Does LLM Have Bias Against Ethnic Minorities or Violence?: Topic-Specific Bias in LLM Synthetic Data Comparing Non-Multicultural and Multicultural Youths

Anonymous ACL submission

Abstract

Previous studies examined whether LLMs can replicate majority group opinions, raising concerns about effectiveness for minorities and bias on sensitive topics. To address these concerns, we evaluated LLM-generated synthetic survey data for South Korean adolescents from ethnic majority and minority families using 1,600 respondents from 2012. Results showed no group differences, but synthetic data systematically overestimated sensitive topics (violence, delinquency, depression) while accurately matching non-sensitive ones (family relationships). Findings suggest LLM bias varies by topic sensitivity rather than minority status, and reveal limitations in capturing sensitive personal experiences through demographic information alone.

1 Introduction

Synthetic data is LLM-generated data using GPT as a survey respondent, in order to address data-related challenges across various scientific fields. Recently, the potential of synthetic data is being increasingly explored in social sciences research (Bail, 2024; Grossmann et al., 2023; Dillion et al., 2023; Veselovsky et al., 2023; Horton, 2023), by predicting political opinions (Argyle et al., 2023) or survey responses to questions that could not be asked due to missing values or historical social perceptions (e.g. questions regarding support for homosexuality) (Kim and Lee, 2023). This allowed researchers to infer past public perceptions that no longer exist in the documents, preventing the risk of unintentional or deliberate exposure of respondents' private information (Raghunathan, 2021).

Synthetic data allows for the prediction of data through LLMs without the need for direct surveying. This offers significant cost-efficiency advantages for survey research (Horton, 2023). Additionally, it allows for the estimation of *non-existent* data that cannot be collected due to physical limitations, producing results that closely resemble reality. However, LLMs depict the world with greater bias than reality (Wang et al., 2024), which is also problematically evident in synthetic data.

On the other hand, surveying minority groups has limitations. Their smaller sample sizes make it difficult to capture their characteristics sufficiently through surveys (Herzing et al., 2019). Also, fear of identity disclosure, concerns about stigmatization, and distrust stemming from discriminatory experiences can lead to survey avoidance or defensive responses (George et al., 2014). The subtle nature of daily discrimination is often ambiguous and difficult to interpret; victims are frequently unable to clearly recognize the discrimination they experience or its emotional impact (Sue et al., 2007), and this ambiguity results in under-reporting or misinterpretation of experiences in self-report surveys, complicating accurate measurement of microaggression effects (Wong et al., 2014; Castillo et al., 2025).

Moreover, stigma inherently centers on negative concepts, and societal stereotypes have historically linked minority groups with negative attributes such as violence, substance abuse, and mental illness (Yang et al., 2017). Large language models not only reflect these real-world biases but amplify them, producing more stereotyped representations of minorities than exist in their training data (Fazelpour and Danks, 2021; Wang et al., 2024). When combined with the difficulty of obtaining accurate responses to sensitive questions in minority surveys due to social desirability bias (Krumpal, 2013), this algorithmic amplification raises concerns that synthetic data generated by LLMs may exhibit heightened bias on sensitive topics, potentially distorting rather than faithfully reproducing the true experiences of marginalized populations.

Previous studies have not adequately validated the accuracy of predicting minority opinions using synthetic data compared to those of majority

groups. While LLM-generated synthetic data has been moderately effective in general surveys, its effectiveness in predicting minority opinions remains unclear due to fundamental challenges in minority research and an insufficient understanding of the biases LLMs have in generating minority data. The possibility to predict outcomes like depression and domestic violence in hard-to-reach minority populations using minimal demographic information would have significant implications for early identification and prevention of public health issues.

We examined the efficacy of synthetic data by generating survey responses for ethnic majority (non-multicultural) and minority (multicultural) adolescents in South Korea. We focused on whether the accuracy varied between sensitive and non-sensitive questions. Using 2012 comparative survey data on sociodemographic information, we found no significant difference in synthetic data accuracy between groups. However, the synthetic data significantly overestimated the prevalence of socially sensitive topics (e.g., violence, delinquency, and depression) compared to the actual data. These results suggest that LLMs' performance in generating synthetic data varies by question sensitivity rather than by the respondents' status—minority or majority.

2 Experiment

2.1 Data

The ground truth dataset used survey data from *the Comparative Study on Delinquency Between Multicultural and General Family Adolescents* (Jeon et al., 2016). The survey included questions about socio-demographic information, such as family background, socioeconomic status, and school life. According to Republic of Korea (2020), the multicultural family is defined as one composed of immigrants who have married with Korean citizens or foreigners who have acquired Korean nationality or received permission to be naturalized. The survey was given to 800 non-multicultural and 800 multicultural adolescents, aged 11–19, who attend elementary, middle, and high schools, as well as community children's centers in Korea (excluding Gangwon and Jeju). A key aspect of the survey is that the same set of questions was asked of both the ethnic majority group of non-multicultural adolescents and the ethnic minority group of multicultural adolescents.

2.2 Analytic Strategy

Synthetic data is generated by creating virtual personas based on demographic information and using them to prompt LLMs to respond to survey questions (Bisbee et al., 2024; Kim and Lee, 2023). We created virtual personas with 11 demographic attributes: gender, age, grade level, country of birth, length of residence in Korea, father's and mother's birth country, parents' marital status, household members, residential administrative district, and subjective economic status. Prior research shows synthetic data commonly exhibits means closely resembling real-world data with substantially reduced variation (Bisbee et al., 2024). Using mean and variance as primary indicators, we examined mean differences between synthetic and actual ground truth data for 77 questions across multicultural and non-multicultural adolescents. We then analyzed how mean differences vary according to minority characteristics and identified systematic patterns in distribution differences between original and synthetic data.

3 The Process of Creating Synthetic Data

3.1 Prompts Design

The persona prompts in this study comprised 10 variables based on the Korea Population and Housing Census (Korea, 2016) and one survey-specific variable. The question prompts were structured with response options identical to those in the ground truth dataset survey. However, for questionnaires where having experience was critical (e.g., violence), the original response categories were binarized into two categories: those who experienced the event at least once (Yes) and those who never experienced it (No). The original categories were: never experienced (1), experienced 1–2 times (2), 3–4 times (3), 5–6 times (4), 7 or more times (5). The prompts were presented in Korean, consistent with the original survey.

3.2 Model

According to Ma et al. (2025), the llama-2-13b-chat-hf model outperforms other models (gemma-7B-IT and mixtral-8x7b-instruct-v0.1) in reproducing the sociocultural context of human responses. Therefore, this study employed the llama3:latest model, which is a more advanced version of the Llama model series.

3.3 Preprocessing

The ground truth dataset, also referred to as the original dataset, contained 16 missing values (0.001% of the total) in personal variables and 254 missing values (0.002% of the total) in question variables. Missing country-related persona variable values were replaced with South Korea, while difficult-to-estimate values, such as economic status, were left as non-responses. For the question variables, the K-nearest neighbors (KNN) machine learning technique was used to estimate similar values by considering all the variables.

To conduct a comparative analysis, both the original and synthetic data were aggregated by topic. For relationship-related questions, averages were calculated to indicate overall relationship quality. Questions about violence, abuse, and delinquency were converted through dummy coding. The scale ranged from never experienced (1) to experienced seven or more times (5). These values were then transformed into a binary scale of never experienced (0) or experienced at least once (1) and summed across all items. Depression levels were measured by evaluating self-esteem (e.g., "I am as important as other people") and negative cognition (e.g., "I have thoughts of wanting to die") on a scale from Not at all true (1) to Very true (5). Since depression measurement focuses on symptom accumulation, the scores were summed after coding more positive responses as 1 and more negative responses as 5. Finally, min-max normalization was applied to all variables based on the theoretically possible range for each survey item.

4 Results

Overall, there is no significant difference in the mean gap between original and synthetic data for non-multicultural or multicultural students. The difference between the averages of the original and synthetic data indicates how closely the LLM generates data that resembles real data. In other words, a larger difference indicates that the LLM is less able to generate data similar to real-world data, while a smaller difference indicates that the LLM is more successful in doing so. According to *Figure 1*, the difference between the original and synthetic data was less than 0.1 for all items and for both non-multicultural and multicultural students. In other words, the LLM does not generate synthetic data that particularly resembles real-world data for the social majority.

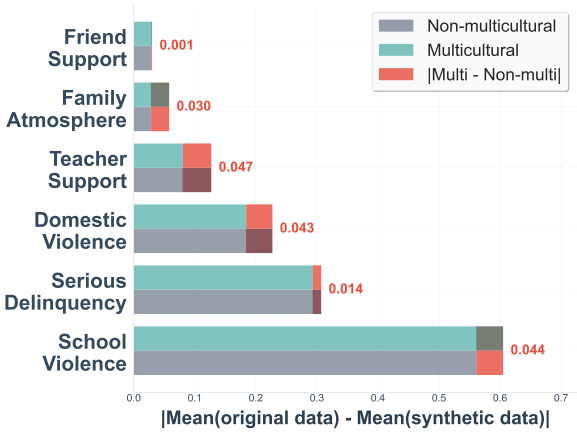

Figure 1: Mean gaps between original and synthetic data comparing non-multicultural and multicultural students. There is no significant difference in the mean gap between original and synthetic data for non-multicultural or multicultural students.

However, the mean difference between original and synthetic data varies significantly across topic categories, particularly for negative topics such as violence, delinquency, and depression. *Figure 2* shows that topics containing socially sensitive concepts, such as school violence, child abuse, juvenile delinquency, depression, and domestic violence, exhibit a much larger average difference between synthetic and actual data than relatively non-sensitive topics, such as family intimacy. The mean gap between synthetic and original data was large for categories containing sensitive concepts, ranging from 0.60 to 0.19. In contrast, all non-sensitive categories had a mean gap less than 0.15. LLMs generated more similar data for relatively non-sensitive topics like family intimacy and less similar data for socially sensitive topics like violence.

Furthermore, the variance of synthetic data differs slightly across topics. Nevertheless, the difference in mean is highly pronounced in sensitive topics, maximizing the disparity in distribution. This study also found that all synthetic data exhibited low variance values, regardless of topic. On average, the variance difference between synthetic and original data was 0.026 for non-multicultural students and 0.035 for multicultural students in sensitive topics. For non-sensitive topics, the variance difference was 0.043 for non-multicultural students and 0.049 for multicultural students. *Figure 3* reveals significant differences in distribution between sensitive and non-sensitive topics. This appears to be because large mean differences across topics significantly influence the overall distribution.

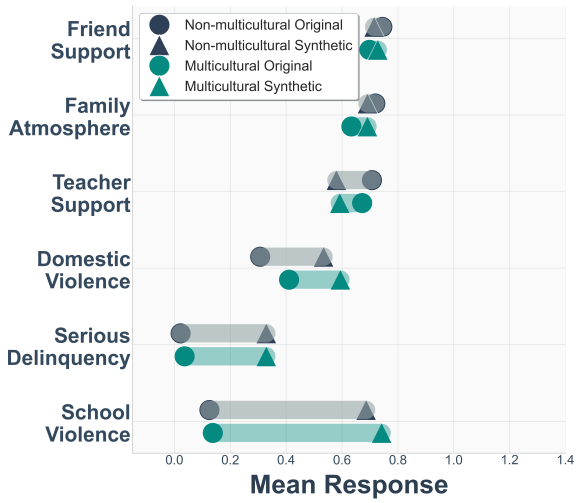

Figure 2: Mean gaps between original and synthetic data for non-multicultural and multicultural students across different topic categories. The mean difference between original and synthetic data varies significantly across topics.

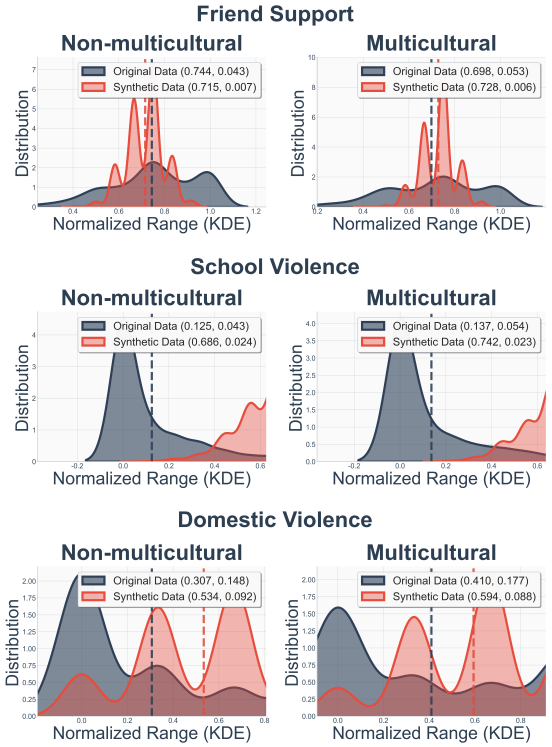

Figure 3: Distribution comparison of original and synthetic data for non-multicultural (top) and multicultural (bottom) students. KDE curves may extend beyond [0, 1] due to boundary smoothing effects. Dashed lines indicate mean values.

5 Conclusion

5.1 Discussion

We compared how well synthetic data replicated actual data for adolescents from majority non-multicultural and minority multicultural families. Although performance differences between groups were not significant, synthetic data scores were notably higher than actual data scores for socially sensitive topics such as violence, delinquency, and depression. These results indicate that LLMs' performance varies based on question sensitivity rather than respondents' multicultural backgrounds. We propose two interpretations. First, LLMs exhibit systematic bias toward socially sensitive topics. Attention mechanisms assign greater weight to exceptional and extreme cases (Mehrabi et al., 2021), making shocking situations more likely to capture the model's attention (Gehman et al., 2020; Kiritchenko and Mohammad, 2018). Second, gauging opinions on sensitive topics is inherently difficult due to limitations in both LLMs and survey methodology itself. Social desirability bias strongly influences responses, causing respondents to provide socially acceptable answers regardless of their actual experiences (Bispo Júnior, 2022; Latkin et al., 2017). Therefore, both synthetic data generation and actual surveys may be insufficient for obtaining truthful information on sensitive topics, suggesting that accurate information requires in-depth interviews or qualitative research methods.

5.2 Limitations

This study reveals limitations of using demographic information for synthetic data generation. Sensitive personal experiences—such as violence, abuse, and depression—relate more to experiential context and relationships than demographics like age or gender. Experiences like domestic violence link to specific family dynamics rather than demographic attributes, making demographic information alone inadequate for replicating response patterns in personal domains.

5.3 Ethical Considerations

Our study reveals important risks in using LLM-generated synthetic data. The systematic overestimation of sensitive topics could mislead policy decisions without proper validation. While we found no bias against multicultural adolescents, demographic prompts alone may perpetuate oversimplified representations of minority experiences.

References

Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.

Christopher A Bail. 2024. Can generative ai improve social science? *Proceedings of the National Academy of Sciences*, 121(21):e2314021121.

James Bisbee, Joshua D Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer M Larson. 2024. Synthetic replacements for human survey data? the perils of large language models. *Political Analysis*, 32(4):401–416.

José Patrício Bispo Júnior. 2022. Social desirability bias in qualitative health research. *Revista de saude publica*, 56:101.

Wendy Castillo, Rachel Renbarger, Sasha Mejia-Bradford, Christen Priddie, Juan Cruz, Brein Mosely, and Katherine Aragon. 2025. Unveiling racism: A systematic review of survey measures of racism in education. edworkingpaper no. 25-1168. *Annenberg Institute for School Reform at Brown University*.

Danica Dillion, Niket Tandon, Yuling Gu, and Kurt Gray. 2023. Can ai language models replace human participants? *Trends in Cognitive Sciences*, 27(7):597–600.

Sina Fazelpour and David Danks. 2021. Algorithmic bias: Senses, sources, solutions. *Philosophy Compass*, 16(8):e12760.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *arXiv preprint arXiv:2009.11462*.

Sheba George, Nelida Duran, and Keith Norris. 2014. A systematic review of barriers and facilitators to minority research participation among african americans, latinos, asian americans, and pacific islanders. *American journal of public health*, 104(2):e16–e31.

Igor Grossmann, Matthew Feinberg, Dawn C Parker, Nicholas A Christakis, Philip E Tetlock, and William A Cunningham. 2023. Ai and the transformation of social science research. *Science*, 380(6650):1108–1109.

Jessica ME Herzing, Guy Elcheroth, Oliver Lipps, and Brian Kleiner. 2019. Surveying national minorities. Technical report, University of Lausanne; FORS.

John J Horton. 2023. Large language models as simulated economic agents: What can we learn from homo silicus? Technical report, National Bureau of Economic Research.

Young-sil Jeon, Dongjun Shin, and Sanghee Park. 2016. Comparative survey on delinquency of juveniles from multicultural families and juveniles from non-multicultural families, 2012. Korean Social Science Data Archive (KOSSDA). Dataset, released April 11, 2016.

Junsol Kim and Byungkyu Lee. 2023. Ai-augmented surveys: Leveraging large language models and surveys for opinion prediction. *arXiv preprint arXiv:2305.09620*.

Svetlana Kiritchenko and Saif M Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. *arXiv preprint arXiv:1805.04508*.

Statistics Korea. 2016. *Population and Housing Census Report: Complete Enumeration Results*, volume 1. Daejeon: Statistics Korea. Partially published from 2016-2021.

Ivar Krumpal. 2013. Determinants of social desirability bias in sensitive surveys: a literature review. *Quality & Quantity*, 47:2025–2047.

Carl A Latkin, Catie Edwards, Melissa A Davey-Rothwell, and Karin E Tobin. 2017. The relationship between social desirability bias and self-reports of health, substance use, and social network factors among urban substance users in baltimore, maryland. *Addictive behaviors*, 73:133–136.

Bolei Ma, Berk Yoztyurk, Anna-Carolina Haensch, Xinpeng Wang, Markus Herklotz, Frauke Kreuter, Barbara Plank, and Matthias Aßenmacher. 2025. Algorithmic fidelity of large language models in generating synthetic German public opinions: A case study. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1785–1809, Vienna, Austria. Association for Computational Linguistics.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35.

Trivellore E Raghunathan. 2021. Synthetic data. *Annual review of statistics and its application*, 8(1):129–140.

Republic of Korea. 2020. Multicultural families support act. Korean Law Information Center. Act No. 17281, enacted March 21, 2008, last amended May 19, 2020.

Derald Wing Sue, Christina M Capodilupo, Gina C Torino, Jennifer M Bucceri, Aisha Holder, Kevin L Nadal, and Marta Esquilin. 2007. Racial microaggressions in everyday life: implications for clinical practice. *American psychologist*, 62(4):271.

Veniamin Veselovsky, Manoel Horta Ribeiro, Akhil Arora, Martin Josifoski, Ashton Anderson, and Robert West. 2023. Generating faithful synthetic data with large language models: A case study in computational social science. *arXiv preprint arXiv:2305.15041*.

Ze Wang, Zekun Wu, Jeremy Zhang, Xin Guan, Navya Jain, Skylar Lu, Saloni Gupta, and Adriano Koshiyama. 2024. Bias amplification: Large language models as increasingly biased media. *arXiv preprint arXiv:2410.15234*.

Gloria Wong, Annie O Derthick, EJR David, Anne Saw, and Sumie Okazaki. 2014. The what, the why, and the how: A review of racial microaggressions research in psychology. *Race and social problems*, 6(2):181–200.

Lawrence H Yang, Linda Y Wong, Melissa M Grivel, and Deborah S Hasin. 2017. Stigma and substance use disorders: an international phenomenon. *Current Opinion in Psychiatry*, 30(5):378.

A Prompt Examples

A.1 English version

Persona: It is now July 2012. You are 15 years old, female, born in China, and have been living in Seoul, a capital city in the Republic of Korea for 8 years as a student in middle school 3rd grade from a multicultural family. Your mother's nationality is China, and your father's nationality is South Korea. Your parents are currently married, and you live with both parents. Your family's income is middle class.

Question: Do you think you have no interest or enthusiasm in anything?
Options: (a) Not at all (b) Not really (c) Neither agree nor disagree (d) Mostly agree (e) Strongly agree

Constraint: Answer using the provided options only. Respond in Korean only; do not use English.

Command: Answer the question from the perspective of the respondent with the specified persona characteristics.

A.2 Korean version

페르소나: 지금은 2012년 7월이다. 너는 13살 여성이며, 중국에서 태어나 대한민국의 수도인 서울에 8년째 살고 있는 다문화 가정 중학교 3학년 학생이다. 어머니의 국적은 중국(조선족)이고, 아버지의 국적은 한국이다. 부모님은 현재 결혼 중이며, 너는 친부모 두 분과 살고 있다. 가족의 소득은 중간 정도이다.

질문: 당신은 당신 스스로가 모든 일에 관심과 흥미가 없다고 생각하시나요?
보기: (a) 전혀 그렇지 않다 (b) 별로 그렇지 않다 (c) 그저 그렇다 (d) 대체로 그렇다 (e) 매우 그렇다

제약 조건: 주어진 보기 안에서 대답해줘. 한국어로만 답변하고 영어는 사용하지 마.

명령: 명시된 페르소나 특성을 지닌 응답자의 관점에서 질문에 답해줘.

B Comparison of Descriptive Statistics between Synthetic and Ground-Truth (Original) Data

B.1 Non-multicultural Youths

Table 1: Comparison of Descriptive Statistics between Synthetic and Original Data of Non-multicultural Youths

Variable	Data Type	Mean	Variance	SE
Family Atmosphere	Original	3.77	0.21	0.01
	Synthetic	3.88	0.77	0.03
Father Intimacy	Original	3.56	1.09	0.04
	Synthetic	4.01	0.38	0.01
Mother Intimacy	Original	4.07	0.89	0.03
	Synthetic	4.19	0.30	0.01
Teacher Support	Original	3.83	0.96	0.03
	Synthetic	3.52	0.54	0.02
Friend Support	Original	3.98	0.83	0.03
	Synthetic	3.86	0.33	0.01
Domestic Violence	Original	0.92	1.15	0.04
	Synthetic	1.60	0.91	0.03
Child Abuse	Original	1.21	2.16	0.08
	Synthetic	4.66	1.77	0.06
School Violence	Original	1.13	1.86	0.07
	Synthetic	6.18	1.40	0.05
Status Delinquency	Original	0.22	0.64	0.02
	Synthetic	1.85	1.33	0.05
Serious Delinquency	Original	0.24	0.70	0.02
	Synthetic	3.61	1.60	0.06
Depression	Original	30.35	10.24	0.36
	Synthetic	47.88	3.26	0.12

B.2 Multicultural Youths

Table 2: Comparison of Descriptive Statistics between Synthetic and Original Data of Multicultural Youths

Variable	Data Type	Mean	Variance	SE
Family Atmosphere	Original	3.53	0.78	0.03
	Synthetic	3.76	0.20	0.01
Father Intimacy	Original	3.24	1.08	0.04
	Synthetic	4.06	0.32	0.01
Mother Intimacy	Original	3.71	0.96	0.03
	Synthetic	4.22	0.29	0.01
Teacher Support	Original	3.69	0.94	0.03
	Synthetic	3.37	0.54	0.02
Friend Support	Original	3.79	0.92	0.03
	Synthetic	3.91	0.30	0.01
Domestic Violence	Original	1.23	1.26	0.04
	Synthetic	1.78	0.89	0.03
Child Abuse	Original	1.65	2.77	0.10
	Synthetic	5.92	1.72	0.06
School Violence	Original	1.23	2.09	0.07
	Synthetic	6.67	1.37	0.05
Status Delinquency	Original	0.38	1.01	0.04
	Synthetic	1.93	1.30	0.05
Serious Delinquency	Original	0.40	1.13	0.04
	Synthetic	3.62	1.53	0.05
Depression	Original	31.93	10.27	0.36
	Synthetic	49.21	3.00	0.11

C Theoretical Range of Survey Variables for Normalization

Table 3: Theoretical Range of Survey Variables for Normalization

Variable	Minimum	Maximum
Family Atmosphere	1	5
Father Intimacy	1	5
Mother Intimacy	1	5
Teacher Support	1	5
Friend Support	1	5
Domestic Violence	0	3
Child Abuse	0	10
School Violence	0	9
Status Delinquency	0	8
Serious Delinquency	0	11
Depression	15	75

Note: These values represent the theoretical minimum and maximum scores possible in the survey instrument for each variable, used for normalization purposes.

D Comparison of Descriptive Statistics between Synthetic and Ground-Truth (Original) Data (Normalized)

D.1 Non-multicultural Youths (Normalized)

Table 4: Comparison of Descriptive Statistics between Synthetic and Original Data of Non-multicultural Youths (Normalized)

Variable	Data Type	Mean	Variance	SE
Family Atmosphere	Original	0.72	0.19	0.01
	Synthetic	0.69	0.05	0.00
Father Intimacy	Original	0.64	0.27	0.01
	Synthetic	0.75	0.10	0.00
Mother Intimacy	Original	0.77	0.22	0.01
	Synthetic	0.80	0.07	0.00
Teacher Support	Original	0.71	0.24	0.01
	Synthetic	0.58	0.14	0.00
Friend Support	Original	0.74	0.21	0.01
	Synthetic	0.72	0.08	0.00
Domestic Violence	Original	0.31	0.38	0.01
	Synthetic	0.53	0.30	0.01
Child Abuse	Original	0.12	0.22	0.01
	Synthetic	0.47	0.18	0.01
School Violence	Original	0.13	0.21	0.01
	Synthetic	0.69	0.16	0.01
Status Delinquency	Original	0.03	0.08	0.00
	Synthetic	0.23	0.17	0.01
Serious Delinquency	Original	0.02	0.06	0.00
	Synthetic	0.33	0.15	0.01
Depression	Original	0.26	0.17	0.01
	Synthetic	0.55	0.05	0.00

D.2 Multicultural Youths (Normalized)

Table 5: Comparison of Descriptive Statistics between Synthetic and Original Data of Multicultural Youths (Normalized)

Variable	Data Type	Mean	Variance	SE
Family Atmosphere	Original	0.63	0.20	0.01
	Synthetic	0.69	0.05	0.00
Father Intimacy	Original	0.56	0.27	0.01
	Synthetic	0.77	0.08	0.00
Mother Intimacy	Original	0.68	0.24	0.01
	Synthetic	0.81	0.07	0.00
Teacher Support	Original	0.67	0.24	0.01
	Synthetic	0.59	0.13	0.00
Friend Support	Original	0.70	0.23	0.01
	Synthetic	0.73	0.07	0.00
Domestic Violence	Original	0.41	0.42	0.01
	Synthetic	0.59	0.30	0.01
Child Abuse	Original	0.17	0.28	0.01
	Synthetic	0.59	0.17	0.01
School Violence	Original	0.14	0.23	0.01
	Synthetic	0.74	0.15	0.01
Status Delinquency	Original	0.05	0.13	0.00
	Synthetic	0.24	0.16	0.01
Serious Delinquency	Original	0.04	0.10	0.00
	Synthetic	0.33	0.14	0.00
Depression	Original	0.28	0.17	0.01
	Synthetic	0.57	0.05	0.00

E Data and Implementation Details

E.1 Data Licensing and Access

The original dataset (Jeon et al., 2016) is available through KOSSDA (https://kossda.snu.ac.kr/handle/20.500.12236/15861) under their institutional data use agreement. The data was collected with informed consent from participants and approved by the Korean Institute of Criminology's ethics review board. KOSSDA's terms require: (1) proper citation using the provided DOI, (2) use restricted to academic research purposes, (3) prohibition of re-identification attempts, and (4) no redistribution to third parties without authorization.

The original dataset was pre-anonymized by the data provider prior to public release. All personally identifiable information, including names, addresses, and school identifiers, were removed during the original data collection process. Both the original study and our use of this archival data for synthetic data generation share a common non-commercial academic research purpose. Our use complies with KOSSDA's data use agreement, which permits re-use of archival data for scholarly research activities.

The synthetic data generated in this study was created using only demographic categories (age, gender, family type) and response patterns from the original survey. The resulting synthetic dataset contains no personal information that could be used to identify specific individuals. This synthetic data is intended solely for academic research purposes, specifically for evaluating the capability and limitations of large language models in generating survey data that preserves statistical properties of real-world distributions.

E.2 LLM API Specifications

We used Llama3:latest (Meta AI) for all synthetic data generation. The model was accessed via Ollama with default parameters except for temperature, which was set to 0.7. All other generation parameters (top_p, top_k, max_tokens, etc.) used the default Ollama settings. Total API calls: approximately 123,200 (800 samples per group $\times$ 2 groups $\times$ 77 questions). The generation was conducted on a local server with GPU acceleration.

E.3 Software and Package Versions

All data processing and statistical analyses were conducted using Python 3.13.2 with the following packages:

- pandas (v2.2.3)
- numpy (v2.2.4)
- pyreadstat (v1.2.8)
- scikit-learn (v1.6.1)
- scipy (v1.15.2)
- matplotlib (v3.10.1) and seaborn (v0.13.2)
- ollama (v0.5.1)
- transformers (v4.54.1)
- tqdm (v4.67.1)

E.4 Use of AI Assistants

During the preparation of this work, the authors used the following tools: Claude Sonnet 4.5 (accessed through the Claude and Perplexity websites) and DeepL (an AI translator) to improve the coding and refine the writing (e.g., grammar checking). The authors reviewed and edited all AI-generated content and take full responsibility for the accuracy and integrity of the published work. AI tools were not used for data analysis, interpretation, or research conclusions.