HumanMade: A Platform for Verified Human Creations

Author: Luca Sarif-Kattan
Supervisor: Ligang He
Year of study: 2024
Keywords: AI, Generative AI, Human, Creations, Blockchain, Web App

Abstract: With the dawn of generative AI, the landscape of human creativity and ingenuity is increasingly encroached upon. It is clear to most experts that this trend will continue in some fashion [1], and there is a high chance that the entire creative space will be heavily dominated by AI in the near future. Due to the anthropocentric worldviews many of us hold [2] and to retain the pride of the human identity, having a platform which supports and incentivises human creation seems like it will be an in-demand product, and a beneficial necessity. HumanMade seeks to do this whilst avoiding the flawed approach of directly detecting AI-generated content, by:

- Giving humans an easy way to document and upload their Creations
- Ensuring Creations are tamper-proof and traceable to the original creator
- Enabling and empowering the community of users to decide what they think is human-made and deserves their attention

Acknowledgements: I would like to thank myself for my amazing ideas, skilled programming and genius-level problem solving. I would like to thank Ligang He for a frictionless and enjoyable supervisor experience.

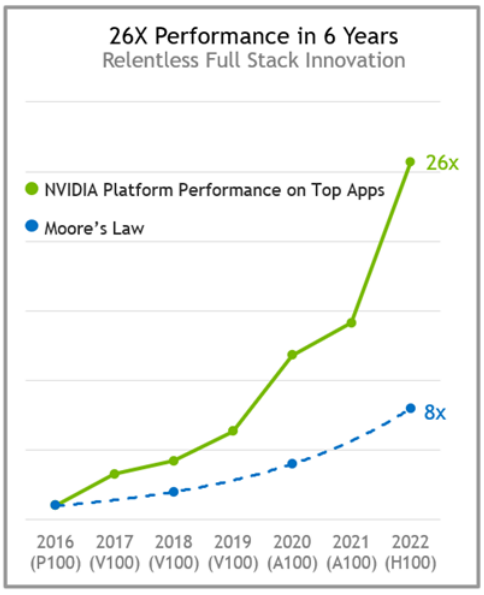
Contents

1 Motivators

1.1 The Generative AI Explosion

We are currently (March of 2024) in the midst of a generative AI (GAI) arms race. ChatGPT managed to reach 100 million users in 2 months [3]. Incredible advancements are being made weekly. More broadly, experts are predicting High-Level Machine Intelligence by 2059 [1], and this is an exponentially decaying timeline, down 8 years from the previous time the survey was given. Here are some of the main reasons for this current AI explosion:

- **Advancements in hardware.** The architectures which a lot of these models are based off of were discovered a while ago, (for example, the seminal Attention Is All You Need paper was released in 2017 [4]), but recently we have witnessed huge improvements in the AI accelerated hardware needed to scale these models. Examples include Google's development of the TPU [5], and Nvidia's development of the Ampere architecture line of GPUs. GAI models require large-scale compute for the training phase.



[6]

Figure 1: A comparison of Nvidia hardware progress compared to Moore's law. The A100 GPU was used to train OpenAI's GPT-4. [7]

- **Lack of regulations and restrictions.** There are currently no re-

strictions on the development of such technologies in the US, and the EU has introduced limited restrictions in the new AI act [8]. There are obvious geopolitical incentives to allow the acceleration of GAI technologies, and open letters calling to pause giant AI experiments have failed to bring action [9].

- **Profit incentives.** Such technologies are applicable across a huge spectrum of high-level tasks. There was 14 billion dollars in funding in 2023 towards GAI technologies.

1.2 A Look at the SOTA

A big part of the motivation for such a project comes from the unbelievable progress and ability of current GAI models, and how we might attempt to improve their detection. Therefore, in this section we take a brief look into the current state-of-the-art in GAI. At the moment there are two dominant model types in the GAI world:

- Transformer models, used for text generation
- Diffusion models, used for image and video generation

Transformers were initially designed for NLP tasks, but have proven to be incredibly versatile and scalable [10]. Performance has increased linearly with model size, even as model sizes have increased to trillions of parameters.

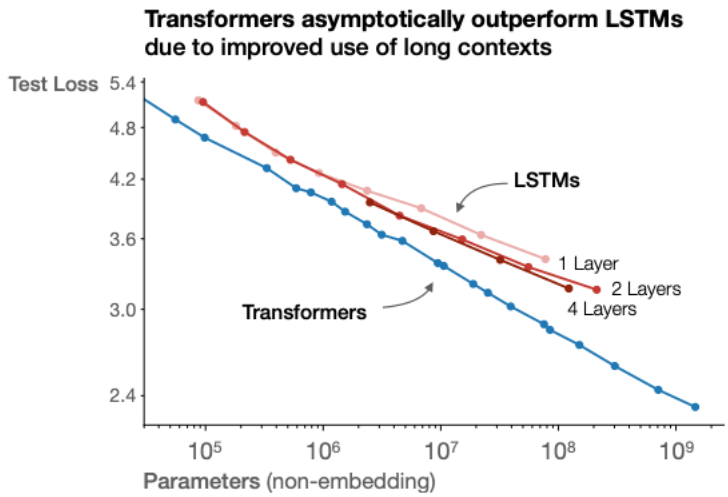


Figure 2: Test loss (performance) against size (parameter count) of a model

SOTA for Transformers is currently Claude Opus, recently released by Anthropic. A few stats include Opus achieving 86.8% on the MMLU (Undergraduate level multiple choice test) and 84.9% on the HumanEval code benchmark [11].

Diffusion models gradually transform noise into a coherent image, based on learned data distributions. The SOTA is inherently more subjective in this area, but Midjourney version 5 is widely regarded as a front-runner in image generation. Midjourney can generate extremely detailed images from highly specific prompts.

Figure 3: Generated using Midjourney with the prompt: 'A scholarly turtle wearing glasses and a graduation cap, sitting in front of a stack of books. He has a wise and contemplative expression, as if lost in philosophical thought'

SOTA in video generation is clearly Sora, a recently released model from OpenAI which took the industry by surprise. Essentially working by diffusion but on multiple frames, Sora can also take in highly specific prompts and output detailed and realistic videos.

Figure 4: A frame from a Sora video, generated with the prompt: 'A stylish woman walks down a Tokyo street filled with warm glowing neon and animated city signage. She wears a black leather jacket, a long red dress, and black boots, and carries a black purse. She wears sunglasses and red lipstick. She walks confidently and casually. The street is damp and reflective, creating a mirror effect of the colorful lights. Many pedestrians walk about.'

1.3 Preserving Human Creativity

The implication is that soon, we will be inundated with synthetic media and art. Whilst it is important to preserve human creations for philosophical reasons like maintaining the human identity, creative spirit and pride, there are two clearer motivators for building HumanMade:

1. There will be free-market demand for authentic, human made creativity
2. There is demand for GAI detection and tracking

Expanding on the first - recent research (perhaps unsurprisingly) shows a clear bias in humans towards human made art [2]. In this study, participants in different groups were presented with the same piece of art, but labelled as either human made or AI made. These labels directly effected a participants' preference to buy the art, via biased perceptions of the creativity and awe they felt towards it.

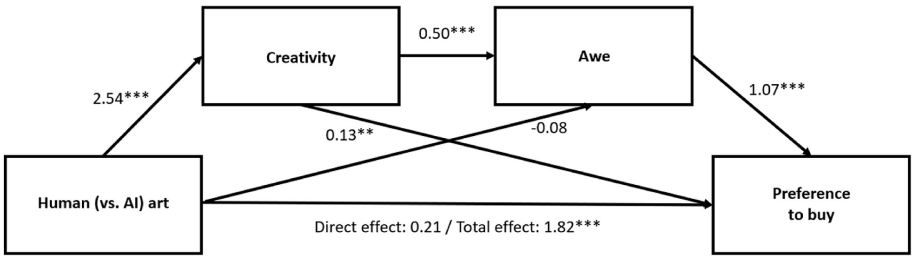


Figure 5: Indirect effect of human (vs. AI) art on preference to buy via perceived creativity and experienced awe.

On the second - tools which attempt to detect GAI are extremely popular (and for obvious reasons). For example GPTZero, a tool which attempts to detect AI generated text, has over 2.5 million users and partnerships with more than 100 organizations [12]. Industry-leading creative software developers Adobe have recently released Content Credentials, which is closer in spirit to what HumanMade is attempting. Both of these tools and more related systems will be further looked into subsequently.

2 Background

2.1 Related systems

We now look to analyse related systems and rate them out of 5 in three different categories - claimed capabilities, accuracy, and ease of use. This will give insight into what the best way to approach the development of HumanMade could be. In general, findings re-iterated the demand for a platform which does more than just attempt direct detection of AI. There were no holistic products with the aim of promoting human creativity.

System	Capabilities	Accuracy	Ease of Use
Adobe Content Credentials	3	2	5

Content credentials [13] is advertised as, first and foremost, a way for creators to attach credit and usage details to their work, and second as a way to be transparent with AI generation. At the former it succeeds, however with the latter, content credentials only indicates the use of generation with regard to Adobe Firefly and other proprietary apps. This does not solve the issue of verifying any piece of AI generated content as specifically human made. Accuracy is poor as the system can easily be cheated - someone could

screenshot a piece of digital work and then export it themselves. Ease of use is high as the feature is built right into industry leading apps used for content creation.

GPTZero	1	4	5

GPTZero is a web interface that detects whether your pasted in text is AI generated or not. From testing by pasting in GPT-4 (the current state of the art for chatbots according to the widely used Huggingface arena leaderboard [14]) generated text, GPTZero performs fairly accurately, predicting 6/7 samples overwhelmingly correctly. GPTZero is a paid service beyond 7 free scans, and does not support any other types of content. Other systems tested perform at or worse than GPTZero.

Sensity	4	4	5

Sensity is a leader in deepfake detection and has a high accuracy according to a recent study [15]. It has API, SDK and UI offerings.

2.2 Issues with Direct Detection

Despite accuracy scores for related systems initially looking positive, research shored up bigger picture concerns when considering a direct detection approach. GAI systems are improving at such a rapid pace that it will likely be impossible to keep up with them.

Sensity mentions on their site, 'As AI technology advances, new and more sophisticated techniques for generating realistic images emerge. Keeping up with these developments requires constant innovation and vigilance.' [16]. Sensity uses mainly ML techniques to identify GAI content [16], and so of course the detection model is only as up-to-date as the data used to train it. When a cutting edge GAI model like Sora drops, content remains undetectable until detection techniques catch up again (assuming it is even possible for them to do so).

Further, during the course of this third year project, Humbot was released [17]. Humbot 'humanizes' AI text to avoid detection by tools like GPTZero and Turnitin. Upon initial testing (7), Humbot works very well, and there are many examples and reviews of functionality on their site.

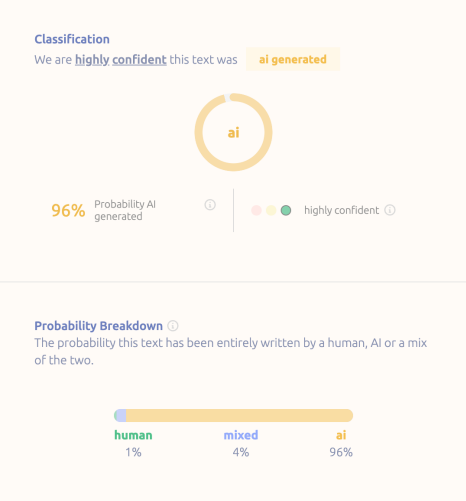


Figure 6: GPTZero classification before using Humbot

Figure 7: Classification after using Humbot

3 Objectives

After adequate research and analysis of the industry, the core objectives for HumanMade were then:

- Creation of an intuitive and modern web app, allowing users to view and support human-made creations
- An easy way for creators to document their project and creative process to the app
- A way to ensure uploaded content is tamper-proof and traceable to a the author

4 Design

4.1 Initial Design Ideas

Initially, to circumvent the extensive issues with direct detection detailed previously, but to still allow for seamless and automatic verification, HumanMade was to work on an 'evidence collection' basis. Evidence relating to the creative process itself would be collected, with the hope that this would give enough of a holistic input to whatever (likely ML) model was to be used to decide whether a creation was human or AI generated. Upon starting the project, there were some large caveats with this approach:

4.1.1 Issues

- Data collection for training would be long & an unproductive use of time. There were no existing datasets for my purpose.
- Data collected would be messy, with few consistent patterns. 'Evidence' could have included screen recordings, screenshots, timelapses, etc. There would not have been many consistent underlying features or patterns to learn to make a fully automated ML approach work well.
- Not actually avoiding the fundamental issues stemming from direct detection. At some point, GAI technologies would probably get so good that they would be able to generate the evidence being used to identify human made creations. This leads to a further philosophical point detailed later, but it was clear a more holistic approach would be required.

4.2 Final Design

The final design relies on some partially automated verification to give users an easier time identifying what might be human made, but the main approach revolves around the key concept of a progression timeline.

- Each progression timeline consists of 'commits', akin to a git branch.
- Each creator-defined commit consists of files, a description, a percentage of completion, and an icon to denote whether AI was used.
- There is a simple interface to make any commit tamper-proof and traceable.
- There is a command line interface for easy publishing to the progress timeline for creatives.

4.2.1 Advantages

This approach had some inherent advantages, addressing the systemic issues we would see when attempting direct detection.

- Generative AI has a very obvious 'diffusion' progression. An image of pure noise is slowly transformed into the final output. This would be easy to spot on a progress timeline.

- The approach allows the community to decide the level of AI involvement in a creation they are comfortable with, as this is clearly a gray area. Certain AI tools which simply help a human creator be more effective in creating may not be inherently negative.
- The approach sidesteps the 'arms race' between AI detectors and AI generators by allowing the community to decide for themselves what they think constitutes a human made creation. The best neural network detection algorithm to use to achieve the aims of human made is by necessity that of the human brain.

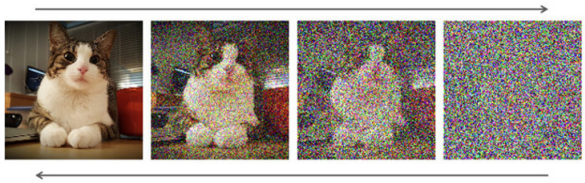


Figure 8: An example of the diffusion progression when generating a photo of a cat [18].

5 Development

5.1 Architecture & Tech Stack

5.2 Development Areas

5.2.1 Web App

5.2.2 Blockchain

5.2.3 Command Line Interface

5.3 Testing

6 Conclusions

- evulation, process of production, lessons learnt, further work, limitations and contributoins

7 Project Management

8 Philosophy of ito

References

[1] A. Impacts. (2022) 2022 expert survey on progress in ai. [Online]. Available: https://wiki.aiimpacts.org/doku.php?id=ai_timelines:predictions_of_human-level_ai_timelines:ai_timeline_surveys:2022_expert_survey_on_progress_in_ai

[2] K. Millet, F. Buehler, G. Du, and M. D. Kokkoris. (2023) Defending humankind: Anthropocentric bias in the appreciation of ai art. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0747563223000584

[3] D. Milmo. (2023) Chatgpt reaches 100 million users two months after launch. [Online]. Available: https://www.theguardian.com/technology/2023/feb/02/chatgpt-100-million-users-open-ai-fastest-growing-app

[4] (2017) Attention is all you need. [Online]. Available: https://arxiv.org/abs/1706.03762

[5] (2023) Google's cloud tpu v4 provides exaflops-scale ml with industry-leading efficiency. [Online]. Available: https://cloud.google.com/blog/topics/systems/tpu-v4-enables-performance-energy-and-co2e-efficiency-gains

[6] (2022) Fueling high-performance computing with full-stack innovation. [Online]. Available: https://developer.nvidia.com/blog/fueling-high-performance-computing-with-full-stack-innovation/

[7] (2023) Gpt-4 technical report. [Online]. Available: https://arxiv.org/abs/2303.08774

[8] (2023) Eu ai act: first regulation on artificial intelligence. [Online]. Available: https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence#:~:text=Unacceptable%20risk%20AI%20systems%20are,encourage%20dangerous%20behaviour%20in%20children

[9] (2023) Pause giant ai experiments: An open letter. [Online]. Available: https://futureoflife.org/open-letter/pause-giant-ai-experiments/

[10] (2020) Scaling laws for neural language models. [Online]. Available: https://arxiv.org/abs/2001.08361

[11] (2024) Introducing the next generation of claude. [Online]. Available: https://www.anthropic.com/news/claude-3-family

[12] M. than an AI detector Preserve what's human. (2023) Gptzero. [Online]. Available: https://gptzero.me/

[13] Adobe. (2023) Content credentials for assets generated with adobe firefly. [Online]. Available: https://helpx.adobe.com/firefly

[14] (2023) Lmsys chatbot arena leaderboard. [Online]. Available: https://huggingface.co/spaces/lmsys/chatbot-arena-leaderboard

[15] (2023) An investigation of the effectiveness of deepfake models and tools. [Online]. Available: https://www.mdpi.com/2224-2708/12/4/61

[16] (2023) How to detect ai generated images with sensity. [Online]. Available: https://sensity.ai/blog/deepfake-detection/how-to-detect-ai-generated-im/#:~:text=Extensive%20Training%20Data%3A%20Sensity%20AI,generation%2C%20ensuring%20high%20detection%20accuracy.

[17] (2024) Humanize ai text and get 100

[18] (2024) What is stable diffusion and how does it work? [Online]. Available: https://www.vegait.co.uk/media-center/knowledge-base/what-is-stable-diffusion-and-how-does-it-work