NATIONAL INSTITUTE ON ARTIFICIAL INTELLIGENCE IN SOCIETY
http://csus.edu/ai

Generative AI Results and Reality: An Assessment of Representation in Image Generation of Graduate Education

AJMAL M. AMINEE, JOSEPH TAYLOR, California State University Sacramento

ABSTRACT

The accuracy of Generative Artificial Intelligence (GAI) tools is the product of the quality of underlying data used to train models, and the models themselves. This interplay between data and models can lead to differences in the accuracy of outputs provided to common prompts across different GAI tools. This study investigates the disparities in accuracy related to representativeness between the outputs of GAI tools and demographic data from National Center for Education Statistics (NCES) and large enrollment, regional comprehensive university in the western United States (CSU). Three GAI platforms - ChatGPT, Co-pilot, and Gemini were evaluated using five samples each, with the same instruction across all platforms: “Show a class of graduate students.” The GAI outputs were analyzed based on three demographic variables: gender, race, and age group. These outputs were then compared to national averages from the NCES for gender and race and the CSU for age group. Notably, the variances in the results showed broader distributions across the demographic variables. To assess accuracy, a representation rate metric was calculated, reflecting the average absolute variance from the NCES and CSU benchmarks. The findings highlight the opportunity of higher quality data in model training, as well as the necessity for improved algorithms and methodologies in GAI systems to represent complex demographic information more accurately.

Keywords: Artificial Intelligence, generative AI, ChatGPT, Co-pilot, Gemini, Image Generation

1. INTRODUCTION

Generative Artificial intelligence (GAI) refers to artificial intelligence (AI) systems that can generate new data, often in the form of images, text, audio, video, or other types of contents (Chien, Chan & Hou, 2024). The GAI systems can create samples that resemble data from the training data sets. The sample data sets are often created by learning the patterns and structures present in the training or source data sets. These patterns and structures create new data streams in forms of image, text, audio, video, and other content types (Gugin, 2023).

As AI and its various subfields, including machine learning (ML), natural language processing (NLP), and computer vision (CV) evolves, the importance of representative sources or training data increases. The GAI reliability depends on data availability, quality of data source(s), generalization, evaluation metrics, ethical consideration, and privacy. The quality of the source data refers to the representativeness of the source data in relation to the predictive intentions of the GAI output (Reddy, 2024). The output is driven by the completeness, high-quality, and unbiased data source. As far as generalization of training data is considered it refers to the representativeness of the data in relation to the real-world distribution of data to facilitate the realistic, dependable, and dependable result. Robust evaluation metrics are important for accessing the contribute to the dependability and reliability of the GAI system. The metric determinedness of all case scenarios determines the robustness. The algorithm should traverse options possible and infer to every possible in the dataset.

Ethical considerations and privacy are other key factors that determine the GAI reliability. While algorithmic interventions are a valuable tool to reduce challenges associated with the quality of data on which models are trained, algorithmic override of training data Is not without risk. On February 22, 2024, Fox Business reported that Google Gemini produced inaccurate demographic images of historic figures such as George Washington and Abraham Lincoln (Fox Business, 2024). In response, Google blocked Gemini's image generation capability. Prabhakar Raghavan, Senior Vice President of Google, acknowledged that Gemini program developed to ensure inclusion and diversity that was over-ridden by algorithms (Raghavan, 2024). OpenAI \faced a similar challenge in July 2022 that led to updates of the algorithms to address demographic inaccuracies (Vynck & Tiku, 2024). The core issue may lie in the representation data rather than the algorithms themselves. With sufficient representation data, generative AI (GAI) can produce more accurate and reliable results.

The importance of ethical considerations cannot be overlooked as it may impact humans socially, economically, and personally. The ethical considerations include considering safety measures, being unbiased and fair, relying on data privacy, protecting personal privacy, practicing transparently and explainability and acting accountably and responsibly (Eiseman & Ortiz, 2023).

2. RESEARCH BACKGROUND

Generative AI (GAI) has been evolving rapidly as technology companies invest substantial resources into research and development. Many of these companies continuously improve their algorithms and data representations. One of the significant controversies in this field arose when Google Gemini created gender-inclusive versions of historic figures. For instance, Google Gemini produced images of George Washington and Abraham Lincoln as Black or African American individuals (Vynck & Tiku, 2024). Additionally, they generated a female image of the Pope, the head of the Catholic Church (Vynck & Tiku, 2024). Companies like Google, Microsoft, and OpenAI are pioneering the development of algorithms that aim to be ethical, gender-balanced, racially, and ethnically inclusive, responsible, usable, and dependable. However, we still lack a GAI platform that can consistently create reliable, dependable, and accurate outputs.

Despite the revolutionary potential of Artificial Intelligence (AI), its reliability, dependability, and accuracy are under scrutiny. For example, Valle-Cruz, García-Contreras, & Gil-Garcia (2024) discussed the negative impacts of AI on government, noting that the dark side of AI systems is influenced by political, legal, institutional, data, and technological factors. Bernard Marr (2023), a contributor to Forbes, identified 15 significant risks associated with AI, highlighting lack of transparency, bias and discrimination, privacy concerns, ethical dilemmas, and security risks as the top five.

AI has played a crucial role in providing recommendations to prospective customers, matching products, and services to their needs. However, recommendation algorithms, despite their benefits, can create information-related stress (Ma et al., 2021). Another study by Sun et al. (2021) discussed balancing the accuracy of recommendations with diversity. While humans retain the ability to make final decisions, using AI and algorithms to achieve desired outcomes, modifying algorithms can diminish the importance of factual accuracy.

Gap in Literature

There is a notable gap in the literature regarding the availability of sufficient training data to enable Generative AI to produce accurate, reliable, and dependable results. To address this gap, researchers instructed three widely used Generative AI platforms (OpenAI ChatGPT, Microsoft Co-pilot, and Google Gemini) to generate images of a graduate class. The outputs were evaluated based on three categorical variables: gender (male and female), race (with federal race categories: American Indian or Alaskan Native, Asian, Black or African American, Hispanic or Latino, Native Hawaiian or Other Pacific Islanders, and White), and age groups (30 years old and younger, 31-40 years old, and 41 years and older). The average results for race and gender from each platform were compared to national averages of graduate schools maintained by the National Center for Educational Statistics (NCES, 2023). The age group results were compared to those of California State University – Sacramento (CSU) graduate school. Finally, the researcher created a representation rate based on the average absolute value variance from the national averages for each category.

3. THEORITICAL FOUNDATION

To compare the difference between the results of Generative AI (GAI) with reality and to determine if there is sufficient training data available for GAI, several theoretical frameworks were evaluated. The theories considered were the Halo Effect, Stratification, Generative Adversarial Networks (GANs), and Entropy.

Information Theory Entropy was chosen as the foundational theory for this study. Information Theory Entropy's emphasis on measuring disorder and uncertainty is particularly suited to evaluating the accuracy and reliability of GAI-generated outputs compared to real-world data. Entropy measures the average amount of information produced by a stochastic source of data. It quantifies the uncertainty or randomness in a data set. In addition, mutual information and entropy measures help evaluate the quality of clustering and classification algorithms (Hershey, 2010). This theory discusses the measures of the average amount of information produced by a stochastic source of data. In addition, the theory quantifies the uncertainty or randomness in a data set. Based on this theory, we can predict the uncertainty in a training data set that is being sourced by the GAI defined algorithms (He & Yao, 2021).

Using Information Theory Entropy, the researchers will investigate the alignment of results from various GAI systems with reality. The study will involve three categorical variables: gender, race, and age group. Three different GAI platforms will be evaluated: OpenAI's ChatGPT, Microsoft's Co-pilot, and Google's Gemini. The validation process will involve calculating margin of error and representation rates.

- Margin of Error: The results from the GAI systems will be validated to see if they fall

within a 5% margin of error compared to national averages for gender and race, and California State University – Sacramento for age groups.

- Representation Rate: The representation rate for each GAI platform will be calculated by measuring the average absolute value variance. This will help determine how well each platform represents different demographic groups.

By focusing on these metrics, the researchers aim to provide a comprehensive assessment of the accuracy and reliability of GAI outputs, thereby offering insights into the sufficiency of training data and the potential biases within these AI systems.

4. PROBLEM STATEMENT

Systemic Errors in Training Data: Systemic errors can be introduced when training AI models on data that is not representative of the population for which the AI model will make predictions.

Amplification of Errors by Generative AI (GAI): Generative AI tends to amplify any systemic errors present in the underlying data on which the model was trained, potentially leading to biased or inaccurate outputs.

Artificial Adjustment of Outputs: GAI developers may create algorithms to artificially adjust outputs to compensate for low-quality training data. While well-intentioned, these adjustments can introduce new complexities and issues.

Unintended Consequences of Algorithmic Adjustments: Algorithmic methods designed to compensate for low-quality training data may have unintended consequences, including the introduction of new biases or inaccuracies.

Improving Data Representativeness: Enhancing the representation of the data used to train AI models is crucial. Better representativeness can lead to the development of more accurate and reliable AI models, reducing biases and improving overall model performance.

This study aims to address these issues by investigating the extent to which GAI systems reflect reality and whether the training data is sufficiently representative. By focusing on the systemic errors, the potential amplification of these errors by GAI, and the consequences of artificial adjustments, this research seeks to highlight the importance of high-quality, representative training data for the development of reliable and fair AI models.

5. PURPOSE STATEMENT

The purpose of this research study is to investigate the impact of systemic errors inherent in the underlying data used to train Generative Artificial Intelligence (GAI) models. Specifically, this study seeks to address the following questions:

1. Does GAI exacerbate systemic errors present in its training data?

2. Is there sufficient data available to enable GAI to generate meaningful and reliable outputs?

To achieve these objectives, the researchers will assess the reliability and dependability of three currently available GAI tools: OpenAI ChatGPT, Microsoft Co-Pilot, and Google Gemini. The study will maintain methodological consistency by posing the same set of instructions to all selected GAI tools.

By analyzing the performance and outputs of these GAI tools across standardized queries, this research aims to provide insights into the extent to which GAI models reflect and potentially amplify underlying data biases and limitations. Additionally, the study will evaluate the adequacy of existing data resources for training GAI systems and generating reliable, contextually appropriate responses.

This investigation contributes to the understanding of GAI systems' capabilities and limitations, with implications for improving the quality and ethical considerations of artificial intelligence technologies.

6. RESEARCH QUESTION AND HYPOTHESIS

The following research questions and hypotheses guided this study:

- RQ 1: To what extent was there a difference in gender of graduate students between the GAI generated result and the national average.
- H01: There is not a meaningful difference in gender of graduate students between the GAI generated results and the national average.
- H1a: There is a meaningful difference in gender of graduate students between the GAI generated results and the national average.
- RQ 2: To what extent was there a difference in race of graduate students between the GAI generated result and the national average.
- H02: There is not a meaningful difference in the race of graduate students between the GAI generated results and the national average.
- H2a: There is a meaningful difference in race of graduate students between the GAI generated results and the national average.
- RQ 3: To what extent was there a difference in age group of graduate students between the GAI generated result and the California State University – Sacramento.
- H03: There is not a meaningful difference in age group of graduate students between the GAI generated results and the California State University – Sacramento.
- H3a: There is a meaningful difference in age group of graduate students between the GAI generated results and the California State University – Sacramento.
- RQ 4: To what extent was there a difference in representation rate of graduate students between the GAI generated result and the national average and California State University – Sacramento.
- H04: There is not a meaningful difference in representation rate of graduate students between the GAI generated results and the national average and the California State University – Sacramento.
- H4a: There is a meaningful difference in representation rate of graduate students between the GAI generated results and the national average and the California State University – Sacramento.

7. RESEARCH METHODOLOGY

This study employed a quantitative methodology using descriptive and advanced statistical models to compare the outputs of Generative Artificial Intelligence (GAI) platforms with real-world data from the National Center for Education Statistics (NCES). The goal was to determine differences in representations of gender, race, and age group.

The researchers conducted a study to evaluate the answers and reliability of outputs from different Generative AI (GAI) platforms. One prompt and instructions guided the three widely used GAI platforms: OpenAI ChatGPT, Microsoft Co-pilot, and Google Gemini. The prompt requested the GAI to: **“generate an image of a class of graduate students”**.

To ensure data accuracy, consistency, and reliability, the researchers implemented the following steps:

1. Selection of Testers: To conduct the study, researchers have assigned five subjects with five different profiles to instruct the GAI to generate an image of graduate students. This multiple-user approach was designed to mitigate individual biases and ensure a more comprehensive evaluation.
2. Standardized Environment: All testers used the same version of each GAI platform and consistently accessed the platforms through Google Chrome installed on machines running Windows 11 operating system. This uniform setup was critical to eliminate variables that could affect the GAI outputs.
3. Controlled Timing: The queries were run within a one-hour period, specifically from 2:00 pm to 3:00 pm, on Friday, May 10, 2024. This controlled timing helped to reduce the impact of any potential temporal variations in the GAI platforms' performance.

4. Consistent Instructions: The researchers provided detailed instructions to the testers to ensure that each query was used identically across all platforms. This consistency in instructions was essential for a fair comparison of the outputs generated by each GAI tool.

By standardizing the environment, timing, and instructions, the researchers aimed to obtain reliable and comparable results from the three GAI platforms. This methodology enabled a rigorous assessment of the platforms' ability to generate accurate and representative images of a class of graduate students, thus providing insights into the platforms' performance and the potential amplification of biases in their training data.

8. DATA COLLECTION

The researchers in this study utilized 5 different users with different profiles in a standardized environment to ensure accuracy, reliability, and dependency of data.

GAI Platforms and User Profiles:

- The study focused on three GAI platforms: OpenAI's ChatGPT, Microsoft's Co-pilot, and Google's Gemini.
- Five different user profiles were created, each using a separate computer to query the GAI platforms.
- This approach ensured a diverse and comprehensive data collection process.
- Each user asked the same question across all three GAI platforms to maintain consistency.

Standardization of Environment and Timing:

- All queries run on Google Chrome web browser running on Windows 11 machine.
- To control potential temporal variations, all queries were performed within a one-hour period from 2:00 pm to 3:00 pm on Friday, May 10, 2024.

9. DATA ANALYSIS

1. Descriptive Statistics:

- Descriptive statistics calculations produced the mean values across the three GAI platforms for gender, race, and age group.
- These statistics provided a baseline understanding of how each platform represented the different demographic categories.

2. Chi-square Test for Homogeneity:

- To validate the results, a Chi-square Test for Homogeneity was applied. This test compared the distributions of gender (male and female), race (American Indian or Alaskan Native, Asian, Black or African American, Hispanic or Latino, Native Hawaiian or other Pacific Islanders, and White), and age groups (30 years old and younger, 31 to 40 years old, and 41 years and older) among the outputs from the three GAI platforms.
- The Chi-square test helped determine if there were significant differences in the demographic representations generated by each platform compared to the expected distributions from the NCES data.

10. RESULTS

Three main Generative Artificial Intelligence (GAI) platforms Open AI ChatGPT 4.0, Microsoft Co-pilot, and Google Gemini were able to generate the picture. ChatGPT image generated a picture with the following gender, race, and age group characteristics. To validate the gender results, the researchers used Chi-square Test for Homogeneity. The result set was identical across all 5 test users within the 5% margin of error (p= .05). The data

obtained from three different GAI platforms reflected different results based on the three categories of analysis: gender, race, and age group. The summary of data gathered is in the following tables.

Table 1. Results from 3 GAI platforms merged – Gender.

Gender	ChatGPT	Co-pilot	Gemini	Average
Male	50%	30%	25%	35%
Female	50%	70%	75%	65%
Total	100%	100%	100%	100%

Table 2. Results from 3 GAI platforms merged – Race.

Race based Federal classifications	ChatGPT	Co-pilot	Gemini	Average
American Indian or Alaskan Native	0%	0%	0%	0
Asian	20%	85%	10%	38%
Black or African American	50%	5%	0%	18%
Hispanic or Latino	10%	0%	0%	3%
Native Hawaiian or other Pacific Islanders	0%	0%	0%	0%
White	20%	10%	90%	40%
Total	100%	100%	100%	100%

Table 3. Results from 3 GAI platforms merged – Age Group.

Age Group	ChatGPT	Co-pilot	Gemini	Average
30 years old and younger	40%	90%	100%	70%
31 to 40 years old	20%	10%	0%	10%
41 years and older	60%	0%	0%	20%
Total	100%	100%	100%	100%

11.SUMMARY OF FINDINGS

The range of results for each demographic category—gender, race, and age groups—varied significantly across the three GAI platforms examined in this study. For instance, ChatGPT's estimates of the male gender representation in graduate programs were 200 percent higher than those of Google Gemini, and 2% higher than the 2021 census data from the National Center for Education Statistics (NCES; 2023). Despite these substantial variances among the platforms, the aggregated average of the results from ChatGPT, Microsoft Co-pilot, and Google Gemini closely approximated the NCES figures.

Table 4. GAI Results comparison with NCES – Gender.

Gender	GAI Average	National Average 2021	Variance
Male	35%	33%	+2%
Female	65%	67%	-2%

The analysis of race demographics across different Generative AI (GAI) platforms revealed substantial variances, both among the platforms themselves and in comparison, with the NCES. Notably, the representation of the Asian demographic varied significantly: Microsoft Co-pilot estimated that Asians comprised 85% of the graduate student population, while ChatGPT and Google Gemini reported much lower proportions of 20% and

10%, respectively. In contrast, the NCES data indicated that Asians constituted only 8% of the graduate student body.

For the Hispanic or Latino demographic, none of the GAI platforms reported any representation, which starkly contrasts with their actual 11% representation in the NCES census. Similarly, the representation of Black or African American students also showed a wide variance: ChatGPT estimated them as 50% of the graduate population, Microsoft Co-pilot at 5%, and Google Gemini reported none. While the actual figure in the NCES was only 12%.

Furthermore, the estimated average representation of White students across the GAI platforms was 12% less than their actual representation in the NCES census. These discrepancies underscore significant challenges in the accuracy and reliability of demographic estimations by current GAI platforms, suggesting a need for improved calibration and methodologies to better align GAI outputs with real-world demographics.

Table 5. GAI Results comparison with NCES – Gender.

Race based Federal classifications	GAI Average	National Average 2021	Variance
American Indian or Alaskan Native	0%	0%	0%
Asian	38%	8%	+30%
Black or African American	18%	12%	+6%
Hispanic or Latino	3%	11%	-8%
Native Hawaiian or other Pacific Islanders	0%	0%	-0%
White	40%	52%	-12%

In the analysis of age group demographics, a significant variance was noted among the GAI platforms studied. Specifically, ChatGPT estimated that 60% of graduate students were 41 years old or older, a stark contrast to the estimates from Microsoft Co-pilot and Google Gemini, both of which represented this age group as nonexistent. The age group values from the National Center for Education Statistics were not available. The CSU-GS was used only in this section of the study.

This paradoxical finding—where individual variances are high, yet the collective average aligns well with actual demographic data—suggests that while individual GAI platforms may have biases or errors in age group estimation, their combined outputs might inadvertently compensate for these inaccuracies, leading to a collective estimate that mirrors reality more closely. This highlights the potential utility of using an ensemble approach in generative AI applications to enhance the reliability of demographic estimations, despite individual platform inconsistencies

Table 6. GAI Results comparison with CSU – Sacramento – Age Group.

Age Group	GAI	CSU – Sac Fall 2023 Census	Variance
30 years old and younger	70%	66%	+3%
31 to 40 years old	10%	15%	-5%
41 years and older	20%	19%	+1%

Picture 1: OpenAI ChatGPT Output

Browsed by: Dell Latitude 5530, Windows 11 OS - Built: 22631.3447, Google Chrome Version 124.0.6367.91 at 21:59 hours on 4/29/2024.

Picture 2: Microsoft Co-pilot Output

Browsed by: Dell Latitude 5530, Windows 11 OS - Built: 22631.3447, Google Chrome Version 124.0.6367.91 at 21:59 hours on 4/29/2024.

Picture 3: Google Gemini Output

Browsed by: Dell Latitude 5530, Windows 11 OS - Built: 22631.3447, Google Chrome Version 124.0.6367.91 at 21:59 hours on 4/29/2024.

12.ANALYSIS

12.1. Algorithmic Differences

The Generative AI (GAI) platforms are the using training data and algorithms to generate results based on the input the end-user inputs. The difference in the training data and algorithms causes different results.

OpenAI ChatGPT

ChatGPT is built on Generative Pre-trained Transformers (GPT) architecture that excels in natural language understanding and natural language generation. The GPT architecture is built on multiple layers of attention mechanism to process user input and generate relevant responses or output.

Microsoft Co-pilot

Co-pilot is built on Codes model of OpenAI. Like ChatGPT, it utilized the GPT architecture. Co-pilot integrates the capabilities of GPT into Microsoft Productivity tools. In addition, the Co-pilot used the Microsoft data eco-system to provide context aware suggestions and automations.

Google Gemini

Gemini is built on transformer-based models like GPT and includes enhancements specific to Google's Research and development. Gemini is part of Google's Artificial Intelligence (AI) suite that integrates various AI models developed by DeepMind. In addition, Gemini combines features from large language models with advancements in multimodal AI.

There are algorithmic differences between GAI platforms. For example, ChatGPT and Co-pilot both use the GPT model, while Gemini uses transformer-based models like GPT. Although all these platforms utilize GPT-like architectures, one might expect identical results from ChatGPT and Co-pilot due to their shared GPT foundation. However, this is not the case. Each platform employs distinct algorithmic designs that lead to different outcomes.

12.2. OpenAI ChatGPT

The Generative AI (GAI) results presented are not meaningful close to the National Center for Education Statistics (NCES) averages. Specifically, the ChatGPT results showed a gender distribution of 50% male and 50% female students. However, the National Average Graduate College (NAGC) student population from NCES comprises 33% male and 67% female students. The results are not meaningfully close to the 5% margin of error.

A similar pattern of discrepancy was observed in the ChatGPT-generated racial distribution. The national average graduate student population comprises 8% Asian, whereas ChatGPT generated 20% Asian students. Further, NAGC reports a Black or African American population of 12%, but ChatGPT generated 50%. Conversely, while NAGC reports that 52% of graduate students are White, ChatGPT results in only 20%. These variances underscore significant deviations in the racial composition provided by ChatGPT compared to national averages.

Due to the lack of national data on age distribution, the researchers compared ChatGPT's results to the California State University - Sacramento Graduate (CSU-G) student population. The ChatGPT results indicated that 60% of the graduate student population is aged 41 years and older, 20% are between 31 and 40 years old, and 20% are below 30. In contrast, the CSU-G population consists of 19% of students aged 41 years and older, 10% between 31 and 40 years old, and 66% below 30. This discrepancy highlights similar variances in age demographics between ChatGPT and the CSU-G.

To quantify the disparities, we calculated the variance across three comparative variables (gender, race, and age) and generated an average score for the ChatGPT-generated values against the NCES and CSU-G populations referred to as representation rate (RR). RR score provides a comprehensive view of how ChatGPT's demographic outputs deviated from actual educational statistics.

Table 7. ChatGPT Representation Rate

Variable measured	Average Variance
Gender	15.00
Race	20.75
Age Group	24.00
Representation Rate - ChatGPT	19.92

12.3. Microsoft Co-pilot

The results from Microsoft Co-pilot generated for gender are meaningfully close, but for race and age groups, the values are not meaningfully close to the National Center for Education Statistics (NCES) averages. Specifically, the Co-pilot results showed a gender distribution of 30% male and 70% female students. The National Average Graduate College (NAGC) student population from NCES comprises 33% male and 67% female students. It is meaningfully close to the 5% margin of error.

A different pattern of discrepancy was observed in the Co-pilot generated racial distribution. The national average graduate student population comprises 8% Asian, whereas Co-pilot generated 80% Asian students. Further, NAGC reports a Black or African American population of 12%, but Co-pilot generated 5%. Conversely, while NAGC reports that

52% of graduate students are White results in only 10%. The values are not meaningfully close within 5% margin of error.

The Co-pilot results indicated that 0% of the graduate student population is aged 41 years and older, 10% are between 31 and 40 years old, and 90% are below 30. In contrast, the CSU population consists of 19% of students aged 41 years and older, 10% between 31 and 40 years old, and 66% below 30. It is not meaningful close within 5% margin of error.

Table 8. ChatGPT Representation Rate.

Variable measured	Average Variance
Gender	6.00
Race	35.00
Age Group	15.67
Representation Rate - Co-pilot	18.89

12.4. Google Gemini

The results from Google Gemini generated for gender, race, and age groups are meaningfully close to the National Center for Education Statistics (NCES) averages. Specifically, the Gemini results showed a gender distribution of 25% male and 55% female students. The National Average Graduate College (NAGC) student population from NCES comprises 33% male and 67% female students. It is not meaningfully close to the 5% margin of error.

A similar pattern of discrepancy with higher variance was observed in the Gemini generated racial distributions. The national average graduate student population comprises 8% Asian, whereas Gemini generated 10% Asian students. Further, NAGC reports a Black or African American population of 12%, but Gemini generated 0%. Conversely, while NAGC reports that 52% of graduate students are White results in 90%. The values are not meaningfully close within 5% margin of error.

The Gemini results indicated that 100% of the graduate students are below the age of 30. In contrast, the CSU-G population consists of 19% of students aged 41 years and older, 10% between 31 and 40 years old, and 66% below 30. It is not meaningful close within 5% margin of error.

Table 9. ChatGPT Representation Rate.

Variable measured	Average Variance
Gender	16.00
Race	15.75
Age Group	22.67
Representation Rate - Gemini	18.14

12.5. Analysis Summary

The results generated from different Generative AI (GAI) produced different values of variance for different variables in this study. The average variance was recorded 15 for ChatGPT, 6 for Co-pilot, and 16 for Gemini. The values are higher than the 5% expected margin of error for all three GAI platforms evaluated. Combined they produce an average of 12.34 that is closer to 5% margin of error, but it is not in the range for meaningful result (Surowiecki, 2004). Therefore, we reject the null hypothesis number 1: There is a meaningful difference in gender of graduate students between the GAI generated results and the national average.

Similarly, the race category was more variant among all 3 GAI platforms. For example, the average variance score for ChatGPT was 20.75, for Co-pilot was 35.00 and for Gemini was 15.75. In each case, higher than the 5% margin of expected error. The age group variable had higher variance rates, also. The variance rate for ChatGPT was 24.00, for Co-pilot it was 15.67 and for Gemini, it was 22.67, higher than the expected 5% margin of error. Therefore, we reject the null hypothesis number 2: There is a meaningful difference in race of graduate students between the GAI generated results and the national average.

The age group category was also variant among all 3 GAI platforms and for each category of variable. The average variant for ChatGPT was the highest, 24. The results followed by Gemini, 22.67 and finally Co-pilot, 15.75. The average variant score for the age group was 20.78. This value is a lot higher than the expected 5% margin of error. Therefore, we reject the null hypothesis number 3: There is a meaningful difference in age groups of graduate students between the GAI generated results and the national average.

The researchers created rankings by calculating the average variance for all three variables (gender, race, and age group) for all three platforms (ChatGPT, Co-pilot, and Gemini), referred to as Representation Rate (RR). The RR for ChatGPT was 19.92, for Co-pilot it was 18.89, and for Gemini, it was 18.14. The average score was 18.98. Although each variable (gender, race, and age group) had higher ranges and variance from the national averages, their RR rates are close. Eventually, the individual GAI platform and the average value of RR was a lot higher than the expected 5% margin of error. Therefore, we reject the null hypothesis number 4: There is a meaningful difference in representation rate of graduate students between the GAI generated results and the national average.

Table 10. Summary of GAI Average Variance and Representation Rate.

Variable/Platform	ChatGPT Average Variance	Co-pilot Average Variance	Gemini Average Variance	Average
Gender	15.00	6.00	16.00	12.34%
Race	20.75	35.00	15.75	23.83%
Age Group	24.00	15.67	22.67	20.78%
Representation Rate	19.92	18.89	18.14	18.98

Figure 1. Generative AI – Average Variances by Variable.

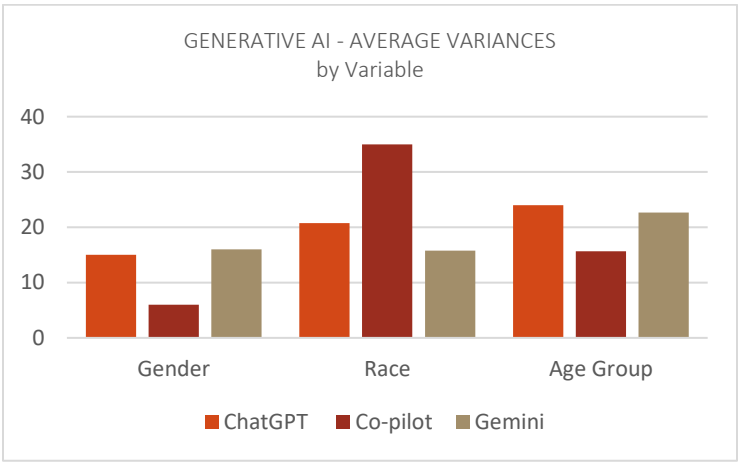
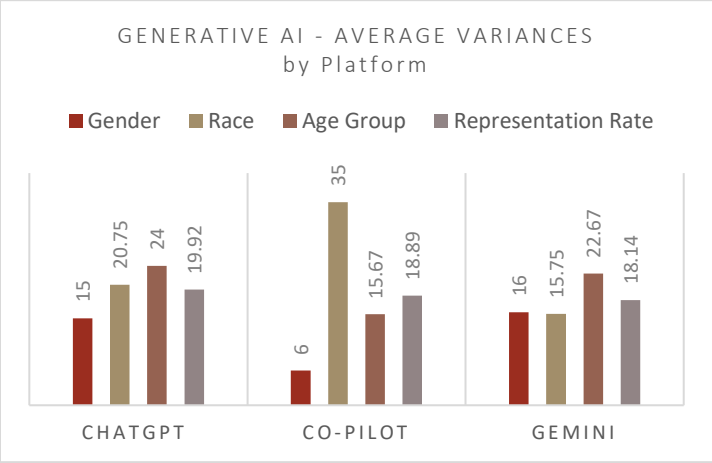


Figure 2. Generative AI – Average Variances by Platform.



13. CONCLUSION

This study assessed the variance in data generated by different Generative AI (GAI) platforms, specifically focusing on OpenAI's ChatGPT, Microsoft Co-pilot, and Google Gemini. The analysis revealed a significant variance in the outputs from these platforms, which escalates with the increasing complexity and number of categories within a variable. This variance becomes notably pronounced, reaching levels that may be considered questionable, particularly when comparing data across the three platforms. We propose that the variance of representation between the generated outputs and reality of racial, gender or age groups may be the product of low-quality training data or ineffective algorithmic intervention GAI developer perceptions of the quality of the underlying training data. In either case we posit that more representative, higher quality training data will improve the quality of GAI output.

To quantify the degree of observed variance, the research calculated the average values derived from all variables and from all three GAI platforms. The researcher also created Representation Rate (RR) based on average variances of variables in each GAI platform compared to the National Center for Education Statistics (NCES) and the California State University – Sacramento – Fall 2023 Graduate School census (CSU-GS). The findings indicate that each platform performed differently for each variable. Although the variance of variable input was extremely high among the three platforms and national averages, the RR value of each GAI platform was close.

These findings suggest an opportunity for better data in training Artificial Intelligence (AI) models, especially in academic settings where representation may influence future participation (Gardner, 2008). In conclusion, GAI tends to magnify any systemic errors present in the data used for training the models. We propose that by developing higher quality, more representative sources of training data higher quality GAI outputs may be achieved. The complexity of the question posed correlates with the risk of inaccuracies or divergence from reality. Enhancing the representativeness of the training data can lead to improved performance and more realistic AI models. Therefore, refining the data used to train GAI models is essential for advancing the quality and accuracy of artificial intelligence systems.

REFERENCES:

Aminee, A. Differences in Readmission and Length of Stay between Income Communities Before and During Covid-19. ProQuest Dissertations Publishing; 2022.

Cheng, X., Lin, X., Shen, X., Zarifis, A., & Mou, J. (2022). The dark sides of AI. National Library of Medicines. NIH. Published online 2022 Feb 22. doi: 10.1007/s12525-022-00531-5

Chien, C.-C., Chan, H.-Y., & Hou, H.-T. (2024). Learning by playing with generative AI: design and evaluation of a role-playing educational game with generative AI as scaffolding for instant feedback interaction. Journal of Research on Technology in Education, 1–20. https://doi-org.proxy.lib.CSU.edu/10.1080/15391523.2024.2338085

Creswell, A., White, T., Dumoulin, V. Arulkumaran, K. Sengupta, B. & Bharath, A. (2017). Generative Adversarial Networks: An Overview. *IEEE Signal Processing Magazine, vol. 35, no. 1, pp. 53-65, Jan. 2018, doi: 10.1109/MSP.2017.2765202*.

Eiseman, J., & Ortiz, N. (2023). GENERATIVE AI & MACHINE LEARNING IN LAW LIBRARIES: The benefits, risks, and ethical issues surrounding these potentially transformative new tools. AALL Spectrum, 27(5), 14–17.

Fox Business. (2024). Google's AI tool taken down after inaccurate historical photos. The Bottom Line with Degen and Duffy – February 22, 2024. Fox Business Network.

Gardner, S. K. (2008). Fitting the mold of graduate school: A qualitative study of socialization in doctoral education. Innovative higher education, 33, 125-138.

Grewal D, Guha A, Satornino C, Schweiger E. (2021). Artificial intelligence: The light and the darkness. Journal of Business Research. 2021; 136:229–236. doi: 10.1016/j.jbusres.2021.07.043.

Gugin, D. (2023). Artificial Intelligence & Generative AI for Beginners. Pacific Asia Inquiry, 14(1), 126–135.

He, Y., & Yao, C. (2021). Information structures and entropy measurement for a fuzzy probabilistic information system. Journal of Intelligent & Fuzzy Systems, 41(6), 6343–6361. https://doi-org.proxy.lib.CSU.edu/10.3233/JIFS-210149

Hershey, D. (2010). Entropy theory of aging systems humans, corporations and the universe / Daniel Hershey. Imperial College Press.

Institutional Research, Effectiveness – Enrollment Dashboard. (2024). Fall-2023 Graduate Program Dashboard. Office of the President. California State University – Sacramento. https://www.CSU.edu/president/institutional-research-effectiveness-planning/dashboards/enrollment.html.

Liu Y, Yan W, Hu B. (2021). Resistance to facial recognition payment in China: The influence of privacy–related factors. Telecommunications Policy. 2021;45(5):1021155. doi: 10.1016/j.telpol.2021.102155.

Ma, X., Sun, Y., Guo, X., Lai, K., & Vogel, D. (2021). Understanding users' negative responses to recommendation algorithms in short-video platforms: A perspective based on the stressor-strain-outcome (SSO) framework. Electronic Markets, 2021. 10.1007/s12525-021-00488-x

Marr, B. (2023). The 15 Biggest Risks of Artificial Intelligence. Enterprise Tech – Forbes. https://www.forbes.com/sites/bernardmarr/2023/06/02/the-15-biggest-risks-of-artificial-intelligence/?sh=11f01ec42706

Microsoft Corporation. (2024). Microsoft is born. Microsoft. https://news.microsoft.com/announcement/microsoft-is-born/

National Center for Education Statistics. (2023). Postbaccalaureate Enrollment. Condition of Education. U.S. Department of Education, Institute of Education Sciences. Retrieved 04/30/2024, from https://nces.ed.gov/programs/coe/indicator/chb.

OpenAI. (2024). ChatGPT Reached 180.5 million Users. OpenAI. https://openai.com/chatgpt/

Raghavan, P. (2024). Gemini image generation got it wrong. We will do better. Google Inc. https://blog.google/products/gemini/gemini-image-generation-issue/

Reddy, S. (2024). Generative AI in healthcare: an implementation science informed translational path on application, integration and governance. Implementation Science, 19(1), 1–15. https://doi-org.proxy.lib.CSU.edu/10.1186/s13012-024-01357-9

Smythe, I. (2023). The impact of generative AI upon the evaluation process. Assessment & Development Matters, 15(3), 30–35. https://doi-org.proxy.lib.CSU.edu/10.53841/bpsadm.2023.15.3.30

Starr, M. (2015). Entropy: the natural order is disorder. United Sates Air Force. https://www.vance.af.mil/News/Commentaries/Display/Article/636873/entropy-the-natural-order-is-disorder/

Sun, J., Song, J., Jiang, Y., Liu, Y., & Li, J. (2021). Prick the filter bubble: A novel cross domain recommendation model with adaptive diversity regularization. Electronic Markets. 10.1007/s12525-021-00492-1

Surowiecki, J. (2004). The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies, and Nations. Doubleday

Ungureanu, C. T., & Amironesei, A. E. (2023). Legal issues concerning Generative AI technologies. Eastern Journal of European Studies, 14(2), 45–75. https://doi-org.proxy.lib.CSU.edu/10.47743/ejes-2023-0203

Valle-Cruz, D., García-Contreras, R., & Gil-Garcia, J. R. (2024). Exploring the negative impacts of artificial intelligence in government: the dark side of intelligent algorithms and cognitive machines. International Review of Administrative Sciences, 90(2), 353-368. https://doi.org/10.1177/00208523231187051

Vynck, G., & Tiku, N. (2024). Google takes down Gemini AI image generator. Here is what you need to know. The Washington Post. https://www.washingtonpost.com/technology/2024/02/22/google-gemini-ai-image-generation-pause/