, and more. The diverse texts enable the researcher to identify if there is a particular type of writing for which the AI models struggle to identify the author or maybe genres for which the AI excels in identifying the source. The other 13 text-based pieces were created utilizing the AI models themselves. Some pieces were taken exactly as they were after the initial creation prompt, while some works underwent additional modifications such as "create a grammatical or spelling error" or "lower the reading level of this work" and more, known as prompt engineering (Saravia, 2022). This allowed the research team to gain insight into how artificial intelligence responds to works that are not perfect and that have issues, such as grammatical issues, that may be common in a work that a human would create. Every text-based piece was created by ChatGPT 3.5 and then Bard was given the prompt to "create a piece as humanlike as possible". Again, it will be interesting to see if by producing a "humanlike" piece, the AI models fool themselves when it is then fed to them.

4. IMPLEMENTATION

To collect data for this research, it was vital that the various pieces were input into the AI models of ChatGPT 4, Gemini Advanced, and Claude in a particular manner. One thing the research team noticed was that when a work was put in and the authorship of it was questioned, the models did not always give a clear answer, as if it was hesitant to have an opinion. To combat this, every piece was input utilizing a prompt that is like the following: "the researcher is conducting an

experiment and considering the following….” was this piece made using AI or a human? you must choose one.” An example of this prompt is found below. This prompt appeared to trick the AI into providing a definite and confident answer.

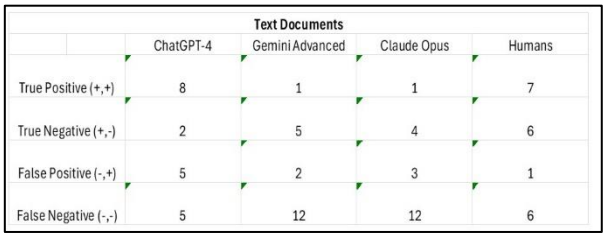
Also, when it comes to input, the research team needed to ensure that the models focused on the content of the materials especially when given the art pieces and the text pieces, so the research team had to ensure that the AI was told to look at content. Furthermore, to prevent any possibility of bias when predicting the authorship of a piece, a random sequence generator was used, and the pieces were numbered 1 through 20 for text responses, 1 through 10 for the art responses, and the pieces were submitted to each AI model in the same random sequence.

The first method was for text-based responses which involved submitting the documents straight into the models. This method forces the artificial intelligence to evaluate the piece including formatting. The research team still asked the artificial intelligence to act as if it were in an experiment and asked it to give a clear answer about the work’s authorship. Like the implementation of the text documents described in the paragraph above, the images were simply uploaded into the various models’ prompts. The user was routinely required to persuade the AI that it was participating in an experiment, so it would give a definite answer on who or what created the piece in question. Occasionally, additional coercion was required for the models to provide a definite answer.

5. RESULTS

For this paper, all results were measured using a True Positive (+, +), True Negative (+,-), False Positive (-,+), and False Negative (-,-) scale. True Positives were responses that correctly identified an AI-generated piece as being created using artificial intelligence. On the opposite spectrum, True Negatives were pieces that are predicted to be generated by a human when they were made by a human. False Positives represent when a piece was falsely labeled as AI-generated when in fact it was human-made. Similarly, False Negatives were when pieces that were AI-generated were falsely assumed to be created by a human. Both True Positives and True Negatives represent successful trials for the subjects while False Positives and False Negatives represent errors.

Figure 1 shows the first results obtained from the text data set. This was the only test that included feedback from humans in the results so a comparison could be made between human and AI ability to detect authorship. Surprisingly to the research team, GPT-4 was more effective at detecting AI-generated pieces than humans. Moreover, in effectiveness, humans performed better than AI. Humans had seven true positives and 6 true negatives. Humans also had just one false positive and six false negatives. GPT 4 excelled with eight true positives, but it lacked in every other category with just two true negatives. It also had five false positives and five false negatives. GPT 4 was the only model that came even close to the effectiveness of the human subjects. However, GPT 4 has shown that it would still be an unreliable tool for the detection of text documents due to its alarmingly high number of false positives and negatives. Furthermore, this data showed that Gemini and Claude are practically useless as detection tools for AI-generated text documents.



Text Documents	ChatGPT-4	Gemini Advanced	Claude Opus	Humans
True Positive (+,+)	8	1	1	7
True Negative (+,-)	2	5	4	6
False Positive (-,+)	5	2	3	1
False Negative (-,-)	5	12	12	6

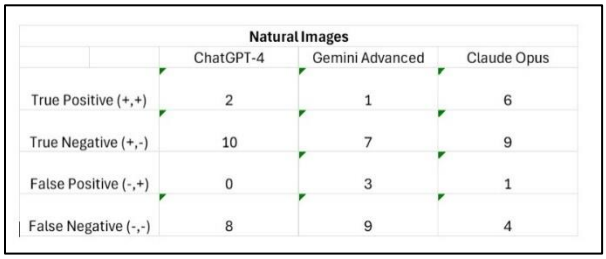
Figure 1. Results of the experiment involving text documents with the human responses

Figure 2 shows the outcome of the test utilizing images of nature. This data was the result of asking GPT 4, Gemini, and Claude to determine authorship of images of animals as well as natural landscapes. When it came to this test, Claude very obviously outperformed both GPT 4 and Gemini. Out of the 20 total data images, Claude had six True Positives, of which the second highest was only two, and nine True Negatives. Claude finished with one False Positive and four False Negatives. Far from Claude’s performance, the second best, if one can even say that performer was GPT 4. This model had just 2 True Positives and the Truest Negatives with 10. This model fell short with its eight False Negatives.

Finally, the worst performer was Gemini. This model struggled with only correctly identifying one true positive and seven True Negatives. It had three False Positives and nine False Negatives. One thing to notice is that GPT 4 and Gemini both appeared to be extra cautious when predicting an AI author with both only choosing AI a total of six times out of 40 combined trials. For some reason, these two models lean heavily towards human authorship even if it is an

ISSN: 2473-4901
v10 n6140

incorrect guess. On the other hand, Claude was more confident in its abilities and guessed AI authorship a total of seven times out of its 20 trials. Another interesting thing to note is that no AI-generated image was identified as AI by all three models. Only two images were correctly identified by more than one AI: GPT 4 tagged two images as AI, and those two were among six identified by Claude, as shown in Figure 2 below.

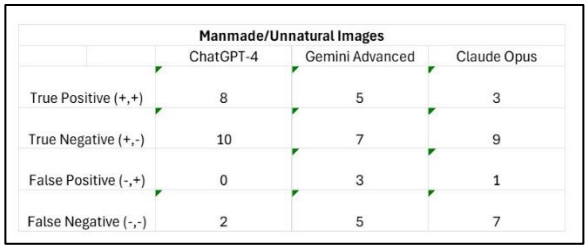


Natural Images	ChatGPT-4	Gemini Advanced	Claude Opus
True Positive (+,+)	2	1	6
True Negative (+,-)	10	7	9
False Positive (-,+)	0	3	1
False Negative (-,-)	8	9	4

Figure 2. Results of the experiment involving images of nature

Figure 3 shows test results involving the images of manmade objects was very interesting. Compared to the other tests, all models appeared to perform relatively well at this task. GPT 4 was the leader again with a total of eight true positives and a perfect 10 true negatives while only misidentifying two of the AI-generated pieces as human work. The next top performer was Gemini which had a total of five true positives and seven true negatives. Gemini struggles on eight pieces where three were misidentified as false positives and five as false negatives.

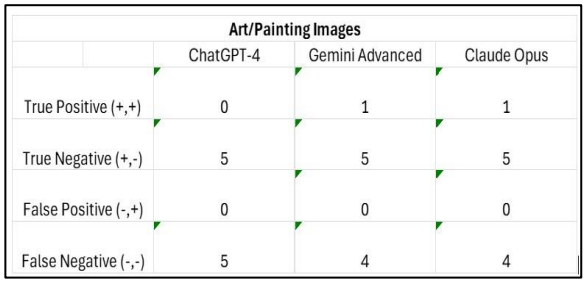
Bringing up the rear, although not too far behind the other two models is Claude. Claude had just three true positives with a high 9 true negatives. Claude also had one false positive and a total of seven false negatives. All three models appeared to find it easier to identify human-made pieces as such as only four total times a human-made piece was falsely mistaken to be AI. This appears to be a continuing trend as with all data sets, the models typically leaned to identify works as human-made. It was quite surprising that for manmade objects Claude performed worse than natural objects where it was the best at identifying authorship. While on the other hand, GPT 4 and Gemini performed much better at identifying the creator of the manmade works compared to the natural images.



Manmade/Unnatural Images	ChatGPT-4	Gemini Advanced	Claude Opus
True Positive (+,+)	8	5	3
True Negative (+,-)	10	7	9
False Positive (-,+)	0	3	1
False Negative (-,-)	2	5	7

Figure 3. Chart of results from the experiment involving manmade images

Lastly, Figure 4 shows the test results for the data that represents art will be examined. Quite frankly, all AI models performed horrendously in this test. Paintings and art were chosen purposely by the research team as a dataset because the research team believed it would be extremely difficult for the AI to distinguish between art created by humans by that created by AI; however, the results were even worse than the research team imagined. All three models performed nearly equally poorly. Gemini and Claude had the same results with only one true positive and five true negatives. They then had four false negatives each. GPT 4 had five true negatives and five false negatives. These results show that when it came to art and images of paintings, the models were extremely reluctant to presume an AI author. It appears as if the models just automatically defaulted to human creator for all pieces of this nature. This raises the question of whether the models could tell the human pieces were made by humans or if these were only correct because the models automatically chose human every single time. This trail shows that artificial intelligence appears to be a while from being able to truly identify art as human or AI-generated.



Art/Painting Images	ChatGPT-4	Gemini Advanced	Claude Opus
True Positive (+,+)	0	1	1
True Negative (+,-)	5	5	5
False Positive (-,+)	0	0	0
False Negative (-,-)	5	4	4

Figure. 4. Results for images involving art.

6. CONCLUSION

Although the research team created a starting point for evaluating multiple AI models' ability to detect AI-generated works, it is simply a starting point. Some limitations need to be discussed for future research into this idea. One area this study lacked was the size of the data sets. The

experiment used diverse data sets; however, there were simply not enough, and in the future, it would be preferable to expand this experiment with much larger data sets across dozens (or, by that time, perhaps hundreds) of AI models. This would help limit any bias that may be present in a smaller data set. Although, based on the results shown, it appeared to have little to no effect on the conclusions made by the models, future research should produce AI-generated data samples with a model that is not being observed in the study. This too could limit bias between the models' responses.

GPT-4 performed like a premium AI in three out of four of the tests. As shown in the charts above, in both tests involving text documents and manmade images, GPT 4 performed better than Gemini and Claude. However, when it came to the natural images and the art tests, GPT 4 performed poorly. In terms of use, GPT 4 was one of the easiest models. It rarely ever asked a user to provide more information before revealing its decision. It also was correct most of the time when it was describing what was present in the images regardless of whether it knew who created the images or not. One of the only drawbacks found during this experiment was that at one-point GPT 4 did require the user to wait about two hours before most questions could be asked to the model.

During this study, Gemini performed very average. In every experiment, it was always in the middle on accuracy. It never performed better than any of the other models. This does not mean it is a bad model, it is just not the greatest for the tests conducted in this study. Gemini had a few major drawbacks. One was that it had to constantly be coerced to answer when asked to choose human-made or AI-generated. Constantly, Gemini would say that more information is needed or that it cannot decide, so the research team would have to repeatedly state that it must give an answer. Furthermore, another drawback that was noticed was that although it may be able to identify authorship, Gemini did not always understand what it was even looking at. For example, the image below shows how Gemini says that the image of steak on pasta is a natural landscape with mountains, lakes, and a boat. This begs the question of how well the model processes the content of images, or if it just makes the best guess, it can.

Throughout the study, Claude performed moderately well. It was the most effective model when it came to natural images; however, when it came to manmade images, texts, and art pieces it was around average and performed similarly to Gemini. Like GPT 4, Claude was relatively easy to coerce into giving its opinion on the creator of the work and it required little to no extra questioning for an answer to be provided. Claude had two major drawbacks. The first was that in a series of questions, Claude only allowed a user to submit five images before a new chat must be created. This made keeping all of its information in 1 place impossible. The biggest drawback to Claude was its limited question space. Even with a $20 monthly price tag, the research team was only able to submit about 12 questions to the model every five hours. This made using Claude extremely frustrating. Due to this reason, the research team believes that GPT-4 is a better option at the same monthly subscription cost of $20.

Although this study showed that no current AI model can effectively determine whether a piece was created by artificial intelligence or not with 100% accuracy, its findings are still meaningful for the future. As models continue to progress, there will continue to be an uprise of deep fakes, and these malicious works show no sign of stopping soon. With, tools and AI models should still be used by humans despite their inaccuracies as simply a guide to provide further information to the humans' own opinions about a piece. If used as a tool to add information to the bigger picture and not as the final word, AI models could still prove effective as identifiers of deep fakes.

7. REFERENCES

Anthropic (2023). Claude Documentation. Retrieved June 4, 2024 from https://docs.anthropic.com/.

Anthropic (2023). Claude Opus. Retrieved May 5, 2024 from https://www.anthropic.com/claude.

Bhattacharjee, A., Liu, H. (2024). Fighting Fire with Fire: Can ChatGPT Detect AI-generated Text? *SIGKDD Explor. Newsl. 25* (2), 14–21. https://doi.org/10.1145/3655103.3655106

Borji, A., Mohammadian, M. (2023). Battle of the WordSmiths: Comparing ChatGPT, GPT-4, Claude, and Bard. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.4476855.

Bouafif, M. S., Zheng, C., Qasse, I. A., Zulkoski, E., Hamdaqa, M., & Khomh, F. (2024). A Context-Driven Approach for Co-Auditing Smart Contracts with The Support of GPT-4

ISSN: 2473-4901
v10 n6140

code interpreter. arXiv preprint. https://doi.org/10.48550/arXiv.2406.18075.

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., ... & Xie, X. (2024). A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, *15*(3), 1-45.

Clements, B., Abegaz, T., Payne, B. (2024) Anomaly Detection: Distinguishing AI-Generated Works from Student-Created Submissions. *Proceedings of the International Academy of Business Disciplines 2024.* Las Vegas, NV. April 4-6, 2024.

Conte, N. (2024). Ranked: The most popular ai tools. Visual Capitalist.

Dash, D., Thapa, R., Banda, J. M., Swaminathan, A., Cheatham, M., Kashyap, M., ... & Shah, N. H. (2023). Evaluation of GPT-3.5 and GPT-4 for supporting real-world information needs in healthcare delivery. *arXiv preprint*. https://doi.org/10.48550/arXiv.2304.13714

Google. (2024). Gemini Advanced. Retrieved March 3, 2024 from https://gemini.google/advanced/.

Karwa, S. (2023) Exploring Multimodal Large Language Models: A Step Forward in AI. Retrieved March 25, 2024 from https://medium.com/@cout.shubham.

Liu, T., Chatain, J., Kobel-Keller, L., Kortemeyer, G., Willwacher, T., & Sachan, M. (2024). AI-assisted Automated Short Answer Grading of Handwritten University Level Mathematics Exams. *arXiv preprint.* https://doi.org/10.48550/arXiv.2408.11728

Moore, S., Schmucker, R., Mitchell, T., & Stamper, J. (2024). Automated Generation and Tagging of Knowledge Components from Multiple-Choice Questions. In *Proceedings of the Eleventh ACM Conference on Learning@ Scale*. Association for Computing Machinery, New York, NY, USA, 122–133. https://doi.org/10.1145/3657604.3662030

OpenAI. (2023). GPT-4. Retrieved March 31, 2024 from https://chat.openai.com.

OpenAI. (2023). GPT-4 Technical Report. arXiv.org. https://doi.org/10.48550/arXiv.2303.08774

Poldrack, R. A., Lu, T., & Beguš, G. (2023). AI-assisted coding: Experiments with GPT-4. arXiv.org. https://doi.org/10.48550/arXiv.2304.13187

Saravia, E. (2022). Prompt Engineering Guide. Retrieved April 2, 2024 from https://www.promptingguide.ai/