SAECULUM
Vol. 58, nr. 2, 2024, p. 59 - 70
Online ISSN 2601-1182

EVALUATING PHOTOGRAPHIC AUTHENTICITY: HOW WELL DO CHATGPT 4O AND GEMINI DISTINGUISH BETWEEN REAL AND AI-GENERATED IMAGES?

EDUARD-CLAUDIU GROSS
Universitatea „Lucian Blaga" din Sibiu
eduardclaudiu.gross@ulbsibiu.ro

DOI: 10.2478/saec-2024-0018
ORCID: 0000-0002-6842-1173

Title: Evaluating photographic authenticity: How well do ChatGPT 4o and Gemini distinguish between real and AI-generated images?

Abstract: As artificial intelligence (AI) advances, differentiating between actual and AI-generated images has become an increasingly difficult task. The present exploratory research looks at the capacity of two leading AI models, ChatGPT and Gemini, to correctly classify genuine vs. synthetic photographs in a variety of categories, including wildlife, portraiture, and abstract art. The present research evaluates the accuracy, mistake rates, and handling of complex visual content of both models, shedding light on their strengths and limits and providing significant insights into their possible uses in combatting visual falsehood.

Keywords: *Artificial Intelligence*; *Generative Art*; *AI-generated images*; *ChatGPT*; *Authenticity*;

Research questions

- How accurately can AI models like ChatGPT and Gemini distinguish between real and AI-generated images?
- What is the error rate in identifying AI-generated images versus real stock photographs?
- Are there specific types of images or manipulation techniques that AI models struggle to identify accurately?
- How do the performance and accuracy of ChatGPT compare to Gemini in identifying real versus AI-generated images?

Essay summary

- ChatGPT outperformed Gemini with 54 correct predictions out of 60 (90% accuracy), excelling in identifying both AI-generated and real photographs. In contrast, Gemini had only 26 correct predictions out of 50 (52% accuracy), struggling particularly with real photographs, where it frequently misclassified them as AI-generated.

- Both models performed well in distinguishing images with clear realism, such as wildlife and still life, but struggled with abstract categories like food photography and landscapes, where AI-generated content closely mimicked reality.
- Gemini showed a tendency to make confident yet inaccurate predictions, especially with AI-generated images, indicating an overestimation of its capabilities. In contrast, ChatGPT demonstrated greater reliability, even when making high-confidence judgments.
- These findings suggest ChatGPT's potential as a more effective tool for image authenticity verification, while Gemini may need further refinement to handle complex, abstract, or highly stylized images.

Implications

The rise of generative AI models, such as ChatGPT and Stable Diffusion, has transformed the landscape of visual disinformation, bringing significant challenges in detecting and mitigating misinformation. As mentioned by DiResta and Goldstein (2024), AI-generated images are used by scammers to increase engagement on Facebook pages and to perform complex deceptive scams and spam aimed at misleading the user. These advanced technologies have made it easier for malicious actors to create convincing disinformation across multiple modalities, including images, texts, audios, and videos, thereby complicating traditional detection methods (Shoaib, Wang, Ahvanooey, & Zhao, 2023; Xu, Fan, & Kankanhalli, 2023). The proliferation and sophistication of AI-generated content (AIGC) have raised substantial concerns, particularly regarding the rapid generation of large volumes of targeted disinformation. This includes realistic images and fabricated testimonials, which pose serious risks to public health and safety (Menz et al., 2023). As the quality of AI-generated content becomes increasingly convincing, it is becoming more difficult for humans to distinguish between real and synthetic information, further amplifying the threat of disinformation (Karinshak & Jin, 2023; Spitale, Biller-Andorno, & Germani, 2023).

Traditional machine learning techniques, which primarily focus on automatic classification and feature extraction, have been central to combating disinformation. However, their practical application outside research settings remains limited due to a reliance on narrow datasets. This limitation underscores the need for future efforts to develop more reliable AI-based systems that can assist humans in the early detection of disinformation, rather than depending solely on fully automated solutions (Montoro-Montarroso et al., 2023). The evolution of visual disinformation, driven by the increasing sophistication and volume of AI-generated content, poses significant challenges for existing detection methods. In response, there is a critical need for advanced defense mechanisms, robust AI vigilance, and organizational preparedness to effectively counter the risks associated with AI-driven disinformation. Collaborative efforts, combining technological innovation with regulatory oversight, are essential to safeguard against the pervasive effects of generative AI-enabled disinformation campaigns.

Research indicates that human accuracy in detecting deepfake images is only marginally better than chance, with a detection rate of 62%. Moreover, confidence in these judgments does not align with accuracy, showing that people often overestimate their ability to identify manipulated content (Bray, Johnson, & Kleinberg, 2022). This underscores the risks of relying solely on human judgment to detect visual disinformation. Additionally, AI models, especially those utilizing deep learning, have outperformed humans in image classification tasks. For example, AI achieved a 97% accuracy rate in determining sex from fundus photographs, a task in which human performance lags significantly (Sakamoto, 2020). This highlights AI's ability to exceed human performance in complex visual tasks. Moreover, AI systems that replicate human visual detection show strong correlations with human performance. These models can predict human detection patterns and interpret images in ways that resonate with human intuition, making AI a valuable asset in complementing human judgment for image analysis (DeKraker, 2022).

Practical recommendations for users

Amid scammers and malicious actors using artificial intelligence (AI) to generate credible images, the phenomenon of misinformation has taken on a new dimension. This paper proposes an accessible and efficient way of working with images, where audiences can learn to verify and evaluate the visual authenticity of visual content. Given that both ChatGPT and Gemini platforms are available on a freemium basis, this makes them accessible to numerous users. This is essential as it democratizes access to cutting-edge technology, giving ordinary users a powerful tool set to assess and combat misleading content.

Accessibility of freemium platforms: Users can access these language models at little or no cost, allowing them to learn how to enter suspicious images and get detailed answers about their authenticity. For example, they can upload an image to a system such as ChatGPT or Gemini and receive an analysis that highlights possible signs of forgery (such as unnatural distortions, lighting inconsistencies, or unnatural textures).

Self-directed learning and independent verification: A major advantage of these platforms is the possibility for people to learn to verify their information sources themselves. By interacting with these AI models, users can develop critical thinking skills and learn to identify the methods and techniques used by scammers to produce misleading content. In this way, users become better equipped to distinguish between real and artificially generated images, reducing the risk of being fooled by fake news or disinformation campaigns.

User-friendliness and AI efficiency: Both ChatGPT and Gemini are designed to be intuitive and easy to use, even for those without advanced technical knowledge. This makes them a practical tool in the hands of everyday users. They can quickly verify the authenticity of an image without calling in experts or complex programs. This capability can change the way people interact with information online, giving them a shield against visual misinformation.

Findings

Finding 1: The first finding reveals that ChatGPT outperformed Gemini in image classification, making 54 correct predictions out of 60 images, while Gemini correctly identified only 26 out of 50 images.

The present finding from Table 1 evaluates the comparative performance of two large language models (LLMs), ChatGPT-4 and Gemini, in distinguishing between AI-generated and real images across six distinct categories: wildlife, still life, portrait, landscape, food photography, and abstract art. Each category comprised an equal number of real and AI-generated images, with 10 images per category—5 real and 5 generated by AI. The primary finding reveals that ChatGPT-4 consistently outperformed Gemini in accurately identifying both real and AI-generated images. ChatGPT correctly classified 54 out of 60 images, achieving near-perfect accuracy across all categories. In contrast, Gemini correctly identified 26 out of 50 images, exhibiting significantly lower accuracy, particularly in more complex categories such as food photography and abstract art. For instance, in the food photography category, Gemini made six incorrect classifications, while ChatGPT made only three errors. Notably, in the portrait category, Gemini was excluded due to its inability to analyze human images, leaving ChatGPT as the sole model tested. In this category, ChatGPT demonstrated flawless performance, accurately identifying all 10 images. Across all categories, Gemini's performance was weaker, as evidenced by 24 incorrect classifications out of 50 images, compared to ChatGPT's six errors out of 60. These findings underscore ChatGPT-4's superior ability to detect AI-generated images across diverse visual categories, positioning it as a more reliable tool for image authenticity verification. In contrast, Gemini's lower accuracy, particularly in ambiguous image types, suggests the need for further enhancement to achieve performance parity.

Table 1 *This table presents a crosstabulation of image predictions (correct vs. wrong) across different categories (wildlife, still life, portrait, landscape, food photography, and abstract art) for two large language models, ChatGPT and Gemini.*

Image prediction

Count

Image prediction			LLM Used		Total
			ChatGPT	Gemini	
Correct	Image category	Wildlife	10	5	15
		Still life	10	6	16
		Portrait	10	0	10
		Landscape	7	5	12
		Food photography	7	4	11
		Abstract art	10	6	16
	Total		54	26	80
Wrong	Image category	Wildlife	0	5	5
		Still life	0	4	4
		Landscape	3	5	8
		Food photography	3	6	9
		Abstract art	0	4	4
	Total		6	24	30
Total	Image category	Wildlife	10	10	20
		Still life	10	10	20
		Portrait	10	0	10
		Landscape	10	10	20
		Food photography	10	10	20
		Abstract art	10	10	20
	Total		60	50	110

Finding 2: ChatGPT outperformed Gemini in accurately distinguishing between real and AI-generated images, particularly at higher confidence levels, while Gemini exhibited more frequent overconfidence and errors.

The crosstabulation data from Table 2 reflects the performance of ChatGPT-4 and Gemini in distinguishing between real and AI-generated images, with their confidence levels also factored into the analysis. The findings suggest notable differences in accuracy and confidence between the two models. When both models correctly identified images, the majority of responses were concentrated at confidence levels of 8 and 9, with ChatGPT showing greater accuracy at higher confidence levels. For instance, 28 out of 54 correct responses from ChatGPT occurred at a confidence level of 9, compared to Gemini's 13. Similarly, at a confidence level of 8, ChatGPT had 16 correct identifications, while Gemini had only 6. This indicates that when ChatGPT had high confidence, it was more likely to be accurate compared to Gemini, which had a lower rate of accuracy, especially in the

mid-confidence ranges. The highest confidence level (10) was relatively rare, but still reflected more accuracy for Gemini (7 correct) than ChatGPT (4 correct).

However, the wrong predictions offer another layer of insight. Gemini exhibited more incorrect predictions at higher confidence levels, particularly at a confidence level of 8, where 15 out of 24 wrong assessments were made. In contrast, ChatGPT had fewer incorrect responses overall, with only 6 wrong predictions, of which 2 occurred at a confidence level of 9. This suggests that Gemini was overconfident in several cases, especially when distinguishing AI-generated images. The data reveals that ChatGPT is generally more reliable and cautious in its assessments, with fewer errors at higher confidence levels. Gemini, on the other hand, showed a tendency to produce incorrect assessments despite higher confidence, indicating a potential overestimation of its capabilities. This disparity highlights ChatGPT's superior accuracy in distinguishing real from AI-generated images in this study.

Table 2 *Crosstabulation of large language models (LLMs) ChatGPT and Gemini's image prediction accuracy, categorized by confidence levels ranging from 7 to 10. The table is structured into two main sections: correct and incorrect predictions.*

How confident are you in your assessment of this image's origin (AI or real)?

Image prediction			LLM Used		Total
			ChatGPT	Gemini	
Correct	How confident are you in your assessment of this image's origin (AI or real)? Provide a confidence level from 1 to 10, and explain your reasoning.	7	6	0	6
		8	16	6	22
		9	28	13	41
		10	4	7	11
	Total		54	26	80
Wrong	How confident are you in your assessment of this image's origin (AI or real)? Provide a confidence level from 1 to 10, and explain your reasoning.	7	1	4	5
		8	2	15	17
		9	2	1	3
		10	1	4	5
	Total		6	24	30

Total	How confident are you in your assessment of this image's origin (AI or real)? Provide a confidence level from 1 to 10, and explain your reasoning.	7	7	4	11
		8	18	21	39
		9	30	14	44
		10	5	11	16
	Total		60	50	110

Finding 3:

The final data from Table 3 reveals significant insights into how accurately the LLMs evaluated the realism of images across different categories. A clear pattern emerges, showing that some image types were more easily identified as real or AI-generated, while others, particularly more abstract or interpretive categories, presented substantial challenges. In terms of realism, portrait and wildlife categories stood out, with models demonstrating a strong ability to distinguish real from AI-generated images. In the portrait category, ChatGPT-4 alone achieved a perfect accuracy, correctly identifying all images. This suggests that realism in portraits—where subtle details such as facial expressions, lighting, and human features are crucial—was easier for the model to assess. Similarly, the wildlife category showed strong performance, with only a few misclassifications, indicating that the models were generally effective at detecting natural elements like fur textures, animal shapes, and environments as realistic or AI-generated.

On the other hand, categories like food photography and landscape revealed substantial difficulties in judging realism. In food photography, there was a notably high error rate, with 9 out of 20 images misclassified. This suggests that AI-generated food images achieved a high degree of realism, making it challenging for the LLMs to detect subtle inconsistencies, such as the way light interacts with textures or the finer details of food presentation. Similarly, landscape images saw 8 misclassifications, indicating that the models struggled to differentiate realistic scenery from AI-generated environments, likely due to the convincing replication of natural lighting and geographic features by the AI.

In the abstract art category, realism was harder to evaluate, with more errors occurring in mid-range scores (6–7). This is likely due to the subjective nature of abstract art, where realism is not necessarily tied to clear visual cues, making it difficult for the models to distinguish between authentic and AI-created works. Overall, the LLMs excelled in categories with more concrete realism cues, such as human faces and wildlife, but struggled with more abstract or highly stylized images. This suggests that while AI can replicate many realistic features convincingly, particularly in categories where authenticity is less subjective, it still poses challenges in evaluating more complex or artistically interpretive visuals.

Table 3 *Crosstabulation of image categories and realism ratings, including correct and incorrect predictions*

Image realism

Count

On a scale from 1 to 10, how would you rate the realism of this image?			Image prediction		Total
			Correct	Wrong	
4	Image category	Abstract art	3		3
	Total		3		3
5	Image category	Abstract art	2		2
	Total		2		2
6	Image category	Wildlife	1		1
		Abstract art	2		2
	Total		3		3
7	Image category	Wildlife	2	0	2
		Still life	2	0	2
		Landscape	2	0	2
		Food photography	1	0	1
		Abstract art	1	4	5
	Total		8	4	12
8	Image category	Wildlife	2	4	6
		Still life	8	0	8
		Portrait	4	0	4
		Landscape	5	5	10
		Food photography	3	7	10
		Abstract art	5	0	5
	Total		27	16	43
9	Image category	Wildlife	10	1	11
		Still life	3	4	7
		Portrait	3	0	3
		Landscape	3	2	5
		Food photography	3	1	4
		Abstract art	3	0	3
	Total		25	8	33

10	Image category	Still life	3	0	3
		Portrait	3	0	3
		Landscape	2	1	3
		Food photography	4	1	5
	Total		12	2	14
Total	Image category	Wildlife	15	5	20
		Still life	16	4	20
		Portrait	10	0	10
		Landscape	12	8	20
		Food photography	11	9	20
		Abstract art	16	4	20
	Total		80	30	110

Methods

The present exploratory study examines the capability of large language models (LLMs) in discerning between real and AI-generated images. A key objective was to evaluate whether AI tools can assist users in determining the authenticity of visual content. To achieve this, a dual-phased method was designed, consisting of image selection, AI replication, and the subsequent analysis of images by two prominent LLMs, ChatGPT-4 and Gemini.

Image selection and categorization

The first step involved curating a diverse set of images from a publicly accessible source. Six distinct categories were established to ensure a comprehensive range of visual styles. These categories included landscape photography, still life, abstract art, food photography, portrait, and wildlife covering both concrete and more interpretive subjects. From the free stock photography website *Unsplash*, five representative images were selected for each category, resulting in a total of n=30 images. This variety was crucial for testing the robustness of the LLMs in detecting subtle differences between real and AI-generated imagery. Factors such as lighting, texture, and composition were considered to ensure that the selected images provided a suitable challenge for AI analysis.

Figure 1 *Side-by-side comparison between a real photograph (Unsplash) and a generated image (Midjourney v6.1)*

AI Image replication

Once the real images were selected, they were replicated using the AI tool *MidJourney*. This software was employed to generate artificial versions of the original images using the "describe" function, which creates text prompts based on the visual content of the original photograph. These prompts were then used to generate AI versions of the images, with the goal of producing images that closely mimic the original photographs in terms of composition and visual quality. The replication process aimed to create AI-generated images that could potentially deceive human viewers. MidJourney's advanced image synthesis capabilities were utilized to ensure that the artificial images maintained a high level of visual realism while still potentially incorporating subtle inconsistencies or artifacts characteristic of AI-generated content. The resulting dataset comprised two sets of images for each category: one consisting of real photographs, and one of their AI-generated counterparts (N=60)

Evaluation by Large Language Models

The next phase of the study involved assessing the images using two widely accessible LLMs, ChatGPT-4 and Gemini. These models were chosen due to their user accessibility. A standardized evaluation guide was developed to direct the models' analysis, consisting of a series of questions that targeted different aspects of the images, including authenticity, realism, and technical characteristics.

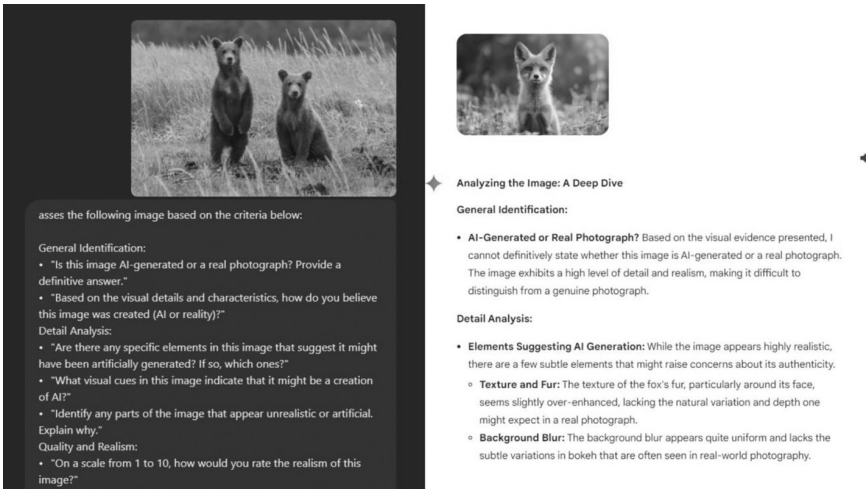

Figure 2 *Side-by-Side comparison of ChatGPT-4 and Gemini in image veracity assessment This table presents a comparative analysis of the performance of two large language models (ChatGPT-4 and Gemini) in assessing the veracity of images.*

Analytical approach

The responses from ChatGPT-4 and Gemini were analyzed in terms of accuracy, consistency, and reasoning. Special attention was given to the models' ability to detect specific indicators of artificial generation, such as unnatural lighting, inconsistent textures, or distortions in proportions. Furthermore, the confidence levels provided by the models were cross-referenced with their assessments to determine any correlation between certainty and accuracy. This structured methodology provided a comprehensive framework for assessing the capabilities of AI models in visual analysis. The approach not only tested the technical proficiency of LLMs in distinguishing between real and AI-generated content but also explored the extent to which these models can assist users in making informed judgments about the authenticity of visual media.

Bibliografie /Bibliography:

Bray, S., Johnson, S., & Kleinberg, B. (2022). Testing Human Ability To Detect Deepfake Images of Human Faces. *J. Cybersecur.*, 9. https://doi.org/10.48550/arXiv.2212.05056.

DeKraker, Z. (2022). Modeling Human Visual Detection Using Deep Networks. https://doi.org/10.37099/mtu.dc.etdr/1312.

DiResta, R., & Goldstein, J. A. (2024). How spammers and scammers leverage AI-generated images on Facebook for audience growth. *Harvard Kennedy School Misinformation Review*. https://doi.org/10.37016/mr-2020-151

Karinshak, E., & Jin, Y. (2023). AI-driven disinformation: a framework for organizational preparation and response. *Journal of Communication Management*. https://doi.org/10.1108/jcom-09-2022-0113.

Menz, B., Modi, N., Sorich, M., & Hopkins, A. (2023). Health Disinformation Use Case Highlighting the Urgent Need for Artificial Intelligence Vigilance: Weapons of Mass Disinformation.. *JAMA internal medicine*. https://doi.org/10.1001/jamainternmed.2023.5947.

Montoro-Montarroso, A., Cantón-Correa, J., Rosso, P., Chulvi, B., Panizo-Lledot, Á., Huertas-Tato, J., Calvo-Figueras, B., Rementeria, M., & Gómez-Romero, J. (2023). Fighting disinformation with artificial intelligence: fundamentals, advances and challenges. *El Profesional de la información*. https://doi.org/10.3145/epi.2023.may.22.

Sakamoto, T. (2020). Author Response: Factors in Color Fundus Photographs That Can Be Used by Humans to Determine Sex of Individuals. *Translational Vision Science & Technology*, 9. https://doi.org/10.1167/tvst.9.7.11.

Shoaib, M., Wang, Z., Ahvanooey, M., & Zhao, J. (2023). Deepfakes, Misinformation, and Disinformation in the Era of Frontier AI, Generative AI, and Large AI Models. *ArXiv*, abs/2311.17394. https://doi.org/10.48550/arXiv.2311.17394.

Spitale, G., Biller-Andorno, N., & Germani, F. (2023). AI model GPT-3 (dis) informs us better than humans. *Science Advances*, 9. https://doi.org/10.1126/sciadv.adh1850.

Xu, D., Fan, S., & Kankanhalli, M. (2023). Combating Misinformation in the Era of Generative AI Models. *Proceedings of the 31st ACM International Conference on Multimedia*. https://doi.org/10.1145/3581783.3612704.

Funding

No funding has been received to conduct this research.

Competing interests

The author(s) declare(s) no competing interests.

Ethics

Ethical approval was not required for this research as it did not involve human subjects.

Data availability: https://doi.org/10.6084/m9.figshare.27308799.v1

Gross, Eduard (2024). Evaluating photographic authenticity: How well do ChatGPT 4o and Gemini distinguish between real and AI-Generated images?. figshare. Media. https://doi.org/10.6084/m9.figshare.27308799.v1