Results from NIST's GenAI Text-to-Text (T2T) Discriminator Challenge

Trung Lam, MSc
Lusine Petrosyan, MSc
Dr. Michael Akinwumi

NFHA NATIONAL FAIR HOUSING ALLIANCE™

Results from NIST's GenAI Text-to-Text (T2T) Discriminator Challenge

Authors: Trung Lam[1], MSc, Lusine Petrosyan[2], MSc, and Dr. Michael Akinwumi[3]

Executive Summary

The rapid advancement of Generative AI (GenAI) technologies offers significant benefits and presents substantial risks, particularly in the housing sector. The increasing sophistication of Large Language Models (LLMs) makes it challenging to distinguish between AI-generated and human-generated text, potentially leading to discriminatory practices like racial steering, misleading property listings, and restricted access to housing resources. The National Fair Housing Alliance (NFHA) participated in the NIST GenAI Text-to-Text (T2T) Discriminator Challenge to develop building blocks that can be used to address these concerns. Our findings from the Challenge indicate that certain machine learning models, such as support vector machine (SVM) and extreme gradient boosting (XGBoost), show promise in differentiating between AI-generated and human-generated text. However, we encountered persistent difficulties in tracing the origin of AI-generated content, raising concerns about accountability and transparency. The implications of these findings extend to platform regulation, transparency measures, auditing and detection tools, literacy and awareness, and high-risk use cases. NFHA's ongoing participation in this initiative and related research aims to further explore these implications and contribute to the development of responsible GenAI practices in the housing sector.

Introduction

Generative AI (GenAI) technologies have advanced rapidly, offering significant creative and practical benefits across various industries. However, they also pose substantial risks, such as the potential spread of misinformation and disinformation, and the challenge to content provenance. The proliferation of deepfakes, fake news, and

[1] Associate Software Engineer, Responsible AI Lab, National Fair Housing Alliance. He conducted the research in this report.
[2] AI Policy Researcher, Responsible AI Lab, National Fair Housing Alliance. She worked on the policy implications of this study.
[3] Chief AI Officer and Rita Allen Fellow, Responsible AI Lab, National Fair Housing Alliance. He supervised the research. Inquiries about the study can be sent to him at makinwumi@nationalfairhousing.org.

generative AI tools used in critical sectors like medicine and housing has raised concerns about the accuracy and trustworthiness of digital content.

In the housing sector, the challenges posed by AI-generated content are particularly pronounced:

- **Racial Steering:** Large Language Models (LLMs) may exhibit biases that lead to discriminatory practices, such as racial steering, where AI-generated **property descriptions** may unintentionally guide potential **homebuyers** towards or away from certain areas based on biased algorithms. This can perpetuate segregation and inequity in housing markets. For instance, GPT-4 demonstrates racial steering by recommending neighborhoods based on race, steering White and Black home seekers in highly segregated cities like New York City and Chicago towards neighborhoods populated by people of the same race. Additionally, Black seekers are directed to lower socioeconomic areas and White seekers to higher opportunity areas.[4]

- **Misleading Property Listings and Fraudulent Offers:** AI-generated content can create convincing but deceptive property listings or fraudulent offers. This misinformation can mislead potential buyers or renters, distorting market dynamics and causing financial harm. These fraudulent practices can undermine trust in the housing market and affect the integrity of property transactions. In one instance, a University of Rhode Island student was tricked into wiring $1,800 for an apartment in Wakefield after responding to a listing on Facebook Marketplace.[5] Although this specific listing may not be AI-generated, the potential for AI to amplify rental and real estate scams by learning from human-generated scam data is a serious concern.[6]

- **Restricting Access and Skewing Public Service Distribution:** The deployment of LLMs in the housing sector can result in the restriction of information and essential resources based on user data or prompt engineering. This can create unjust and

[4] Gabriel, Saadia, Jessy Xinyi Han, Eric Liu, Isha Puri, Wonyoung So, Fotini Christia, Munzer Dahleh, et al. 2024. "Advancing Equality: Harnessing Generative AI to Combat Systemic Racism." An MIT Exploration of Generative AI, March. https://doi.org/10.21428/e4baedd9.7dc53bbf.
[5] Hamberg, A., Lima, M., and Munthe, S. (March, 2024). *False Promises, Real Losses: The Reality of Housing Scams in RI*. https://sites.brown.edu/tip/2024/03/26/false-promises-real-losses-the-reality-of-housing-scams-in-ri/
[6] FBI. (2022, July 12). *FBI warns of spike in rental and real estate scams*. FBI. https://www.fbi.gov/contact-us/field-offices/boston/news/press-releases/fbi-warns-of-spike-in-rental-and-real-estate-scams

unequitable outcomes. For instance, in Evanston, IL, Open Communities resolved federal Fair Housing Act litigation that revealed AI systematically rejected rental applicants receiving housing assistance payments, predominantly affecting Black renters. This AI-driven denial restricted access to housing for these individuals, illustrating how AI can amplify existing biases and lead to discriminatory practices. Such high-stake decisions highlight the necessity for mechanisms of liability and accountability to ensure fair and equitable distribution of goods and services.[7]

Challenges

LLMs are becoming increasingly sophisticated, making it difficult to distinguish between AI-generated and human-generated text. As these models improve, they produce text that is increasingly indistinguishable from human writing, complicating the detection process. As such, current provenance methods might have difficulty distinguishing between AI-generated text and human-generated text. In addition, reliance on older texts that predate LLMs for training the model poses challenges, as contemporary texts might be generated by GenAI, making it harder to differentiate between sources and sourcing data for training.

Enforcing the Fair Housing Act and the Equal Credit Opportunity Act (ECOA) in the context of AI use cases in housing is becoming increasingly challenging due to the advancements in LLMs. This is because LLMs are rapidly evolving, creating text so sophisticated that it is often indistinguishable from that written by humans. As these technologies continue to improve, detecting whether content is AI-generated or human-generated becomes increasingly difficult. This creates significant hurdles for regulators and watchdogs tasked with ensuring compliance with the Fair Housing Act (FHA) and Equal Credit Opportunity Act (ECOA), which mandate that housing-related communications and lending practices are free from discrimination and bias.

Given the current limitations of AI provenance methods, accurately identifying AI-generated text becomes problematic. This means that discriminatory or biased content generated by AI can slip through the cracks, undetected and unchallenged. Consequently, enforcing fair housing and lending laws becomes more complex as it is harder to pinpoint the source and intent behind potentially harmful communications.

[7] https://www.open-communities.org/post/press-release-open-communities-reaches-accord-in-case-addressing-artificial-intelligence-communicat

As the increasing indistinguishability of AI-generated text from human-written content makes it difficult to identify instances where discriminatory language or practices might be subtly embedded within AI-generated communications or decision-making processes, the rise of LLMs necessitates a proactive and adaptive approach to regulation and oversight, emphasizing transparency, explainability, and ongoing monitoring of AI systems to uphold the principles of fairness and equity in housing.

Solution

To address these issues, the National Institute of Science and Technology's (NIST) AI challenge focuses on differentiating between AI-generated and human-generated text. The National Fair Housing Alliance (NFHA) participated in this challenge to better prepare the civil rights group and its members for mitigating risks that come with the adoption of LLMs in housing and lending, and to provide research-based evidence to support housing policymakers' efforts to advance AI innovation in housing in a safe, secure and trustworthy manner.

To this end, we trained a model specifically designed to identify the provenance of text-based content, which is crucial for ensuring accurate information and holding content creators accountable. This model aims to improve the detection of AI-generated text and distinguish it from human-generated text.

The implications of this exercise for platform regulation in the context of AI-generated text are profound and far-reaching. While an AI-generated text often exhibits distinct patterns that can set it apart from human-generated content, identifying the specific source or platform—for example, whether it is Gemini, ChatGPT, Claude, MetaAI, or Perplexity—remains a significant challenge. Hence, we developed models that can detect the likelihood of the source of a given text using a labeled dataset. Each text in the dataset is labeled according to its source, whether generated by Gemini, ChatGPT, Claude, MetaAI, Perplexity, or a human. We then trained the dataset using various machine learning models to determine which model performs best in platform detection. We used both accuracy and Area Under the Receiver Operating Characteristic curve (AUC) to evaluate model performance. AUC measures a model's ability to separate or distinguish between classes (value closer to 1, the better the model can distinguish between AI-generated text and human-generated text), while accuracy represents the proportion of correct predictions out of all predictions (value closer to 1, the better the predictions).

The dataset used to train the model comprises a balanced collection of 5,000 entries, evenly split between AI-generated and human-generated texts. It includes 2,500 AI-generated samples, with 500 texts sourced from each of the following platforms: ChatGPT, Perplexity, Claude, Gemini, and MetaAI. The remaining 2,500 entries are human-generated texts, meticulously collected from Wikipedia pages to ensure a diverse and representative range of natural language usage. This comprehensive dataset facilitates a robust training process, aimed at enhancing the model's ability to differentiate between AI and human-generated content accurately.

Results

Training Results

The results in Table 1 show the performance of various machine learning models in distinguishing between AI-generated and human-generated text. The models are ranked by AUC, a metric that combines sensitivity and specificity to evaluate a model's overall performance.

The Support Vector Machine (SVM) model achieved the highest AUC of 0.969854, indicating that it is the most effective model overall at differentiating between the two types of text. Extreme gradient boosting (XGBoost) is close behind with an AUC of 0.969682. The other models performed well, with AUCs ranging from 0.966134 for Gradient Boosting Classifier to 0.845 for Decision Tree.

The results also show the accuracy of each model, which measures the proportion of correctly classified text samples. However, accuracy can be misleading when the classes are imbalanced, as it may not reflect the model's ability to correctly identify the minority class. Therefore, AUC is a more reliable metric for evaluating model performance in this case.

Overall, the results suggest that several machine learning models can effectively distinguish between AI-generated and human-generated text. However, the SVM and XGBoost models are the most promising and warrant further investigation.

Table 1: Results (Classification between AI vs human)

Machine Learning Models (Sorted by AUC)	Accuracy	AUC
Support Vector Machine	0.912	0.969854
XGBoost	0.9074	0.969682
Gradient Boosting Classifier	0.9052	0.966134
Random Forest	0.901	0.96357
Logistic Regression	0.9012	0.957048
Stochastic Gradient Descent Classifier	0.8914	0.949703
KNN (k-nearest neighbor) k=5	0.8838	0.942571
KNN (k-nearest neighbor) k=3	0.8806	0.928814
Gaussian Naive Bayes	0.8222	0.901979
KNN (k-nearest neighbor) k=1	0.8634	0.8634
Decision Tree	0.845	0.845

While AI-generated text often exhibits distinct patterns that set it apart from human-generated content, pinpointing the specific source or platform used to generate the text, whether it is Gemini, ChatGPT, Claude, MetaAI, or Perplexity for example, proves to be notably complex. This difficulty arises from the nuanced similarities across various AI platforms and the sophisticated nature of their text generation. As a result, attributing responsibility for specific pieces of text becomes problematic, complicating efforts to ensure accountability and traceability in the use of AI technologies. These complexities create a significant obstacle in attributing responsibility for specific pieces of text. Without the ability to accurately trace and identify the origin of AI-generated content, ensuring accountability becomes exceedingly difficult. This lack of traceability poses a major regulatory challenge, as it undermines efforts to enforce compliance and maintain oversight over the use of AI technologies.

Table 2 summarizes key findings from developing machine learning models to identifying the source of text, distinguishing between six potential origins.

- **Model Performance:** The models exhibited varying degrees of success in classifying text by source. The XGB Classifier achieved the highest AUC of 0.9498, suggesting strong overall performance in distinguishing between sources. However, accuracy varied considerably across models, with XGB Classifier at 74.6% and Gaussian Naive Bayes at 33.5%.
- **Challenges in Attribution:** Despite some models demonstrating relatively high AUC scores, accuracy results indicate the complexity of definitively attributing text to a specific AI platform or human origin. This difficulty underscores the nuanced similarities in text generation across platforms, making source identification a challenging task.
- **Top-Performing Models:** XGB Classifier, Gradient Boosting Classifier, and Support Vector Machine consistently ranked highest in terms of AUC, indicating their potential for further refinement and application in source attribution tasks.
- **Limitations:** Lower-performing models, such as Gaussian Naive Bayes and Decision Tree, may require further optimization or alternative approaches to improve their accuracy in source identification.

These results highlight the ongoing challenge of accurately identifying the source of AI-generated text, particularly among various platforms. While some models show promise, further research and development are needed to enhance the precision and reliability of source attribution. This is crucial for ensuring accountability, transparency, and ethical use of AI-generated content. Additionally, the findings emphasize the need for comprehensive guidelines and regulations surrounding the disclosure and labeling of AI-generated text to maintain transparency and trust in digital communications.

Table 2: Classifying Text by Source: ChatGPT, Gemini, MetaAI, Claude, Perplexity, or Human

Machine Learning Models (Sorted by AUC)	Accuracy	AUC
XGB Classifier	0.746	0.949807467
Gradient Boosting Classifier	0.741	0.9481536
Support Vector Machine	0.735666667	0.946622667
Random Forest	0.736	0.9447896
Logistic Regression	0.708333333	0.934908267
Stochastic Gradient Descent Classifier	0.656	0.908985733
KNN (k-nearest neighbor) k=5	0.685333333	0.901839267
KNN (k-nearest neighbor) k=3	0.668333333	0.8766696
Gaussian Naive Bayes	0.335333333	0.8425726
KNN (k-nearest neighbor) k=1	0.643	0.7858
Decision Tree	0.62	0.772

Evaluation Results

To validate the performance of our trained classifiers, we applied them to a hold-out test dataset provided by NIST. This dataset contains 104 text summaries on various topics, with each text potentially generated by either AI or a human. The results of this test are summarized in the table below:

Table 3: Results from applying trained models to test data provided by NIST

Machine Learning Models (sorted by AUC)	AUC
Gradient Boosting	0.9219
XGBoost	0.9172
Support Vector Machine	0.9016

The Gradient Boosting model, while maintaining high performance, shows a slight decrease in AUC when applied to the NIST test data, from 0.966134 to 0.9219. This suggests the model remains robust but might encounter slight variations in unseen data. Similarly, the XGBoost model exhibits a reduction in AUC from 0.969682 to 0.9172 when tested on the NIST dataset, indicating strong but slightly diminished performance on new data. The SVM model's AUC decreased from 0.969854 in the training data to 0.9016 on the NIST test set, reflecting a consistent pattern of slight performance drop when encountering new data.

The application of our trained classifiers to the NIST hold-out test dataset confirms the robustness of our models. The Gradient Boosting, XGBoost, and SVM models all maintained high AUC values, although there was a noticeable decrease compared to their performance on the training data. This reduction suggests that while the models are highly effective, there is some variability when applied to entirely new data sets.

Overall, these results underscore the importance of continuous evaluation and refinement of AI detection models to ensure they remain effective across diverse datasets. The consistent performance across different datasets reinforces our models' reliability in distinguishing between AI-generated and human-generated text.

Policy Implications

1. Platform Policy and Content Moderation

As referenced previously, detecting the source of AI-generated content is increasingly difficult, especially as training datasets become more comprised of synthetic content. The lack of transparency regarding model datasets and their usage exacerbates this challenge, making it harder to track and mitigate harms while establishing accountability. For example, without clear visibility into the origins of harmful text, it becomes impossible to address concerns with the relevant platform. The opacity prevents investigations into the guardrails and limitations that allow misuse and harm. In cases where harm is observed, identifying liable parties becomes a persistent issue, complicating efforts to hold actors responsible. As evidenced in industry practices, LLM platforms are beginning to set disclosure and liability disclaimers to inform users of their limited role in content production and ownership, reflecting the growing complexity of algorithmic accountability.

The provenance of data and content is also crucial in reassessing how current platform protections under Section 230[8] will be challenged when public forums and networks serve both as producers and circulators of AI-generated content. For instance, Meta's use of LLMs to generate user content and its role as a platform for sharing and facilitating engagement around information sharing exemplifies this issue. As AI-generated content proliferates, the dual role of platforms like Meta complicates the enforcement of information intermediaries, which were originally designed to shield platforms from liability concerns around user-generated content. This evolving challenge necessitates a reconsideration of legal complexities introduced by AI and the policy opportunities to balance innovation with platform responsibilities in a proactive manner to ensure public safety.

2. Transparency and Accountability

The findings provide practical insights to support the regulatory developments in transparency and accountability measures for AI-generated content and datasets. The transparency of digital content provenance will help set accountability expectations based on potential impacts. Considering a scenario where AI-generated content is used to create real estate listings, clarifying whether a property listing is AI-generated or created by a human would allow potential buyers to better assess the authenticity of

[8] Section 230 Overview: https://crsreports.congress.gov/product/pdf/R/R46751

information. For example, based on the training dataset, AI might produce biased representations of certain properties, leading to potential misconceptions. Transparency about the origin of these listings helps buyers avoid falling for misleading information and ensures that real estate platforms are held accountable for the accuracy of their content. This accountability is essential to protect consumers and maintain trust in the housing market. This process will also expand on the policy discussions of differentiating between AI authorship and ownership. For instance, in the scenario mentioned, it's crucial to determine whether the responsibility for the outcomes of AI-generated content falls on the listing owner, the platform hosting the listing, both parties, or neither. Insights from this challenge will help clarify these responsibilities and guide the development of appropriate regulations to address such issues effectively.

3. Auditing and Detection Tools

With the rise of AI audits, our findings help illustrate how to integrate content provenance checks into both in-house processes, such as detection of AI-generated content used for solution developments, and sensitive applications. This process further supports human awareness of synthetic content and applicable instances of contestability. For example, consider a real estate platform using AI to generate property descriptions. This challenge will help determine how to track and verify whether a listing's content is AI-generated, which is crucial for ensuring accuracy and transparency. It will also address the limitations of current detection software, such as biases in identifying AI-generated text and assumptions about the quality of its outputs. Insights from the challenge will guide improvements in both detection tools and regulatory practices, enhancing human oversight and accountability in sensitive contexts.

4. Literacy and Awareness

The presented findings demonstrate a compelling opportunity to advance our ability to distinguish effectively between AI-generated content, human-authored text, and static or fixed content. By improving the precision of these differentiations, we can address core assumptions about the nature of content creation in a more nuanced way. This involves not only recognizing the distinct characteristics of each type of content but also understanding the collective assumptions underpinning these distinctions, such as the perceived authenticity of human voices versus machine outputs. Moving forward, it is crucial to develop frameworks for categorizing acceptable and unacceptable use cases of these technologies. This includes defining ethical boundaries and practical guidelines for deploying detection tools in various contexts, from academic integrity to public

misinformation. By establishing clear parameters, we can ensure that the technology serves to enhance transparency and accountability without undermining the integrity of content creation processes or infringing on individual freedoms.

5. High-Risk Use Cases

Understanding the performance and limitations of existing detection methods for distinguishing between AI-generated and human-authored content is crucial for comprehensively evaluating the risks associated with synthetic data in high-stakes applications. For instance, consider the use of AI-generated content in legal proceedings, where the authenticity of evidence is paramount. Current detection tools might flag certain AI-generated texts as potentially synthetic, but they may also struggle with subtle manipulations or advanced generative techniques that evade detection. This highlights the broader implications of relying solely on automated systems, as false positives or negatives could impact judicial outcomes and undermine trust in legal processes. Consequently, it becomes evident that human oversight is essential to complement automated tools. Decision-makers must remain vigilant against automation bias – where overreliance on technology leads to overlooked errors or inaccuracies – and ensure that the significance of the task is not diminished by misplaced confidence in detection algorithms. By incorporating rigorous human review and maintaining a nuanced understanding of the technology's limitations, we can better manage the potential risks and ensure responsible application of GenAI across critical domains.

Conclusion

As observed, GenAI is increasingly utilized in housing and real estate contexts, raising concerns about potential biases and discrimination. To address transparency issues, NFHA participated in NIST's GenAI Text-to-Text (T2T) Discriminator Challenge, aiming to evaluate the effectiveness of current detection models in distinguishing between AI and human-generated content. This effort is also critical for enhancing AI literacy, and accountability in cases of liability. The findings indicate that several machine learning models, particularly SVM and XGBoost, are effective at differentiating between AI-generated and human-generated text, suggesting these models are promising candidates for further investigation. Additionally, using these models to trace content origins underscored the persistent challenges of data provenance in the era of synthetic media. While some models demonstrated potential, additional research is necessary to improve the precision and reliability of source attribution. These results offer important policy considerations for broader AI governance, particularly concerning the increasing

presence of AI-generated content in housing transactions and related opportunities. As NFHA's participation continues in the initiative, we are keen to share the ongoing results and outcome implications.

NationalFairHousing.org

1331 Pennsylvania Ave.
NW #650
Washington, DC 20004

202.898.1661

NFHA NATIONAL FAIR HOUSING ALLIANCE™