Limited Differences in LLM Performance for Social Majorities and Minorities in Survey Synthetic Data

Anonymous ACL submission

Abstract

Large Language Models (LLMs) are increasingly used to generate synthetic survey data, yet it remains unclear whether they perform equally well for social minorities and majorities. Using a unique survey conducted in South Korea in 2012 that administered identical questionnaires to adolescents from multicultural and monocultural families, we examine group-level differences in the accuracy of LLM-generated synthetic data. We construct personas that mirror individual-level demographic profiles and prompt two LLMs, `llama3:latest` and `ChatGPT-4o`, to answer 77 survey questions across 11 question domains. We find LLMs' limited performance differences between minority and majority groups. Instead, accuracy varies by question domain and LLM. Across all domains, synthetic data generally exhibit lower variance than human responses. These findings highlight both the potential and the limitations of using LLMs to generate synthetic survey data for studying social minorities.

1 Introduction

Because traditional survey methods require substantial time, energy, and money, scholars have increasingly explored the use of Large Language Models (LLMs) to generate survey responses (Bail, 2024; Grossmann et al., 2023; Dillion et al., 2023; Veselovsky et al., 2023; Horton, 2023). As these responses are produced through the synthesis of patterns learned from existing text data, they are commonly referred to as synthetic data. Even when provided with only limited demographic information such as age, gender, and place of residence, synthetic data can relatively successfully approximate the average opinions of the general public (Bail, 2024; Grossmann et al., 2023; Argyle et al., 2023; Bisbee et al., 2024; Sun et al., 2024; Ma et al., 2025; Kim and Lee, 2023).

Little research has examined whether LLMs perform equally well in generating synthetic data for social majorities and minorities. When data for a particular group are sparse, algorithms tend to exhibit lower accuracy and reliability for that group (Shafayat et al., 2024; Buolamwini, 2017; Raji et al., 2020; Hofmann et al., 2024). Moreover, algorithms not only reflect real-world biases but may also amplify them in stereotypical ways (Fazelpour and Danks, 2021; Wang et al., 2024). In this context, synthetic data generated by LLMs may show lower predictive performance for minorities than for those of majorities. Such unequal performance may undermine the utility of synthetic data because social minorities are often harder to reach due to their small numerical representation (Herzing et al., 2019) and lower levels of social visibility or participation (George et al., 2014; Song, 2020; Bernasconi, 2000). Yet, they are frequently groups that require government support (Weiss, 1977; Herzing et al., 2019). Therefore, it is important to assess whether LLM performance in generating synthetic data is equitable for social majorities and minorities, and whether such equity holds across different domains of survey questionnaires.

Accordingly, we ask whether synthetic data can predict the survey responses of social minorities as reliably as those of social majorities. We use a unique survey dataset collected in South Korea in 2012 that includes equal numbers of adolescents from social minority groups, specifically individuals from multicultural families who are first- or second-generation immigrants, and, as a comparison group, adolescents from monocultural families in which both parents are Korean. Treating this dataset as ground truth, we generate synthetic data by using the survey's demographic information as personas and evaluate performance by comparing the means and variances of the synthetic responses with those of the original survey. We find little overall difference in LLM performance between social minorities and majorities; however, performance varies by the type of questions asked.

2 Data and Methods

2.1 Data

Here, we define a social minority as adolescents in multicultural families in South Korea(Republic of Korea, 2020). While population size alone does not define a minority, it is an important consideration in our context because we examine a case in which minorities generate less data needed for LLM training. In 2015, individuals in multicultural families are approximately 1.53% of the total population(Korean Statistical Information Service, 2015b,a). Multicultural families have been subject to various forms of discrimination, including sexual violence, social prejudice, and verbal abuse, in both personal relationships and public services(Kim, 2015).

In our study, a key requirement for ground truth data is to have the comparative survey results of minorities and majorities. We therefore select a survey on delinquent behaviors administered to both monocultural and multicultural adolescents(Jeon et al., 2016). The survey includes questions on sociodemographic characteristics such as family background, and school and life satisfaction. Respondents were adolescents aged 11 to 19 who attended elementary, middle, and high schools in South Korea, with each survey including responses from 800 respondents.

2.2 Analytic Strategy

Following the previous approaches in creating synthetic data(Argyle et al., 2023), we design hypothetical personas with demographic information from surveys and ask LLMs to answer survey questionnaires with the given personas. We use 11 sociodemographic variables (Table F.3 for the full list). We create 1,600 personas, including 800 adolescents from multicultural and 800 from monocultural families, by directly mapping each observation in the ground truth data to a unique persona with identical demographic attributes.

Then, using these personas, we ask an LLM to answer 77 survey questions, which are categorized into eleven domains such as experiencing domestic violence, child abuse or school violence (Table F.3 for the full list). We pose each question for every persona, generating 246,400 responses in total (800 responses $\times$ 2 groups $\times$ 77 questions $\times$ 2 models). Answers to these questions are measured either on a continuous Likert scale or as dummy variables. Details on question measurement are provided in Table F.3. We test two LLMs for comparison with the default parameter settings. Specifically, we employ `llama3:latest` and `ChatGPT-4o`.

2.3 Prompt Design and Preprocessing

The persona prompts are composed of the 11 sociodemographic variables. The possible answer options in the question prompts are largely identical to those used in the ground truth survey. All prompts are provided in Korean, as the original surveys. We also include the survey date at the beginning of each prompt, corresponding to the date on which the original survey was conducted.

```
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.
```

Figure 1: Prompt Example (English Version)

Before proceeding, we first address missing values. The details of how we treat missing values are in Appendix B.

2.4 Evaluation Metrics

The primary goal of our experiment is to evaluate the performance of LLMs in generating synthetic data relative to the ground truth. To achieve this goal, we use Cohen's d, which measures the practical significance of differences between two groups (Cohen, 1960), in this case, the synthetic data and the ground truth.

In addition, we compare the distributions of responses between the synthetic data and the ground truth by examining their means and variances. Differences in means between human responses and synthetic data indicate how closely LLMs replicate average human responses, whereas differences in variances reflect how well LLMs capture the diversity of human experiences and opinions.

3 Results

In Table 1, we compare Cohen's d values between adolescents from multicultural and monocultural families across 11 domains. We find no clear systematic pattern in Cohen's d between the

two groups. LLMs sometimes exhibit better performance in predicting responses for adolescents from multicultural families and, in other cases, for those from monocultural families. For example, this pattern is observed in the domains of family atmosphere and father intimacy for `ChatGPT-4o`, as well as teacher support for `llama3:latest`. In contrast, in the domains of family atmosphere, father intimacy, and mother intimacy for `llama3:latest`, LLM-generated responses are more similar to those of human respondents for adolescents from monocultural families than for those from multicultural families. At the same time, in many domains, there are only small differences in LLM performance between the two groups.

Table 1: Cohen's D Results by Groups and Categories

		ChatGPT		Llama	
Domain	**Group**	Cohen's D	Interpretation	Cohen's D	Interpretation
Family Atmosphere	Monocultural	0.38	small	0.20	negligible
	Multicultural	0.02	negligible	−0.41	small
Father Intimacy	Monocultural	0.56	medium	−0.55	medium
	Multicultural	0.19	negligible	−1.03	large
Mother Intimacy	Monocultural	1.00	large	−0.18	negligible
	Multicultural	0.43	small	−0.72	medium
Teacher Support	Monocultural	1.11	large	0.65	medium
	Multicultural	0.95	large	0.42	small
Friend Support	Monocultural	1.14	large	0.18	negligible
	Multicultural	0.99	large	−0.18	negligible
Domestic Violence	Monocultural	0.83	large	−0.66	medium
	Multicultural	1.00	large	−0.51	medium
Child Abuse	Monocultural	0.65	medium	−1.75	large
	Multicultural	0.60	medium	−1.85	large
School Violence	Monocultural	0.49	small	−3.06	large
	Multicultural	−0.15	negligible	−3.08	large
Status Delinquency	Monocultural	0.09	negligible	−1.56	large
	Multicultural	0.20	negligible	−1.33	large
Serious Delinquency	Monocultural	−0.16	negligible	−2.73	large
	Multicultural	0.20	small	−2.39	large
Depression	Monocultural	−0.77	medium	−2.31	large
	Multicultural	−0.86	large	−2.28	large

It is also notable that the domains in which LLMs fail to generate reliable responses vary by model. For `llama3:latest`, the gaps between human responses and synthetic data are substantially larger in socially stigmatized domains such as violence, delinquency, and depression. The left panel of Figure 2 presents the three domains with the smallest gaps and the three domains with the largest gaps. Full results for all 11 domains are reported in Appendix D. The figure shows that Llama tends to overestimate the frequency of experiencing these socially stigmatized behaviors. However, the performance of `llama3:latest` does not differ by multicultural or monocultural adolescents.

The results of `ChatGPT-4o` are different. While `ChatGPT-4o` generally makes the better prediction of the average response, the accuracy level of prediction differs by question domains. `ChatGPT-4o` shows the better performance in predicting responses of socially stigmatized behaviors such as status delinquency or serious delinquency, but it generally underestimates teacher support or friend support. Again, we find little evidence that `ChatGPT-4o` performs better for social majorities.

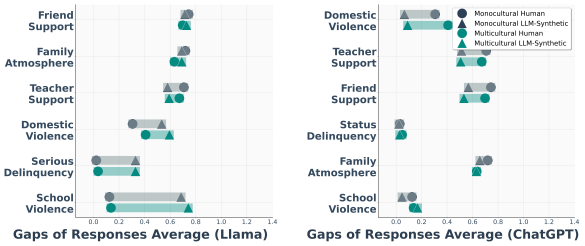

Figure 2: Gaps between the average of human responses and synthetic data. The gray lines indicate the gaps of monocultural group and the green lines indicate the gaps of multicultural adolescents. Left: `llama3:latest`; right: `ChatGPT-4o`.

Lastly, we compare the variances of the response distributions. Figures 3 and 4 present example distributions of synthetic and ground truth data for adolescents from multicultural and monocultural families. Consistent with prior research (Bisbee et al., 2024), the variance of the synthetic data distribution (yellow) is smaller than that of the original data (blue). Regardless of the accuracy of average predictions, the variance of synthetic data is consistently narrower than that of the ground truth. We do not find systematic patterns in how the magnitude of this variance reduction varies across question domains or LLMs.

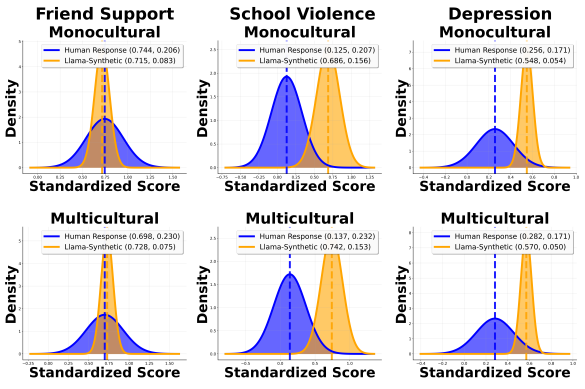

Figure 3: The distribution of scores generated by `llama` in three exemplary domains: friend support, school violence, and depression. The yellow color shows the results of synthetic data and the blue color shows those of the ground truth data. The first row includes the results of monocultural adolescents and the second row includes those of multicultural adolescents.

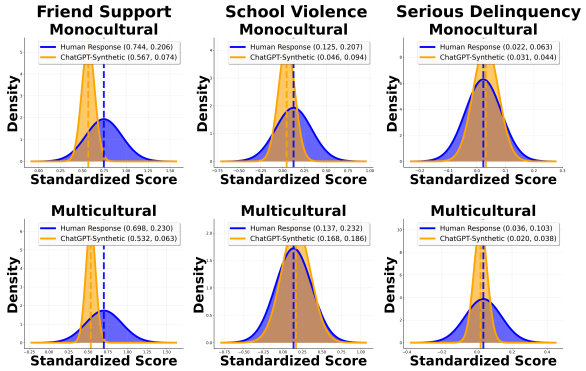

Figure 4: The distribution of scores generated by `ChatGPT-4o` in three exemplary domains: friend support, school violence, and serious delinquency. The yellow color shows the results of synthetic data and the blue color shows those of the ground truth data. The first row includes the results of monocultural adolescents and the second row includes those of multicultural adolescents.

4 Discussion and Conclusion

So far, we have examined how well synthetic data reproduce human survey responses for minority groups, specifically multicultural adolescents, compared to majority groups, namely monocultural adolescents, in the context of South Korean society. Using a unique survey that was administered equally to both groups, we generate synthetic data with `llama` and `ChatGPT-4o` and compare the results with the ground truth data from the original survey. Overall, we find limited differences in LLM performance between social majorities and minorities across question domains and models. Instead, the accuracy of mean predictions varies primarily by the type of questions asked. The variance of synthetic data distributions generally remains smaller than that of the ground truth data.

Our results highlight both the opportunities and limitations of using synthetic data. Regarding opportunities, it is notable that LLMs do not produce substantially less accurate predictions for social minorities than for majorities. Given the lower representation of minorities in digital data and their smaller population size, this result is somewhat unexpected. Another opportunity we identify is the unexpected accuracy of LLMs in predicting private and potentially socially stigmatized experiences. For instance, `ChatGPT-4o` predicts the average level of school violence experienced by multicultural adolescents very closely to the ground truth data.

Regarding limitations, we do not find systematic patterns indicating when and where LLMs perform well or poorly in generating synthetic data, which complicates their application in practical settings. While `llama` tends to overestimate the frequency of socially stigmatized behaviors, `ChatGPT-4o` produces more reliable results in these domains.

Our findings contribute to future research on synthetic data. Given the limited performance differences between social majorities and minorities, future work should focus on how synthetic data performance varies across question domains, particularly with respect to question sensitivity. In doing so, it is also important to recognize that ground truth survey data may themselves be biased due to social desirability concerns among respondents (Bispo Júnior, 2022; Latkin et al., 2017). To address this limitation, future research may benefit from incorporating in-depth interviews and qualitative methods.

4.1 Limitations

Our results are limited in their ability to predict synthetic data accuracy when relying solely on demographic information to generate personas. Prior studies of synthetic data in the social sciences often incorporate political attitudes or ideological positions when modeling public opinion (Argyle et al., 2023; Ma et al., 2025; Kim and Lee, 2023). This suggests that when the prediction target of synthetic data concerns personal experiences rather than political stances, the components used to construct personas may need to differ. In addition, we rely on a basic zero-shot prompting approach for model comparison, and our findings should therefore be interpreted with caution.

4.2 Ethical Considerations

While we conduct experiments using a comparative survey dataset on delinquent behaviors responded to by monocultural and multicultural adolescents (Jeon et al., 2016), the data contain no personal identifiers. We input survey questions in a manner that closely mirrors the structure of the original ground truth data, and we do not include any personal opinions in the prompts. Although our findings indicate that bias is more pronounced for socially stigmatized issues than for minority group status, we do not identify systematic mechanisms within LLMs that explain this pattern. This bias tendency appears to be specific to synthetic data generation in our setting and should not be generalized to LLM systems as a whole.

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.

Robert Bernasconi. 2000. The invisibility of racial minorities in the public realm of appearances. In *Phenomenology of the Political*, pages 169–187. Springer.

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.

Joy Adowaa Buolamwini. 2017. *Gender shades: intersectional phenotypic and demographic evaluation of face datasets and gender classifiers*. Ph.D. thesis, Massachusetts Institute of Technology.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

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.

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.

Valentin Hofmann, Pratyusha Ria Kalluri, Dan Jurafsky, and Sharese King. 2024. Dialect prejudice predicts ai decisions about people's character, employability, and criminality. *arXiv preprint arXiv:2403.00742*.

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*.

Seokjun Kim. 2015. A study on the effects of multicultural family adolescents' characteristics on their experiences of discrimination: Focusing on the 2012 national survey on multicultural families. *The Journal of Asiatic Studies*, 58(3):6–41.

Korean Statistical Information Service. 2015a. Population and housing census: Households and members by household type - eup/myeon/dong (years ending in 0, 5), si/gun/gu (other years). Korean Statistical Information Service (KOSIS). Retrieved November 1, 2025.

Korean Statistical Information Service. 2015b. Population and housing census: Multicultural households and members - si/gun/gu. Korean Statistical Information Service (KOSIS). Retrieved November 1, 2025.

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.

Inioluwa Deborah Raji, Timnit Gebru, Margaret Mitchell, Joy Buolamwini, Joonseok Lee, and Remi Denton. 2020. Saving face: Investigating the ethical concerns of facial recognition auditing. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 145–151.

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

Sheikh Shafayat, Eunsu Kim, Juhyun Oh, and Alice Oh. 2024. Multi-fact: Assessing multilingual llms' multi-regional knowledge using factscore. *arXiv preprint arXiv:2402.18045*.

Miri Song. 2020. Rethinking minority status and 'visibility'. *Comparative Migration Studies*, 8(1):5.

Seungjong Sun, Eungu Lee, Dongyan Nan, Xiangying Zhao, Wonbyung Lee, Bernard J Jansen, and Jang Hyun Kim. 2024. Random silicon sampling: Simulating human sub-population opinion using a large language model based on group-level demographic information. *arXiv preprint arXiv:2402.18144*.

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*.

Carol H Weiss. 1977. Survey researchers and minority communities. *Journal of Social Issues*, 33(4):20–35.

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 adolescent 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 Preprocessing

B.1 Missing Values

Missing values are rare in the sociodemographic variables used to construct personas, accounting for 0.001% of all responses. When missing values occur, they are replaced based on the context of the questions or left as non-responses. For example, missing values for nationality-related variables are replaced with "South Korea," as respondents are assumed to have lived in South Korea for a sufficiently long period and to have been educated in public schools. For question variables from which we draw responses, there are 254 missing values, corresponding to 0.002% of the total responses. For these cases, we estimate missing values using the K-nearest neighbors technique. We then aggregate both human and synthetic responses into 11 domains and normalize each domain score using dummy coding or reverse coding.

B.2 Scores Aggregation Methodology by Categories

B.2.1 Questions about Relationships

To measure overall relationship quality, composite scores for interpersonal relationship categories such as family atmosphere and friend support were calculated by averaging response values across all related items.

B.2.2 Questions about Stigmatic Experiences

For questions assessing whether respondents had any related experiences, such as violence, abuse, or delinquency, response scores were calculated

by summing the dummy-coded values across all related questions. Original responses ranged from "never experienced" (1) to "experienced seven or more times" (5), which were then recorded into a binary scale: "never experienced" (0) or "experienced at least once" (1).

B.2.3 Questions about Depression Levels

For the questions asking depression levels, they were measured by questions about self-esteem and negative thoughts based on responses to statements like "I am as important as other people" and "I have thoughts that I want to die" respectively. They were measured with scale from not at all (1) to strongly agree (5). Given that depression measurement focuses on symptom accumulation, depression scores were then calculated by summing all related items after parts of responses were reverse-coded so that higher scores consistently indicated more severe depressive symptoms.

C Theoretical Range of Survey Variables for Normalization

Table C.2: 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 Gaps between the Average of Human Responses and Synthetic Data

D.1 Llama-Synthetic Data

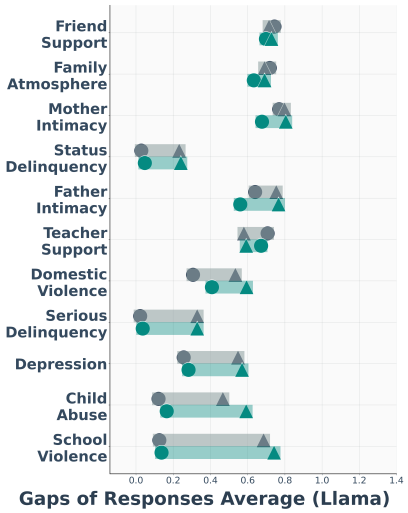

Figure D.5: Gaps between the average of human responses and Llama-synthetic data. The gray lines indicate the gaps of monocultural group and the green lines indicate the gaps of multicultural adolescents.

D.2 ChatGPT-Synthetic Data

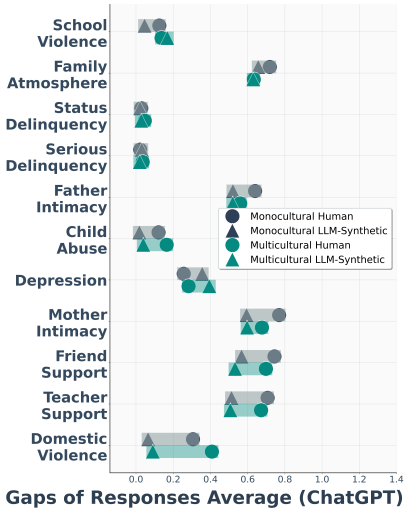

Figure D.6: Gaps between the average of human responses and ChatGPT-synthetic data. The gray lines indicate the gaps of monocultural group and the green lines indicate the gaps of multicultural adolescents.

E The Distribution Graphs of Scores Across the Domain Categories

E.1 Comparison of Human Responses and Llama-Synthetic Data

E.1.1 Non-Stigmatic Categories

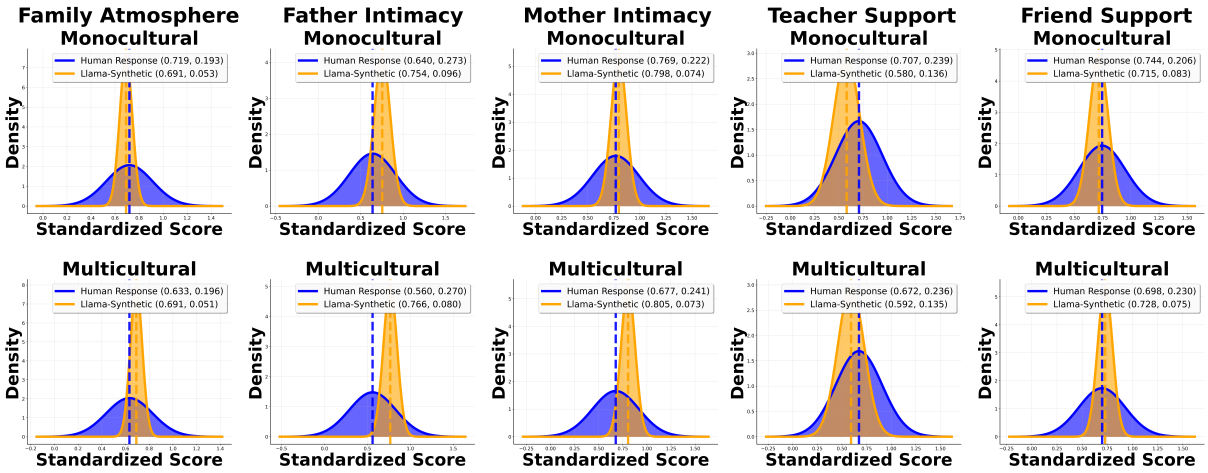

Figure E.7: Distribution of scores in socially non-stigmatic domains (Llama-synthetic data).

E.1.2 Stigmatic Categories

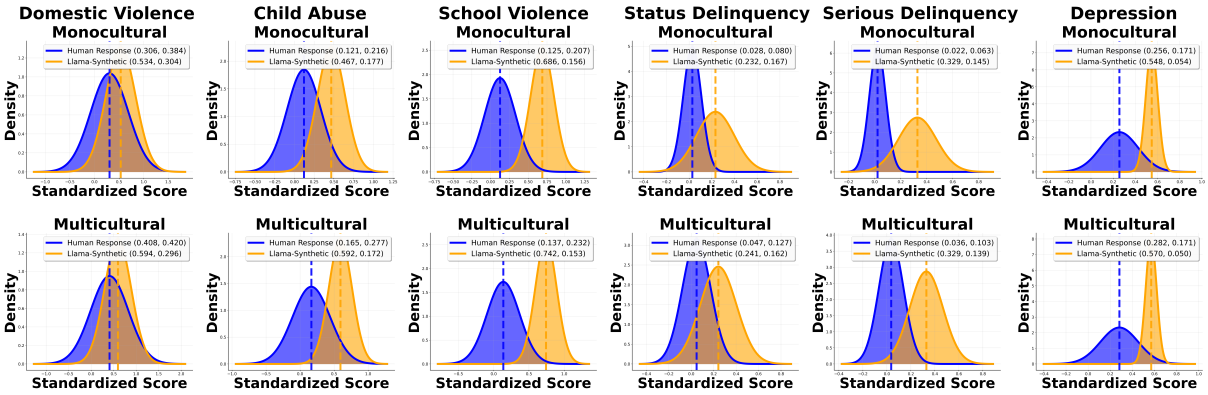

Figure E.8: Distribution of scores in socially stigmatic domains (Llama-synthetic data).

E.2 Comparison of Human Responses and ChatGPT-Synthetic Data

E.2.1 Non-Stigmatic Categories

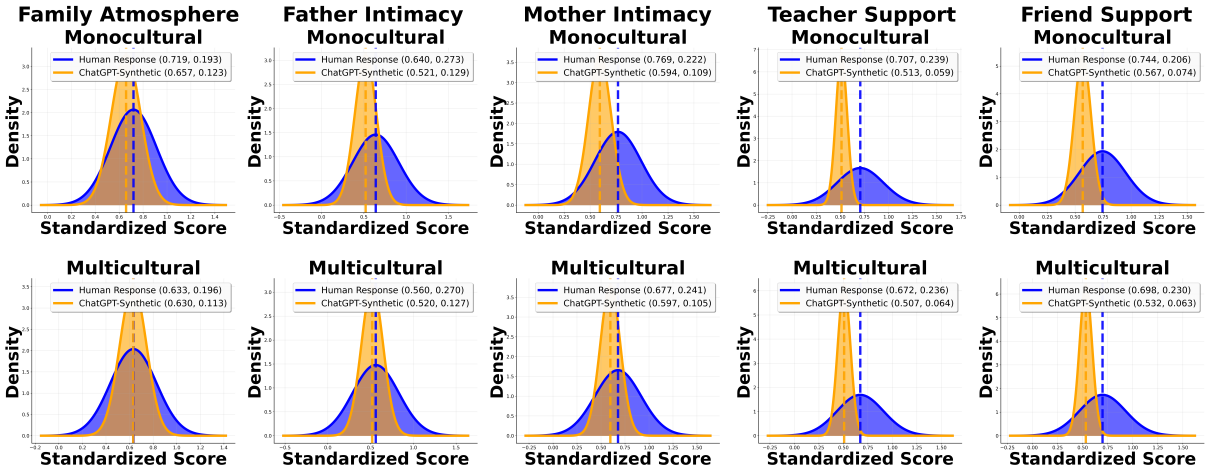

Figure E.9: Distribution of scores in socially non-stigmatic domains (ChatGPT-synthetic data).

E.2.2 Stigmatic Categories

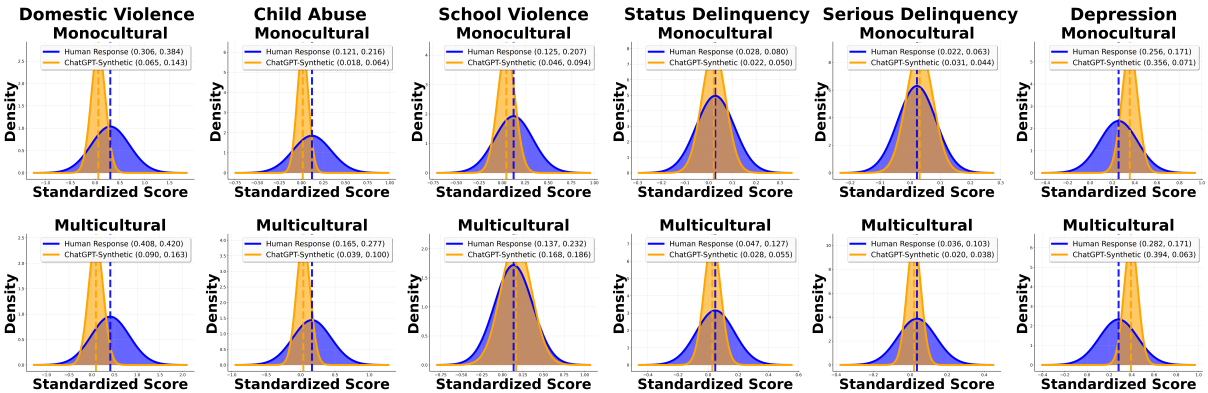

Figure E.10: Distribution of scores in socially stigmatic domains (ChatGPT-synthetic data).

F Selected Variables for Prompts

Used Prompts	Variable Category	Variables' name	Remarks	Examples
Persona	Gender	SQ1		
	Age	SQ2		
	School Grade	SQ3		
	Birth Country	F1_1		
	Residence Length in Korea	F1_1A		
	Parents' Marital Status	F5		
	Parents' Birth Country	B1_1, B1_2		
	Household Members	F6		
	Self-Perceived Household Economic Status	F12		
	Administrative District of Residence	DQ2	dummy variables	
Questionnaires	Family Atmosphere	B2_1 ∼ B2_9	Likert scale, calculated as average	*My family has conversation with each other a lot* (a) Not at all (b) Not really (c) Neither agree nor disagree (d) Mostly agree (e) Strongly agree
	Intimacy with Parents	B2_10 ∼ B2_15	Likert scale, calculated as average	*I want to be someone like my mother* (a) Not at all (b) Not really (c) Neither agree nor disagree (d) Mostly agree (e) Strongly agree
	Teacher Support	B12_3, B12_4	Likert scale, calculated as average	*Our teacher treat every adolescent equally.* (a) Not at all (b) Not really (c) Neither agree nor disagree (d) Mostly agree (e) Strongly agree
	Friend Support	B15_1, B15_2, B15_3	Likert scale, calculated as average	*When I asked for help, my friends help me.* (a) Not at all (b) Not really (c) Neither agree nor disagree (d) Mostly agree (e) Strongly agree
	Domestic Violence	B8_1, B8_2, B8_3	dummy variables, calculated as sum	*Do your parents hit each other when they are arguing?* (a) Never experienced (b) Experienced at least once
	School Violence	D1_1 ∼ D1_9	dummy variables, calculated as sum	*How often have schoolmates spread bad rumors about you in the past year?* (a) Never experienced (b) Experienced at least once
	Child Abuse	B9_1 ∼ B9_10	dummy variables, calculated as sum	*Have you listened swear words or harassment from your family (parents, grand parents, relatives)?* (a) Never experienced (b) Experienced at least once
	Status Delinquency	E1_1 ∼ E1_7, E1_9	dummy variables, calculated as sum	*Have you smoked in the past year?* (a) Never experienced (b) Experienced at least once
	Serious Delinquency	E1_8, E1_10 ∼ E1_19	dummy variables, calculated as sum	*Have you damaged school property in the past year?* (a) Never experienced (b) Experienced at least once
	Depression	A1_1 ∼ A1_15	dummy variables, calculated as sum	*Sometimes, I feel that I am not a good person.* (a) Not at all (b) Not really (c) Neither agree nor disagree (d) Mostly agree (e) Strongly agree

Table F.3: Selected Variables for Prompts and Question Examples

G Descriptive Statistics of Human Responses and Llama-Synthetic Data

G.1 Monocultural Adolescents

Table G.4: Descriptive Statistics of Human Responses and Llama-Synthetic Data by Monocultural Adolescents (n=800, m=800)

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

G.2 Multicultural Adolescents

Table G.5: Descriptive Statistics of Human Responses and Llama-Synthetic Data by Multicultural Adolescents (n=800, m=800)

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

H Descriptive Statistics of Human Responses and Llama-Synthetic Data (Normalized)

H.1 Monocultural Adolescents

Table H.6: Descriptive Statistics of Human Responses and Llama-Synthetic Data by Monocultural Adolescents (Normalized) (n=800, m=800)

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

H.2 Multicultural Adolescents

Table H.7: Descriptive Statistics of Human Responses and Llama-Synthetic Data by Multicultural Adolescents (Normalized) (n=800, m=800)

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

I Descriptive Statistics of Human Responses and ChatGPT-Synthetic Data

I.1 Monocultural Adolescents

Table I.8: Descriptive Statistics of Human Responses and ChatGPT-Synthetic Data by Monocultural Adolescents (n=800, m=800)

Variable	Data Type	Mean	Standard Deviation	SE
Family Atmosphere	Human	3.88	0.77	0.03
	ChatGPT-Synthetic	3.63	0.49	0.02
Father Intimacy	Human	3.56	1.09	0.04
	ChatGPT-Synthetic	3.08	0.51	0.02
Mother Intimacy	Human	4.07	0.89	0.03
	ChatGPT-Synthetic	3.37	0.43	0.02
Teacher Support	Human	3.83	0.96	0.03
	ChatGPT-Synthetic	3.05	0.24	0.01
Friend Support	Human	3.98	0.83	0.03
	ChatGPT-Synthetic	3.27	0.30	0.01
Domestic Violence	Human	0.92	1.15	0.04
	ChatGPT-Synthetic	0.19	0.43	0.02
Child Abuse	Human	1.21	2.16	0.08
	ChatGPT-Synthetic	0.18	0.64	0.02
School Violence	Human	1.13	1.86	0.07
	ChatGPT-Synthetic	0.41	0.85	0.03
Status Delinquency	Human	0.22	0.64	0.02
	ChatGPT-Synthetic	0.17	0.40	0.01
Serious Delinquency	Human	0.24	0.70	0.02
	ChatGPT-Synthetic	0.34	0.49	0.02
Depression	Human	30.35	10.24	0.36
	ChatGPT-Synthetic	36.35	4.27	0.15

I.2 Multicultural Adolescents

Table I.9: Descriptive Statistics of Human Responses and ChatGPT-Synthetic Data by Multicultural Adolescents (n=800, m=800)

Variable	Data Type	Mean	Standard Deviation	SE
Family Atmosphere	Human	3.53	0.78	0.03
	ChatGPT-Synthetic	3.52	0.45	0.02
Father Intimacy	Human	3.24	1.08	0.04
	ChatGPT-Synthetic	3.08	0.51	0.01
Mother Intimacy	Human	3.71	0.96	0.03
	ChatGPT-Synthetic	3.39	0.42	0.01
Teacher Support	Human	3.69	0.94	0.03
	ChatGPT-Synthetic	3.03	0.26	0.01
Friend Support	Human	3.79	0.92	0.03
	ChatGPT-Synthetic	3.13	0.25	0.01
Domestic Violence	Human	1.23	1.26	0.04
	ChatGPT-Synthetic	0.27	0.49	0.02
Child Abuse	Human	1.65	2.77	0.10
	ChatGPT-Synthetic	0.39	1.00	0.04
School Violence	Human	1.23	2.09	0.07
	ChatGPT-Synthetic	1.51	1.67	0.06
Status Delinquency	Human	0.38	1.01	0.04
	ChatGPT-Synthetic	0.22	0.44	0.02
Serious Delinquency	Human	0.40	1.13	0.04
	ChatGPT-Synthetic	0.22	0.42	0.01
Depression	Human	31.93	10.27	0.36
	ChatGPT-Synthetic	38.62	3.75	0.13

J Descriptive Statistics of Human Responses and ChatGPT-Synthetic Data (Normalized)

J.1 Monocultural Adolescents

Table J.10: Descriptive Statistics of Human Responses and ChatGPT-Synthetic Data by Monocultural Adolescents (Normalized) (n=800, m=800)

Variable	Data Type	Mean	Standard Deviation	SE
Family Atmosphere	Human	0.72	0.19	0.01
	ChatGPT-Synthetic	0.66	0.12	0.00
Father Intimacy	Human	0.64	0.27	0.01
	ChatGPT-Synthetic	0.52	0.13	0.00
Mother Intimacy	Human	0.77	0.22	0.01
	ChatGPT-Synthetic	0.59	0.11	0.00
Teacher Support	Human	0.71	0.24	0.01
	ChatGPT-Synthetic	0.51	0.06	0.00
Friend Support	Human	0.74	0.21	0.01
	ChatGPT-Synthetic	0.57	0.07	0.00
Domestic Violence	Human	0.31	0.38	0.01
	ChatGPT-Synthetic	0.06	0.14	0.00
Child Abuse	Human	0.12	0.22	0.01
	ChatGPT-Synthetic	0.02	0.06	0.00
School Violence	Human	0.13	0.21	0.01
	ChatGPT-Synthetic	0.05	0.09	0.00
Status Delinquency	Human	0.03	0.08	0.00
	ChatGPT-Synthetic	0.02	0.05	0.00
Serious Delinquency	Human	0.02	0.06	0.00
	ChatGPT-Synthetic	0.03	0.04	0.00
Depression	Human	0.26	0.17	0.01
	ChatGPT-Synthetic	0.36	0.07	0.00

J.2 Multicultural Adolescents

Table J.11: Descriptive Statistics of Human Responses and ChatGPT-Synthetic Data by Multicultural Adolescents (Normalized) (n=800, m=800)

Variable	Data Type	Mean	Standard Deviation	SE
Family Atmosphere	Human	0.63	0.20	0.01
	ChatGPT-Synthetic	0.63	0.11	0.00
Father Intimacy	Human	0.56	0.27	0.01
	ChatGPT-Synthetic	0.52	0.13	0.00
Mother Intimacy	Human	0.68	0.24	0.01
	ChatGPT-Synthetic	0.60	0.11	0.00
Teacher Support	Human	0.67	0.24	0.01
	ChatGPT-Synthetic	0.51	0.06	0.00
Friend Support	Human	0.70	0.23	0.01
	ChatGPT-Synthetic	0.53	0.06	0.00
Domestic Violence	Human	0.41	0.42	0.01
	ChatGPT-Synthetic	0.09	0.16	0.01
Child Abuse	Human	0.17	0.28	0.01
	ChatGPT-Synthetic	0.04	0.10	0.00
School Violence	Human	0.14	0.23	0.01
	ChatGPT-Synthetic	0.17	0.19	0.01
Status Delinquency	Human	0.05	0.13	0.00
	ChatGPT-Synthetic	0.03	0.06	0.00
Serious Delinquency	Human	0.04	0.10	0.00
	ChatGPT-Synthetic	0.02	0.04	0.00
Depression	Human	0.28	0.17	0.01
	ChatGPT-Synthetic	0.39	0.06	0.00

K Data and Implementation Details

K.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.

K.2 LLM API Specifications

We used `llama3:latest` (Meta AI) and `Chatgpt-4o` (Open AI) for all synthetic data generation. The model was accessed via `Ollama` and `OpenAI` with default parameters. All other generation parameters (top_p, top_k, max_tokens, etc.) used the default Ollama and OpenAI settings. Total API calls: approximately 132,200 (800 samples per group $\times$ 2 groups $\times$ 77 questions) for each model. The generation was conducted on a local server with GPU acceleration.

K.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)
- openai (v1.87.0)
- transformers (v4.54.1)
- tqdm (v4.67.1)

K.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.