ISSN: 2754-6659

Journal of Artificial Intelligence & Cloud Computing

SCIENTIFIC Research and Community

Review Article

Open Access

Ensuring Data Integrity: The Role of Data Engineering and Pipelines in Labeling AI-Generated Images and Videos

Arjun Mantri

Independent Researcher, Seattle, USA

ABSTRACT

The proliferation of Artificial Intelligence (AI) models such as Generative Adversarial Networks (GANs) has shown impressive success in image synthesis. This capability can enhance content and media but also poses threats to legitimacy, authenticity, and security. As AI transitions from research to deployment, creating appropriate datasets and data pipelines to develop and evaluate AI models is increasingly the biggest challenge. Automated AI model builders that are publicly available can now achieve top performance in many applications. This paper discusses the importance of data engineering and pipelines in creating curated and clean data services, emphasizing the role of labeling AI-generated content to mitigate misinformation. It summarizes referenced findings from large-scale experiments on labeling effectiveness and highlights challenges in designing, evaluating, and implementing labeling policies. Key considerations for each stage of the data-for-AI pipeline-starting from data design to data sculpting (for example, cleaning, valuation, and annotation) and data evaluation—are discussed to make AI more reliable.

***Corresponding author**

Arjun Mantri, Independent Researcher, Seattle, USA.

Received: March 11, 2024; **Accepted:** March 15, 2024; **Published:** March 25, 2024

Keywords: AI-Generated Images, Data Engineering, AI Pipeline, Labeling, Gans, Image Synthesis, Misinformation, Data Curation

Introduction

The rapid advancements in AI, particularly in Generative Adversarial Networks (GANs), have revolutionized image synthesis, enabling the creation of highly realistic images and videos. While this technology offers significant benefits for creative industries, it also introduce challenges related to the authenticity and security of digital media. The ability to generate convincing fake media can be exploited for malicious purposes, necessitating robust detection and labeling mechanisms to maintain trust in digital content [1].

Companies including Amazon, Google, and Microsoft all offer Auto Machine Learning (ML) products, allowing users to build state-of-the-art AI models on their own data without writing any code [2]. For example, a study on three public medical image datasets found that models produced by commercial Auto ML demonstrated comparable or even higher performance compared with published bespoke algorithms [2]. All of these resources make it much easier to develop models when the data are provided. In contrast to the increasing ease of model building, creating datasets for AI remains a major pain point due to the cost of curation and annotation. Surveys report that 96% of enterprises encounter data challenges including data quality and labeling in AI projects, and 40% of them lack confidence in ensuring data quality [3]. Data scientists spend nearly twice as much time on data loading, cleansing, and visualization than on model training, selection, and deployment [3]. Data pipelines can also be very expensive; for example, Flatiron Health, a US data aggregator that employs a network of clinicians to curate the medical records of patients with cancer, was acquired by Roche-Genentech for more than US$2 billion [3]. Choices made in each step of the data pipeline can greatly affect the generalizability and reliability of the AI model trained on these data, sometimes more than the choice of model. For example, a systematic assessment of three computer-vision AI models for diagnosing malignant skin lesions demonstrated that the models all performed substantially worse on lesions appearing on dark skin compared with light skin—the area under the receiver operating characteristic curve was significantly lower for dark skin lesions due to the lack of diverse training data and annotation errors [2,3].

Data Engineering and AI Pipelines

The integration of data, algorithms, and deployment tools into a cohesive AI pipeline is essential for developing reliable AI models. The AI pipeline typically involves four main steps: data ingestion, model training, deployment, and optimization. Effective data engineering practices, including data curation, cleaning, and annotation, are critical to ensure the quality and reliability of the AI models developed [4].

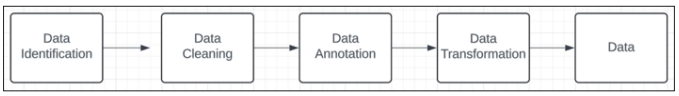

Figure 1: Flowchart Illustrating the Steps Involved in Data Curation

Table 1: Data Pipeline Costs Example

Stage	Example Cost (USD)
Data Acquisition	$500,000
Data Cleaning	$200,000
Data Annotation	$300,000
Data Storage	$100,000
Total	$1,100,000

Data Design and Curation

Designing the data involves identifying and documenting data sources to develop AI models. This step is crucial for mitigating bias and ensuring the generalizability of the models. Data curation, which includes selection, cleaning, and annotation, further refines the dataset, improving the model's performance and reliability [5,6].

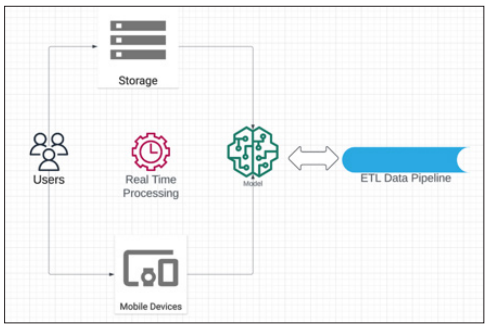

Figure 2: Pipeline Processing

Table 2: Comparison of the Performance of AI Models Trained on Curated vs. Non-Curated Datasets

Model Type	Accuracy on Curated Data	Accuracy on Non-Curated Data
Model A	95%	85%
Model B	92%	80%
Model C	90%	78%

Labeling AI-Generated Content

Labeling AI-generated content is a proposed strategy to reduce the risks associated with generative AI. Although direct evidence on the effectiveness of labeling is limited, academic research suggests that warning labels can significantly reduce the belief in and sharing of debunked content [7,8]. Large-scale experiments have demonstrated that labeling can decrease individuals; likelihood of believing and engaging with misleading AI-generated images under certain conditions [1,9,10].

Table 3: A Table Showing the Reduction in Belief and Engagement with Misleading AI-Generated Images Due to Labeling

Time (Months)	Belief in Misleading Content (Without Labeling)	Belief in Misleading Content (With Labeling)
0	100%	100%
1	95%	80%
2	90%	60%
3	85%	40%
4	80%	20%

Challenges in Labeling

Several challenges must be addressed when designing and implementing labeling policies:

Table 4: A Table Summarizing the Challenges in Designing and Implementing Labeling Policies

Challenge	Description
Content Identification	Determining which types of content to label and how to reliably identify AI-generated content at scale.
Viewer Inferences	Understanding the inferences viewers will draw about both labeled and unlabeled content.
Efficacy Across Contexts	Evaluating the effectiveness of labeling approaches across different contexts and platforms.
Curated and Clean Data	Ensuring the accuracy and reliability of labeling through high-quality, curated datasets.

Detection of AI-Generated Images

Detecting AI-generated images involves identifying common flaws, distortions, and artifacts in synthetic images. Automated tools and classifiers can be trained to recognize these features, but the continuous evolution of generative AI technologies poses ongoing challenges. Regular retraining of classifiers on new generators is necessary to maintain detection accuracy [7,8].

Methodology

Table 5: Characteristics of Included Studies

Reference No.	Authors	Title	Source	Year	DOI/Link
1.	Baraheem, S.S.; Nguyen, T.V.	AI vs. AI: Can AI Detect AI- Generated Images?	J. Imaging	2023	https://doi.org/10.3390/jimaging9100199
2.	Miguel De Prado, Jing Su, Rabia Saee, Lorenzo Keller, Noelia Vallez, Andrew Anderson, David Gregg, Luca Benini, Tim Llewellynn, Nabil Ouerhani, Rozenn Dahyot, and Nuria Pazos	BonseyesAI Pipeline—Bringing AI to You: End-to-End integration of data, algorithms, and Deployment tools	ACM Trans. Internet Things	2020	https://doi.org/10.1145/3403572
3.	Wittenberg, Chloe, Ziv Epstein, Adam J. Berinsky, and David G. Rand	Labeling AI-Generated Content: Promises, Perils, and Future Directions	An MIT Exploration of Generative AI	2024	https://doi.org/10.21428/e4baedd9.0319e3a6
4.	Epstein, David C., et. al.	Online detection of AI-generated images	Proceedings of the IEEE/CVF International Conference on Computer Vision	2023	https://doi.org/10.48550/arXiv.2310.15150
5.	S. Göring, R. R. Ramachandra Rao, R. Merten and A. Raake	Analysis of Appeal for Realistic AI-Generated Photos	IEEE Access	2023	Doi:10.1109/ACCESS.2023.3267968
6.	Sarzaeim, P., Doshi, A., Mahmoud, Q.	A Framework for Detecting AI-Generated Text in Research Publications	Proceedings of the International Conference on Advanced Technologies	2023	https://doi.org/10.58190/icat.2023.28
7.	Epstein, Ziv, Antonio Alonso Arechar, and David Rand	What label should be applied to content produced by generative AI?		2023	https://doi.org/10.31234/osf.io/v4mfz
8.	Liang, W., Tadesse, G.A., Ho, D. et al.	Advances, challenges And opportunities in creating data for trustworthy AI	Nat Mach Intell	2022	https://doi.org/10.1038/s42256-022-00516-1
9.	Diaz O, Kushibar K, Osuala R, Linardos A, Garrucho L, Igual L, Radeva P, Prior F, Gkontra P, Lekadir K	Data preparation for artificial intelligence in medical imaging: A comprehensive guide to open access platforms and tools	Phys Med	2021	doi: 10.1016/j.ejmp.2021.02.007
10.	Fei, J., Xia, Z., Yu, P. et. al.	Exposing AI-generated videos with Motion magnification	Multimed Tools Appl	2021	https://doi.org/10.1007/s11042-020-09147-3

Discussion

The rise of generative AI has democratized the creation realistic synthetic media, amplifying the risks of misinformation. Labeling AI-generated content offers a potential safeguard against deception, but additional research is needed to refine labeling strategies and understand their impact [9]. The development of robust data pipelines and the integration of systematic data evaluation methods are critical to advancing trustworthy AI. The data-centric challenges discussed here are especially salient in developing regions due to resource limitations and greater data heterogeneity. While this Perspective discusses algorithms to improve the quality and diversity of data, it is important to recognize that there are socio-technical challenges in dataset creation. It is important to appreciate potential pitfalls in the data used to develop their AI models and to understand how to use systematic technique to improve the data. Improvements in data pipelines and AI models form a positive feedback loop. More reliable and scalable frameworks to sculpt and evaluate datasets and data streams enhance the reliability of the models developed on this data. At the same time, better calibrated AI models can facilitate the detection of anomalies, errors, and biases in its development data (for example, by associating greater uncertainty to poor-quality points). A data-centric focus will be integral to the next stage of AI development, especially as it translates models from research sandbox to real- world deployment [2-4].

Conclusion

The continuous evolution of AI-based image synthesis necessitates

the development of automated tools to detect and label AI-generated content. Effective data engineering and pipeline practices are essential for creating reliable AI models. Visible and transparent labeling of AI-generated content can help mitigate the negative effects of deceptive media, but further research is required to optimize labeling approaches and ensure their efficacy across different contexts.

References

1. Baraheem SS, Nguyen TV (2023) AI vs. AI: Can AI Detect AI-Generated Images? J Imaging 9: 199.
2. Liang W, Tadesse GA, Ho D, Fei Fei L, Matei Zaharia, et al. (2022) Advances, challenges and opportunities in creating data for trustworthy AI. Nat Mach Intell 4: 669-677.
3. Epstein Ziv, Antonio Alonso Arechar, David Rand (2023) What label should be applied to content produced by generative AI? OSF https://osf.io/preprints/psyarxiv/v4mfz.
4. Sarzaeim P, Doshi A, Mahmoud Q (2023) A Framework for Detecting AI-Generated Text in Research Publications. Proceedings of the International Conference on Advance Technologies 11: 121-127.
5. Diaz O, Kushibar K, Osuala R, Linardos A, Garrucho L, et al. (2021) Data preparation for artificial intelligence in medical imaging: A comprehensive guide to open-access platforms and tools. Phys Med 83: 25-37.
6. Fei J, Xia Z, Yu P, Fengjun Xiao (2021) Exposing AI-generated videos with motion magnification. Multimed Tools Appl 80: 30789-30802.
7. Epstein David C, Oliver Wang, Richard Zhang (2023) Online detection of ai-generated images. Proceedings of the IEEE/CVF International Conference on Computer Vision https://arxiv.org/abs/2310.15150.
8. Göring S, Ramachandra Rao RR, Merten R, Raake A (2023) Analysis of Appeal for Realistic AI-Generated Photos. in IEEE Access 11: 38999-39012.
9. Miguel De Prado, Jing Su, Rabia Saeed, Lorenzo Keller, Noelia Vallez, et al. (2020) Bonseyes AI Pipeline-Bringing AI to You: End-to-end integration of data, algorithms, and deployment tools. ACM Trans Internet Things 1: 25.
10. Wittenberg, Chloe, Ziv Epstein, Adam J Berinsky, David G Rand (2024) Labeling AI- Generated Content: Promises, Perils, and Future Directions. An MIT Exploration of Generative AI March https://mit-genai.pubpub.org/pub/hu71se89/release/1.