The Proficiency of Various AI Models Within Deepfake Detection

Sai Prabhav Vankamamidi, Rishon Ishwa Rajesh, Ali Syed Raza, Daksha Jamallamudi, Arjun Sriram, Tanay Sethiya, Shalaan Umar, Siddharth Rao, Risha Jamallamudi, Aarush Verma and Joseph Kurian
DiscoverSTEM, Plano, TX, USA

spvankamamidi@gmail.com
rishonishwa@gmail.com
alisyedraza18@gmail.com
daksha.jamallamudi@gmail.com
arjun.sriram.p@gmail.com
sethiyatanay@gmail.com
shalaanumar@gmail.com
sidd.r.rao@gmail.com
rishajamallamudi@gmail.com
aarushv153@gmail.com
josephkuran380@gmail.com

Abstract: As AI evolves, it is increasingly challenging to distinguish authentic media from deepfakes. This encourages fraud, misinformation, and legal ambiguity. Despite the existence of specialized deepfake detection tools, consumer-targeted AI agents used by millions have not been systematically evaluated for this task under controlled conditions. This study evaluates the reliability of widespread AI agents in deepfake detection, utilizing human performance as a control group. If AI can generate deepfakes, then it should be equally equipped to detect deepfakes, with a minimum accuracy of 90% across all aspects. To evaluate this, four models were assessed—ChatGPT 5.2, Gemini 3.1 Pro, Claude Sonnet 4.6, and Super Grok—on approximately 300 royalty-free images within four categories: boys, girls, cats, and dogs. To ensure consistency, metadata was stripped using PrivMeta, watermarks were removed with EZRemove V2, file names were randomized with a Java program, and temporary chat sessions were utilized to prevent pattern formation. The decision and confidence score—from 1 to 10—of each AI model were recorded. Three human participants of ages 11, 19, and 48 emulated this for comparison. The hypothesis was disproved by the results. Across all models and trials, the aggregate accuracy on real images was greater than deepfakes, at 89.64% and 65.98% respectively. Gemini demonstrated the greatest performance with an accuracy of 85.79% and also detected deepfakes with more accuracy than real images, at 93.94% and 79.75% respectively. Claude, ChatGPT, and Grok followed with scores of 83.47%, 78.19%, and 68.52% in that order. All models performed better at detecting humans as opposed to animal images. Gender bias was shown towards male faces, with greater accuracy than that of females. These findings reveal that the most publicly-available AI tools are the least qualified to defend against the deepfakes that other AI models generate. As deepfake quality increases, this imbalance reduces public trust of journalism, legal evidence, and political discourse. Reliable deepfake detection should be treated as a premium feature, rather than a public safeguard integrated into the platforms millions use every day.

Keywords: Deepfake detection, AI accuracy, Generative AI, Pattern recognition, Flagship AI

1. Introduction

> ***Hypothesis:*** *If AI can generate deepfakes, then it should be equally capable of detecting deepfakes, with a minimum accuracy of 90% across the board.*

The rapid growth of generative artificial intelligence and deepfakes has decreased digital credibility, with synthetic media being a primary source of distrust in the World Wide Web. A “deepfake” is a synthetic piece of media that is made by generative AI and DL (Deep Learning), which uses techniques to convincingly misrepresent and mimic a person or animal. The social impact of deepfakes is substantial, ranging from people’s views of a political figure changing after seeing a deepfake video to the danger posed by creating non-consensual explicit deepfakes (Alverson et al., 2024). Advances in artificial intelligence have led to the frequent worldwide use of GAN (Generative Adversarial Networks) to enhance deepfakes by creating a generator capable of deceiving a discriminator, which is an AI that attempts to distinguish if images are real or deepfake (Goodfellow et al., 2020). Bray, Johnson & Klienberg (2023) found that humans are inaccurate at detecting deepfake media, and could only correctly detect approximately 60% of deepfakes with human faces, which highlights the increasing reliance on artificial intelligence in order to identify deepfakes.

While prior research in the field of deepfake detection has explored a large range of problems with deepfakes and the detection of deepfakes, several gaps remain about this topic. Trinh & Liu (2021) found that the data

used to train many AI models, FaceForensics++, has racial and gender bias of female Caucasians, which causes ethnicities such as Indians and African Americans to be underrepresented in the expert tools used to train many generative AI models. Additionally, the dual-use nature of many artificial intelligence models (both creating and detecting deepfakes) and the "arms race" between generation and detection of AI models is emphasised, and has been discussed by Sharma & Selwal (2026). However, much existing research covers expert-designed AI training tools, or is theoretical in nature. As a result, the ability of Large Language Models (LLMs) to identify deepfakes in a controlled environment remains untested.

This research paper addresses the large knowledge gap by using flagship LLMs, such as ChatGPT, Gemini, Claude, and Grok, to evaluate their ability to detect deepfakes across the 4 categories of boys, girls, cats, and dogs under controlled conditions. By using control variables such as metadata stripping, file order randomization, and watermark removal, this research paper aims to test these models' capability of using visual clues to identify deepfakes. The confidence of each artificial intelligence model was noted not only to assess accuracy, but also to test each AI's reliability at detecting generative images. The findings of this study provide insight into the reliability of large language models, and whether they can serve as a tool to distinguish deepfakes from real images.

2. Methodology

This is an empirical study that was conducted from November 2025 to May 2026 in a controlled digital environment at DiscoverSTEM. This study did not require any ethical or copyright clearance, since it utilized royalty-free images.

This research employed a detailed experimental design to test and benchmark different AI models' ability to detect synthetic and real images. It involved a multi-step process for each individual AI agent in order to ensure consistent results across all trials. Researchers could classify an image as a deepfake if they generated it, it was explicitly presented as such, or pulled from a dataset. This selection of images encompassed images from various categories, the majority of the general themes being stock photos, pictures or portraits of celebrities, and royalty-free selfies. These images came from sources such as Tiktok, Instagram, and some were generated by Grok to test the dual nature of flagship AIs. Images were excluded if they couldn't be identified as authentic or deepfake, included minors, or were inappropriate for the experiment. The final dataset consisted of exactly 294 images, with 132 of them being deepfake images and 162 of them being authentic. After splitting the images into the main 4 categories (boys, girls, cats, and dogs), the dataset included 76, 71, 75, and 72 images, in that particular order. Utilizing this database of images, researchers tested the accuracy of deepfake detection on four different AI models, being Claude, ChatGPT, Grok, and Gemini. The experiment involved categorizing each image—utilizing Google Drive and the test computers' local storage—in order to be able to compare results holistically as well as specific to each genre. Data was collected from various sources across the web in order to decrease the margin of error of the experiment.

The use of real image samples, metadata stripping through PrivMeta (an open-source metadata stripping software), watermark removal tools such as EZRemove V2, randomized order of images, and the temporary chat feature prevented confounding variables from posing an issue and allowed each AI to rely on its own capabilities. Confounding variables may include low-resolution photos, inconsistent shadows, and certain widespread photos being easily accessible. As such, researchers were able to benchmark and conclude information about each AI's effectiveness in identifying synthetic photos whilst maintaining uniformity.

Data collection and analysis was done in a data sheet with columns for the Unique ID, File Link, the broad category as either Human or Animal, the specific category as Boy/Girl or Dog/Cat, deepfake or real, the AI prediction, AI response, whether it was correct or wrong, any remarks related to the AI and/or dataset, and AI confidence from 1 to 10. The researchers made these columns in order to record the data to include in the results. In order to ensure unpredictability, the images were named based on a random sequence of numbers 1 through 999 sourced from a Java program. The AI models used included ChatGPT 5.2 Plus/Pro, Gemini 3.1 Pro, Claude Sonnet 4.6 Extended, and Super Grok, each being the flagship model for its respective AI. This was done in order to evaluate the maximum capabilities of each AI model. Each AI was fed the following prompt:

> *"I will give you a photo. Based on visual evidence, determine if the image is real or a deepfake and give a ranking of how confident you are on the image, where: 1 = not confident at all; 10 = extremely confident. Briefly explain the key factors that influenced your rating. If evidence is limited, reflect that in your score."*

As such, one image was fed at a time, and data was recorded throughout. Analysis was based on 3 metrics: the percentage that each AI predicted correctly individually and comparatively in broad and specific categories; the average confidence of each respective AI when correct; and the same for incorrect. To ensure consistency and to eliminate any potential gaps within the study, 3 trials were performed for each AI agent, resulting in the creation of 12 unique data collection sheets in total for the AI agents. The comments given by the LLMs were not necessarily included as part of the study, but rather to ensure that no human error was made when inputting the image and/or to diagnose any potential issues with the system of processes.

In order to analyse the impact of the study, comparison to the deepfake detection abilities of humans was necessary. To account for the majority of age groups, 3 participants—aged 11, 19, and 48—chose to participate anonymously in the trials. The same database and experiment given and administered to the flagship AIs, along with the standardized data table for efficient data collection were used to test the humans on their ability to detect deepfake media. Similar to the AI trials, participants were fed 1 image at a time and asked for their confidence on their decision on a scale from 1 to 10. Participants were also given the option to provide additional comments, which would not impact consistency since the LLMs were also allowed to do so. Similar to the metadata stripping process discussed in the AI methodology section, humans were only allowed to see individual images, could not use internet sources, and were instructed to rely on their prior knowledge.

To prevent the confounding variable of fatigue being an issue, participants were given breaks when needed and only 1 trial was performed for each participant, which also addresses the issue of participants using pattern recognition. All 3 trials were performed simultaneously, in separate rooms, within a 3 hour period, and within the same controlled environment that was used for the AI portion of the study. As a result of 1 trial per participant, 3 spreadsheets—in the same format as done with the 4 LLMs—were created to collect data. For each image, if the prediction made by each participant matched the image classification given, then it was recorded correct. Subsequently, this data was used to create the data tables shown in the results section.

3. Results

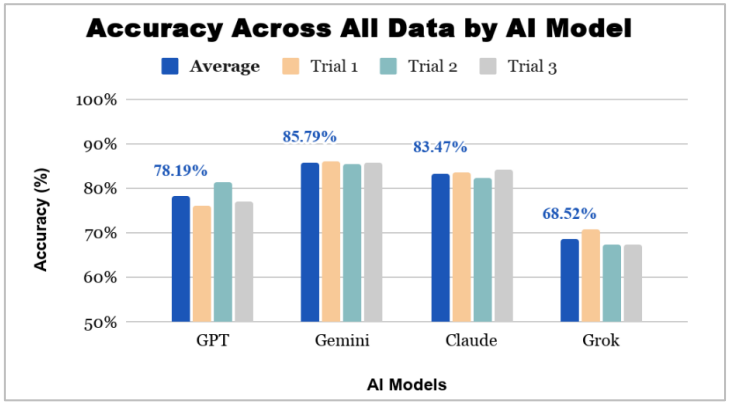

Figure 1: Aggregate full-scale accuracy across all three trials by AI Model, with Gemini leading at 85.79% average accuracy

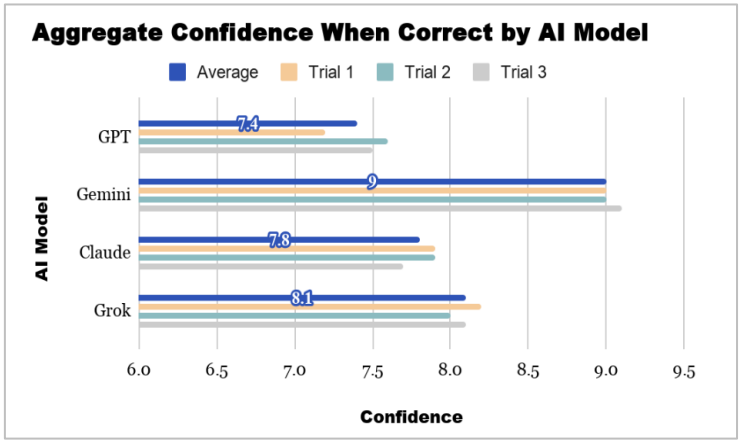

Figure 2: Aggregate confidence across all three trials by AI Model when correct, with Gemini leading with a 0.9 higher average confidence score than the trailing LLM, Grok

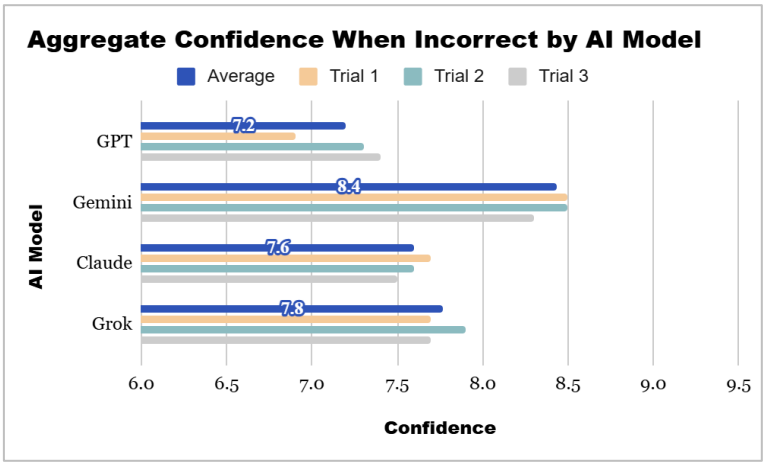

Figure 3: Aggregate confidence across all three trials by AI Model when making an incorrect decision, with Gemini leading with a 0.8 higher average confidence score

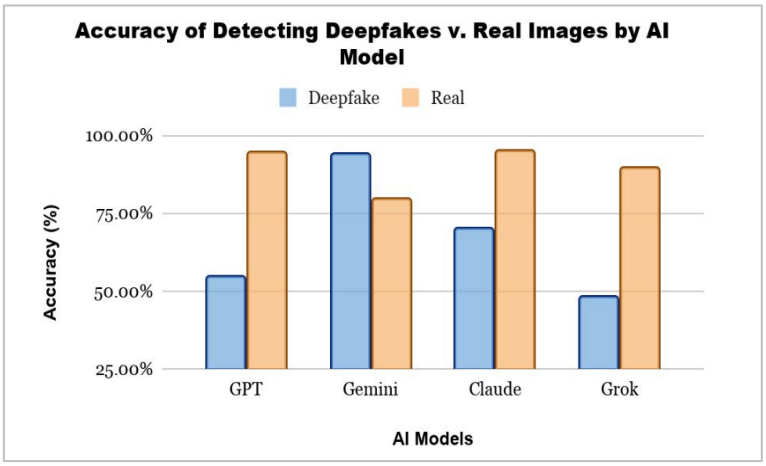

Figure 4: Compared aggregate accuracy across all three trials by AI Model in deepfake v. real image detection, led by Gemini in deepfake detection and Claude in real image detection, with accuracies of 93.94% and 95.06%

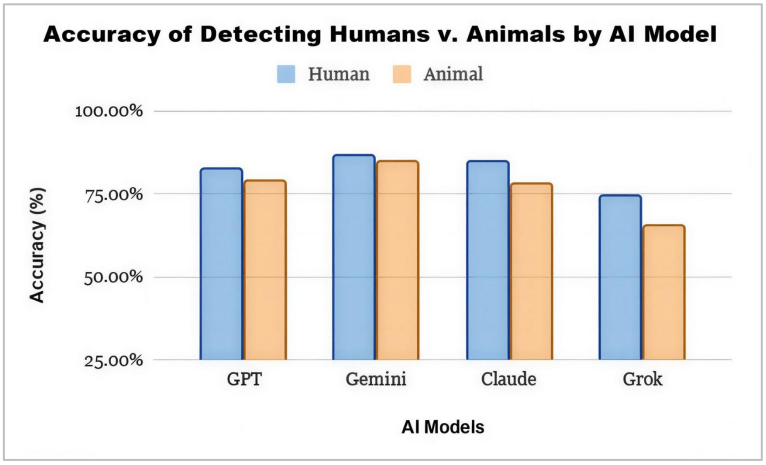

Figure 5: Compared aggregate accuracy across all three trials by AI Model in human v. animal image detection, with Gemini leading by a small margin of 1.93% in human image detection

	ChatGPT	Gemini	Claude	Grok
Average	78.19%	85.79%	83.47%	68.52%
Trial 1	76.27%	86.10%	83.73%	70.85%
Trial 2	81.36%	85.52%	82.37%	67.46%
Trial 3	76.95%	85.76%	84.30%	67.26%

Figure 6: A table demonstrating the aggregate accuracy across all three trials for all four LLMs, with Gemini showcasing the strongest performance with an average accuracy of 85.79%

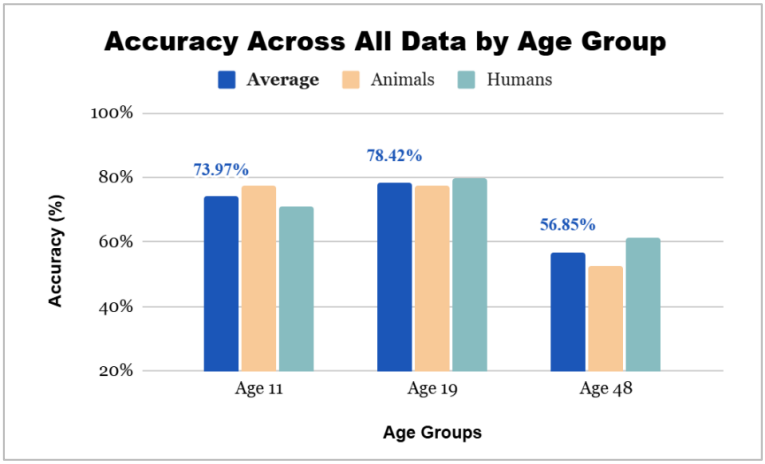

Figure 7: A table demonstrating the aggregate accuracy across both trials for all 3 age groups, with the young adult group (the age 19 participant) leading, with an average accuracy of 78.42%

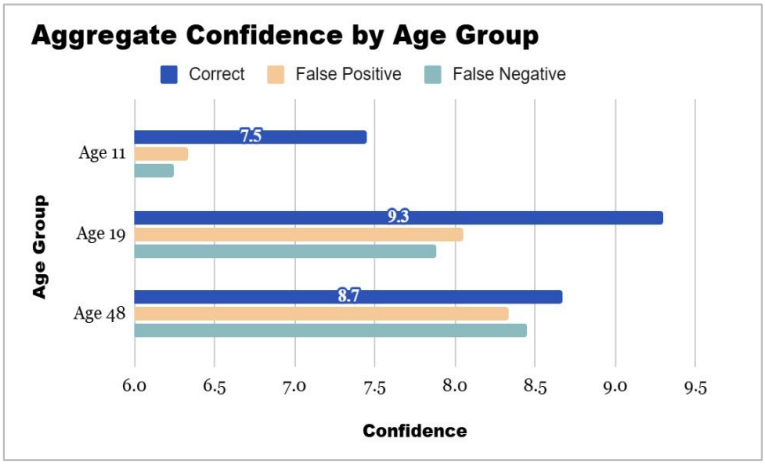

Figure 8: A table demonstrating the aggregate confidence across both trials for all 3 age groups, with the young adult group (the age 19 participant) leading, with an average confidence 0.6 points higher than the trailing age group, the adult group (age 48)

4. Discussion

Across the 3 trials that were conducted, several patterns appeared in the data. Gemini was the most consistent in its results, and had the highest accuracy in detecting both deepfake and real images across all trials. Claude trailed closely behind, with Gemini leading by a bit over Δ+2%. ChatGPT had the most “average” performance, having done very well on real images but significantly less consistently and accurately on deepfakes, especially for animals like cats and dogs. Claude was slightly more consistent than ChatGPT, but they both relied on real

images to keep their average up. Grok had the lowest accuracy and struggled the most with deepfakes. Overall, the AI Models performed better on human images than animal images, and are significantly better at identifying real images as opposed to deepfake detection.

Across all trials, each AI occupied its own "niche" in terms of its capabilities. In terms of the "best trial" for each AI in overall image detection capabilities, Gemini did the best (85.79%), followed by Claude (83.46%), then ChatGPT (78.19%), and finally Grok (68.52%). With the exception of Gemini, all three LLMs were better at detecting real images as opposed to deepfakes, with a substantial increase of Δ+36.34%. On the other hand, Gemini performed noticeably better in deepfake detection as opposed to that of real images, with an aggregate increase of Δ+14.39%. All AI Models did well on real image detection, with Gemini's accuracy at 79.75% being the lowest of the pack and Claude leading at 95.06%, followed by ChatGPT and Grok with aggregate scores of 94.41% and 89.51%, respectively. Despite this, deepfake detection proved to be a problem for these AIs, with Claude scoring the second highest at just 69.92%, followed by ChatGPT at 54.89%, and Grok at 48.12%. However, Gemini particularly shined in terms of deepfake detection as opposed to its mediocre performance in detecting real images, with an aggregate accuracy of 93.94%.

Across the different categories, the AI Models performed significantly better in terms of detecting human images, with all models scoring higher than on animals. Gemini led the pack at 86.49%, with Claude trailing closely behind at 84.56%. ChatGPT and Grok followed at 82.43% and 74.50% accuracy, respectively. This was the subcategory that Grok relatively did best on, with a trail of less than 10% behind the second worst performer, being ChatGPT.

On the other hand, the accuracy of all of the models significantly dropped in terms of correctly detecting animal images. Gemini demonstrated upwards of a Δ+5% lead over the other AI Models with a performance of 84.84%, with ChatGPT and Claude having less than a Δ±1% difference between each other, following suit at 79.07% and 78.08% respectively. Grok, as usual, scored the least, with an average accuracy of 65.22%. Although not as noticeable as the AI gender bias, there was a noticeable variance in accuracy across all AI Models in detecting images within the categories of Animal/Cat and Animal/Dog.

Since the beginning of AI training, databases have featured significantly more male images as opposed to female images, generally leading to "AI bias," a concept in which AI Models perform better in detecting images they were trained on more as opposed to less. This was clearly demonstrated in this experiment, with differences between the genders being as little as Δ-1.83% for ChatGPT to as much as Δ+6.88% in Gemini's case, generally either staying consistent with a slight edge towards Human/Girl detection accuracy or performing significantly better in image detection within the Human/Boy category. Gemini and Claude, the two top-performing AI Models consistently throughout the experiment, generally performed better on human males, with Gemini having performed Δ+6.88% better and Claude performing similarly, with a Δ+6.15% increase in accuracy. However, ChatGPT and Grok took part in the opposite trend, with a decrease in accuracy from females to males of Δ-1.83% and Δ-4.81%, respectively.

Another important note is the consistency of the models' results. Gemini is both the highest-performing and most consistent LLM with its results, with Claude having similar results in terms of consistency. Both of these top-scoring AI Models each had less than a 2% variance between all trials in all aspects of the experiment. As mentioned before, Gemini had the least substantial change in accuracy between real and deepfake images, with an aggregate change of Δ+14.14% across all three trials. Claude also demonstrated relatively consistent results between the categories with a change of Δ-25.14%. Grok and ChatGPT clearly demonstrated strong performance in detecting real images, with ChatGPT having performed the best out of all AI models in real image detection and Grok performing the best in real image detection as opposed to other categories, ranking at 1st and 3rd place, respectively. However, the accuracy of either AI dropped significantly when tasked with deepfake detection, with change rates of Δ-39.52% and Δ-41.39% from real to deepfake images, for ChatGPT and Grok, in that particular order.

There were three humans who volunteered for this experiment of the following ages: 11, 19, and 48. This age range encompasses the majority of the audience who might see deepfakes throughout their lives, with the age groups pertaining to youth, young adults, and adults, respectively, fitting the goal of the experiment. The accuracy of each age group aligns with the amount of exposure to AI that the participants could have had, with the young adult age group scoring the highest at 78.42% across the board, followed by 73.97% and 56.85% for the youth and adult age groups, respectively.

During the experiment, all of the human participants used their prior knowledge to the best of their abilities. The youth participant used their context as opposed to a detailed analysis of each image, and when given some glaringly real images that they were never exposed to, that often hindered their abilities to do better. However, the young adult participant used both their context and analysis, often referencing seeing something on social media or analysing the expression of the human or animal to determine the outcome. The adult participant did a shallow analysis, often being thrown off by things that in their perspective, are anatomically impossible, similar to many of the AIs, as well as referencing celebrities from the current day and in history to help their judgement. Overall, each age group used their own method of decision making based on their knowledge and abilities.

Due to the fact that human accuracy is heavily determined by their exposure to AI, the data was varied across all three participants. The youngest participant had the most noticeable change, with a decrease of Δ-22.28% from deepfake to real images, possibly due to increased exposure to deepfake images and a lack of historical context or access to social media. This is further backed up by the fact that both the young adult and adult participants saw an increase in accuracy, with rates of change of Δ+21.79% and Δ+14.5%, favouring real images. This data can possibly stem from the levels of analysis as well as the methods of decision making used.

Since humans see both men and women every day, theoretically, there should be no gender bias among the volunteers of this experiment. Interestingly enough, there was reverse gender bias, with the youngest participant having a decreased accuracy of Δ-4.82%, followed by Δ-6.78% and Δ-6.9% for the middle and oldest participants, respectively. Between Animal/Dog and Animal/Cat, there was little variance for the two younger age groups but a noticeable decrease from dog to cat with a rate of Δ-8.99%. However, the other two age groups both had variances of less than Δ±2.00%, with the youth representative observing an increase of Δ+1.07% and the young adult having a decrease of Δ-1.69% between dogs and cats. As for the broad categories, due to constant exposure to both animals and humans alike, there was little variance between the categories for all three participants, with the youngest experiencing a Δ-6.49% rate of change between animals and humans, while both the young adult and adult participants were scoring higher in terms of detecting humans, with changes of Δ+2.35% and Δ+8.81% in that order.

Confidence is also an important metric for both AI and humans. The highest scoring participant and AI Model tend to have higher confidence in their decisions, as observed through analysis of the young adult participant and Gemini, the two highest scoring participants in their respective comparisons. Gemini had an average confidence of 9 across the board when correct, while the young adult participant had an aggregate confidence of 9.3, which is backed up by their high accuracy. Surprisingly, the lowest scoring participant and AI Model of each comparison had the second highest confidence, with the adult participant having an average confidence of 8.7 and Grok having an aggregate confidence of 8.1 when correct. This pattern is also present for AI Models in false positive and false negative decisions, with Gemini having a confidence of 8.4 and 8.7 in that order, and Grok, with the second highest confidence across the board, having an average confidence of 7.8 for both false positives and negatives. However, the humans and AI Models with average scores tend to be more “modest” with their confidence in themselves, with the youth participant having confidence scores of 7.5 when correct and 6.3 when detecting false positives or false negatives.

Limitations were present in this experiment that potentially affected the interpretation of the results. SynthID, a tool developed by Google, places a digital watermark invisible to the human eye on AI generated photos made by Google products such as Nano Banana. Due to the nature of the watermark removal software as well as human capabilities, such watermarks were not removable. To ensure consistency, Gemini trials were redone without the tool and the absence of the tool actually improved its accuracy, improving by a bit under Δ+2% on average. This could possibly stem from the nature of Gemini’s reliance on SynthID for many decisions, potentially clouding its judgement. All other LLMs had no such issue.

Additionally, there were limitations in the nature of this experiment as well. Many of the photos that were gathered and categorized within the “Human” category contained famous people. Despite this, very few of them were deepfakes, limiting the uniformity and accuracy of the results of that particular part of the experiment. Furthermore, 2 out of the 4 utilized LLMs had a daily chat limit. ChatGPT had an input quota of 80 messages per hour, whereas Grok had a limit of 100 for every 2 hours. This decreased the efficiency of some aspects of the experiment.

Holistically, the data demonstrates that no single AI was the best at everything. Some models, such as Gemini, were very strong at detecting deepfakes, while others, like ChatGPT, were better at recognizing real images. Simultaneously, Claude showcased consistent results across most categories, while Grok was consistently the

poorest performer except for real image detection. The two top-scoring AIs, Gemini and Claude, not only demonstrated the highest overall score, but also the highest scores in each individual category.

5. Conclusion and Future Work

The accuracy of all AI Models in predicting deepfake images for all trials was 65.98%, while the aggregate score for real images was 89.64%. The hypothesis states that if AI can generate deepfakes, then it should be equally capable of detecting deepfakes; therefore, the AIs should have a grade of at least 90% for all categories. However, very few of the AIs earned close to this score. Additionally, the performance of the AI Models was consistently higher whilst evaluating human images as opposed to animal images, similar to how the performance was also consistently higher with male images as opposed to female images. The idea that there was gender and human bias in the datasets used to train the flagship AIs is heavily supported by these results.

The future of AI implementation in deepfake detection holds great potential to increase accuracy across the board. With a high accuracy observed in this research, it can help in real-world applications. Instead of being a resource, AI can be built directly into many platforms. This could verify whether the content in unique pieces of multimedia is real or false. Ultimately, improvements such as AI being directly built into platforms and verifying if multimedia is real or AI-generated can prevent misinformation in media across the internet, and by extension, the whole world.

Acknowledgements

DiscoverSTEM, a non-profit organization that aids in offering college scholarships to young students and innovators, is the site at which experiments were performed, data was gathered, and provided the framework for the whole project. Hajer Janabi, a rising young innovator and preliminary medical student, provided the guidance to the authors as the head of research at DiscoverSTEM, and showcased unwavering dedication and support throughout the project. Ayaan Haq—a young entrepreneur and a mentor to the authors—, Kayla Goodrich—a scholar immersing herself in the world of AI and the Head of Innovation Mentor at DiscoverSTEM—, Baneen Ali Ahmed—a future dental student and AI hobbyist as well as a mentor to the authors—, and Kashyap Samudrala—a young and rising innovator—, all provided support and utilized their knowledge and expertise to support the authors throughout the entire project.

Ethics Declaration: Ethical clearance was required for this project because human participants were involved. However, no personal or private records were recorded or used, and the 3 human participants remained anonymous, with the only personal detail that was utilized for research being their age. This project was overseen and funded by DiscoverSTEM, from October 2025 until May 2026. Formal IRB Approval was not required as per local guidelines, and all procedures followed standard ethics guidelines to ensure data integrity.

AI Declaration: Artificial intelligence was not used for brainstorming or writing the paper. However, AI was used to aid in developing the visual aids (figures, charts, graphs, etc.) as well as proofreading the paper to rid it of any mistakes. All analysis and final manuscript decisions were taken by the authors.

References

Bray, S.D., Johnson, S.D. and Kleinberg, B. (2023). Testing human ability to detect 'deepfake' images of human faces. *Journal of Cybersecurity*, [online] 9(1). doi:https://doi.org/10.1093/cybsec/tyad011.

Diel, A., Lalgi, T., Schröter, I.C., MacDorman, K.F., Teufel, M. and Bäuerle, A. (2024). Human performance in detecting deepfakes: A systematic review and meta-analysis of 56 papers. *Computers in Human Behavior Reports*, [online] 16, p.100538. doi:https://doi.org/10.1016/j.chbr.2024.100538.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM*, [online] 63(11), pp.139–144. doi:https://doi.org/10.1145/3422622.

Hamilton, E. (2026). *Machines spot deepfake pictures better than humans, but people outperform AI in detecting deepfake videos*. [online] Ufl.edu. Available at: https://news.ufl.edu/2026/02/deepfake-detection/.

Press, G. (2024). Detecting Deepfakes: Fighting AI With AI. *Forbes*. [online] 27 Aug. Available at: https://www.forbes.com/sites/gilpress/2024/08/06/detecting-deepfakes-fighting-ai-with-ai/.

Sharma, S. and Selwal, A. (2026). Potential of artificial intelligence in deepfake media: From generation to detection mechanisms, state-of-the-art, and challenges. *Computer Science Review*, 60, p.100866. doi:https://doi.org/10.1016/j.cosrev.2025.100866.

Trinh, L. and Liu, Y. (2021). An Examination of Fairness of AI Models for Deepfake Detection. *arXiv:2105.00558 [cs]*. [online] Available at: https://arxiv.org/abs/2105.00558.

Villasenor, J. (2019). *Artificial intelligence, deepfakes, and the uncertain future of truth*. [online] Brookings. Available at: https://www.brookings.edu/articles/artificial-intelligence-deepfakes-and-the-uncertain-future-of-truth/.