Evaluating Bias in LLM-Synthetic Data across Majority and Minority Groups on General and Stigmatized Topics: The Case of Multicultural and Monocultural Adolescents in South Korea

Jiyun Beak **Myokyung Han** **Lanu Kim***
KAIST
{bamjy99, mk_han, lanukim}@kaist.ac.kr

Abstract

Large Language Models (LLMs) are increasingly used to replicate human responses in survey data, yet it remains unclear whether they perform equally well for social minorities and majorities. Using a unique survey conducted in South Korea that provided 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 `gpt-4o`, to answer 77 survey questions across 11 domains. We find marginal differences in LLM performance between minority and majority groups. Instead, replication performance varies by whether prompts contain socially stigmatized words and by which model we use. These findings highlight both the potential and the limitations of using LLMs to generate synthetic survey data for the social majorities and minorities.

1 Introduction

Because traditional survey methods require substantial time, effort, and financial resources, 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). They have generated artificial data by designing hypothetical personas with demographic information and asking LLMs to answer survey questionnaires as if they were those personas (Kim and Lee, 2023; Argyle et al., 2023; Bail, 2024; Grossmann et al., 2023; Ma et al., 2025). Because these responses are synthesized from patterns learned from existing training data, they are commonly referred to as synthetic data. Even when provided with solely limited demographic information such as age, gender, and place of residence, synthetic data can approximate the average opinions of the general public relatively well (Argyle et al., 2023; Bisbee et al., 2024; Sun et al., 2024).

While synthetic data has become a solution for complementing scarce and imbalanced datasets (Moe et al., 2025; Singh, 2015; Poucin et al., 2021), research examining whether LLMs generate synthetic data equally well for the social majorities and minorities is still scarce. LLMs' training data for social minorities is sparse because it is often harder to collect due to their small numerical representations (Herzing et al., 2019) and lower levels of social visibilities or participation (George et al., 2014; Song, 2020; Bernasconi, 2000). Algorithms tend to have exhibited lower accuracy and reliability for minority groups (Shafayat et al., 2024; Buolamwini, 2017; Raji et al., 2020; Hofmann et al., 2024; Fazelpour and Danks, 2021; Wang et al., 2024). Therefore, LLMs have replicated opinions unevenly with better performance for dominant and higher-status demographic groups (Qu and Wang, 2024; Wang et al., 2025). As a result, it is important to assess whether LLM performance is equitable for social majorities and minorities in generating synthetic data across various questionnaires.

Accordingly, we examine 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, where 800 respondents were sampled each from social minority and majority groups, specifically teenagers from multicultural families who are first- or second-generation immigrants, and those from monocultural families whose parents are both Korean. Treating this dataset as ground truth, we generate synthetic data using the survey's demographic information to construct hypothetical personas and evaluate the replication performance of the synthetic responses. We find little difference in LLM performance between social minorities

*Corresponding author.

Proceedings of the Fifth Workshop on NLP for Positive Impact (NLP4PI), pages 161–174
October 28, 2026 ©2026 Association for Computational Linguistics

and majorities; however, performance varies by the types of questions asked.

2 Data and Methodology

2.1 Data

Here, we define a social minority group as adolescents from multicultural families in South Korea (Republic of Korea, 2020). When selecting target groups, it is important to consider not only population size but also public perception, particularly whether a group has been subject to marginalization in society. We choose multicultural families in South Korea as target group given small proportion (1.53% of total population in census) (Korean Statistical Information Service, 2015b,a) and discrimination evidences against them (Kim, 2015).

In our study, a key requirement for ground truth data is to have the comparable survey results of the minorities and majorities. We select a survey by the Korean Institute of Criminology and Justice in 2012 which administered identical questions to monocultural and multicultural adolescents (Jeon et al., 2016). In addition to questions on delinquent behaviors, the survey includes questions on sociodemographic characteristics such as family background, and school satisfaction. The 1,600 respondents were aged 11 to 19 students attending elementary, middle, and high schools in South Korea, consisting of 800 students from monocultural families and 800 from multicultural families.

2.2 Analytic Strategy

Following previous approaches to creating synthetic data (Argyle et al., 2023), we design hypothetical personas with 11 demographic variables from the original survey and ask LLMs to answer survey questionnaires from the perspective of the given personas. We create 1,600 personas, including 800 adolescents from multicultural families and 800 from monocultural families, by directly mapping each demographic attribute identically from the ground truth data to an individual persona.

Then, we ask LLMs to answer 77 survey questions, which are categorized into 11 domains such as domestic violence, child abuse, and school violence. Table F.3 indicates the full lists. Answers are measured either on continuous Likert scales or binary scales. We test two LLMs, specifically `llama3:latest` (Dubey et al., 2024) and `gpt-4o` (OpenAI, 2024), using their default parameter settings. We pose each question to every persona, generating 246,400 responses in total (800 individuals $\times$ 2 groups $\times$ 77 questions $\times$ 2 models).

2.3 Prompt Design and Preprocessing

The persona prompts are composed of the 11 sociodemographic variables such as gender, age, and parents' nationality. The answer options in the question prompts are mainly identical to those in the ground truth survey. All prompts are provided in Korean, as same as the original surveys. We also include the date on which the original survey was conducted.

Following the prompt design guidelines for synthetic data (Argyle et al., 2023; Bail, 2024; Bisbee et al., 2024), we finalize the prompts as shown in Figure 1, which separated into system, contextual, and role prompt sections (Boonstra, 2025). We instruct the model to respond from the given persona's perspective to mitigate safety alignment refusals (Shanahan et al., 2023) and answer homogenization across personas.

```
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 (Translated into English)

The ground truth data, referred to as the human respondent data in this study, contain 16 missing values, 0.001% of the persona variables and 254 missing values, 0.002% of the questionnaire variables. For persona variables, missing values were replaced based on the contexts of each variable. The K-nearest neighbors (KNN) technique was mostly used to refill the questionnaire variables.

We aggregate response values as results of scores by domain category in an identical way for both the ground truth and synthetic data. For example, questionnaire items asking whether respondents have experienced stigmatizing events such as violence, abuse, or delinquency were converted to dummy-coded values and aggregated by `SUM`, because it is more important to capture if respondents have ever experienced such traumatic events in their life-

time. The score calculation methodologies differ by question domains as shown in Appendix B.

2.4 Evaluation Metrics

To compare the continuous respondent scores across groups, we use Cohen's *d*, which measures the practical significance of differences between two groups (Cohen, 1988). We also use distribution graphs between the synthetic data and the ground truth, particularly focusing on their mean and variance values. Mean differences between human responses and synthetic data indicates how closely LLMs replicate average human responses, whereas variance differences reflect how well LLMs capture the diversity of human experiences and opinions. Although mean and variability are simple metrics, many commonly used comparison statistics are derived from them, such as *t*-statistics, Cohen's *d*, and the intraclass correlation coefficient (ICC) (Argyle et al., 2023). Thus, meaningful differences in these two statistic values often extend to the effectiveness of other derived metrics, though this does not guarantee equivalence across all measures.

3 Results

Table 1 presents Cohen's *d* values comparing synthetic and human responses for monocultural and multicultural adolescents across 11 domains and two LLMs. Overall, we find no clear pattern indicating that either group is predicted more accurately than the other. In most domains, the differences in Cohen's *d* between the two groups are small.

However, some domain-specific differences appear. For `llama3:latest`, synthetic responses are more similar to human responses for monocultural adolescents than for multicultural adolescents in three domains: family atmosphere, father intimacy, and mother intimacy. In teacher support, however, this pattern reverses, with `llama3:latest` performing better for multicultural adolescents.

For `gpt-4o`, a similar reversal is observed in family atmosphere and father intimacy, where synthetic responses more closely match those of multicultural adolescents than those of monocultural adolescents. Because the direction of the group-level advantage is inconsistent across models and domains, we do not find evidence that either LLM systematically favors the social majority or the social minority group.

Although group-level differences are modest, the accuracy of synthetic data varies considerably by

		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

Table 1: Cohen's *d* Results by Groups and Categories

question domain, and this pattern differs between the two LLMs. For `llama3:latest`, the gap between human and synthetic responses is substantially larger in socially stigmatized domains such as violence, delinquency, and depression than in general domains. The left panel of Figure 2 illustrates the three domains with the smallest gaps and the three domains with the largest gaps, showing that `llama3:latest` tends to overestimate the frequency with which respondents experience these stigmatized behaviors. Notably, this domain-level pattern holds consistently for both monocultural and multicultural adolescents.

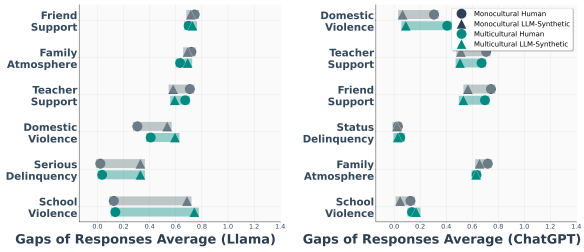


Figure 2: The six domains with the largest and smallest gaps between average human and synthetic responses, generated by `llama3:latest` (left) and `gpt-4o` (right). Gray lines indicate monocultural adolescents; green lines indicate multicultural adolescents. Full results are shown in Appendix D.

The pattern for `gpt-4o` differs. While `gpt-4o` generally produces more accurate predictions of the average response than `llama3:latest`, its accuracy still varies by domain. It performs relatively well in some stigmatized domains, such as status and serious delinquency, but underestimates responses in general-topic domains such as teacher support and friend support. As with

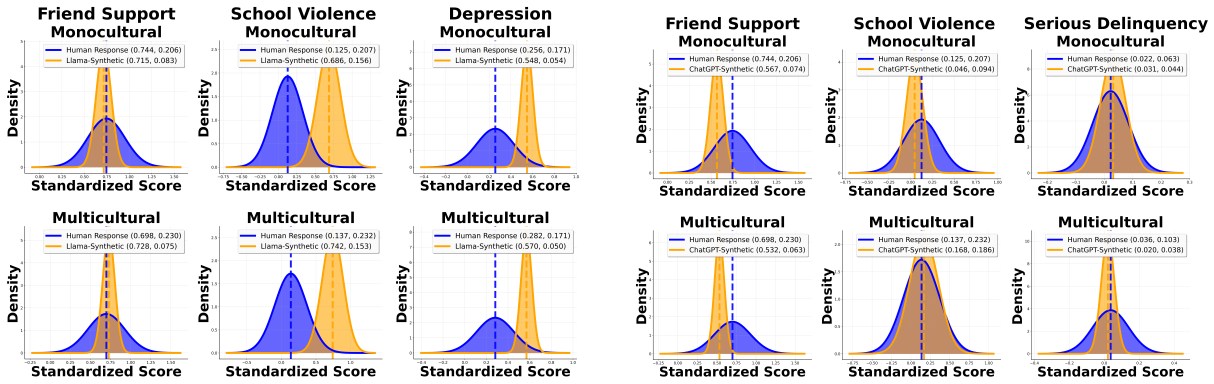


Figure 3: Distribution graphs, generated by `llama3:latest` (left) and `gpt-4o` (right), showing aggregated scores for three domains (friend support, school violence, and depression/serious delinquency). Synthetic data appears in orange lines; ground truth data is in blue lines. Monocultural adolescents appear at the top, multicultural at the bottom. Full results are in Appendix E.

`llama3:latest`, we find little evidence that this domain-level accuracy differs systematically between the social majority and minority groups.

Lastly, we compare the variance of the response distributions between synthetic and ground-truth data. Figure 3 presents example distributions for adolescents from multicultural and monocultural families. Consistent with prior research (Bisbee et al., 2024), the variance of the synthetic data distribution (orange) is consistently narrower than that of the ground truth (blue), regardless of how accurately the average response is predicted. This reduction in variance does not vary systematically by question domain or by LLM.

4 Discussion and Conclusion

So far, we have examined how well synthetic data reproduces human survey responses for minority groups compared to majority groups, namely multicultural adolescents compared to monocultural adolescents in South Korea. Using a unique survey with the same sample size for both target groups, we generate synthetic data with `llama3:latest` and `gpt-4o` and evaluate how closely it matches the ground truth.

Overall, we find marginal differences in LLM performance between majority and minority groups; rather, the accuracy of predicted means varies mainly by question domain. Meanwhile, the variances of synthetic data distributions generally remain much smaller than those of the original data, consistent with previous research (Argyle et al., 2023).

Our results highlight that LLMs do not produce substantially less accurate predictions for social minorities than for majorities—an unexpected finding given the lower representation of minorities in digital data and their smaller population size. It is also surprising that the accuracy of synthetic data on stigmatized questions depends on the specific model: while `llama3:latest` tends to overestimate the frequency of socially stigmatized behaviors, `gpt-4o` predicts more reliably in these domains, closely matching the average level of school violence among both multicultural and monocultural adolescents.

Given these mixed patterns, future work should examine how synthetic data performance varies across question domains, particularly with respect to question sensitivity, while recognizing that ground truth survey data may itself be biased by social desirability concerns (Bispo Júnior, 2022; Latkin et al., 2017). Incorporating in-depth interviews and qualitative methods may help address this.

We construct personas based solely on demographic information, whereas prior studies of synthetic data often incorporate political attitudes or ideological positions when modeling public opinion (Argyle et al., 2023; Ma et al., 2025; Kim and Lee, 2023). This raises the possibility that persona components may need to differ when synthetic data does not concern political stances. Also, we rely on basic zero-shot prompting, so results should be interpreted with caution.

This study uses a survey without personal identifiers (Jeon et al., 2016), and we include no personal opinions in any prompt. The marginal group differences observed here 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.

Lee Boonstra. 2025. Prompt engineering. *Google, https://www. kaggle. com/whitepaper-prompt-engineering*.

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. 1988. *Statistical Power Analysis for the Behavioral Sciences*, 2nd edition. Lawrence Erlbaum Associates, Hillsdale, NJ.

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.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

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.

Lwin Moe, Uyen Trang Nguyen, and Boi Trinh Luu. 2025. Mitigating class imbalance in fact-checking datasets through llm-based synthetic data generation. In *Proceedings of the 4th ACM International Workshop on Multimedia AI against Disinformation*, pages 73–80.

OpenAI. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Florentin Poucin, Andrea Kraus, and Martin Simon. 2021. Boosting instance segmentation with synthetic data: A study to overcome the limits of real world data sets. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 945–953.

Yao Qu and Jue Wang. 2024. Performance and biases of large language models in public opinion simulation. *Humanities and Social Sciences Communications*, 11(1):1–13.

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*.

Murray Shanahan, Kyle McDonell, and Laria Reynolds. 2023. Role play with large language models. *Nature*, 623(7987):493–498.

Hardial Singh. 2015. Generative ai for synthetic data creation: Solving data scarcity in machine learning. *Available at SSRN 5267914*.

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*.

Jia Wang, Ziyu Zhao, Tingjuntao Ni, and Zhongyu Wei. 2025. Sociobench: Modeling human behavior in sociological surveys with large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 26268–26300.

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*.

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 an adolescent in the 3rd grade of middle school 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 standardized 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 Stigmatized 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

Depression levels were measured using questions about self-esteem and negative thoughts, based on responses to statements such as "I am as important as other people" and "I have thoughts that I want to die," respectively. They were measured with a scale ranging 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 Standardization

Table C.2: Theoretical Range of Survey Variables for Standardization

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 standardization purposes.

D Gaps between the Average of Human Responses and Synthetic Data

D.1 Llama-Synthetic Data

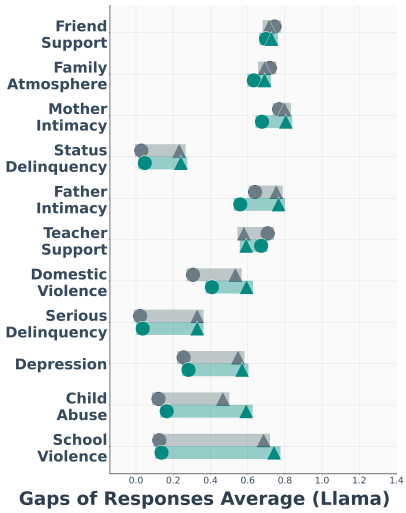


Figure D.4: Gaps between the average of human responses and Llama-synthetic data. The gray lines indicate the gaps of monocultural group and the green lines indicate the gaps of multicultural adolescents.

D.2 ChatGPT-Synthetic Data

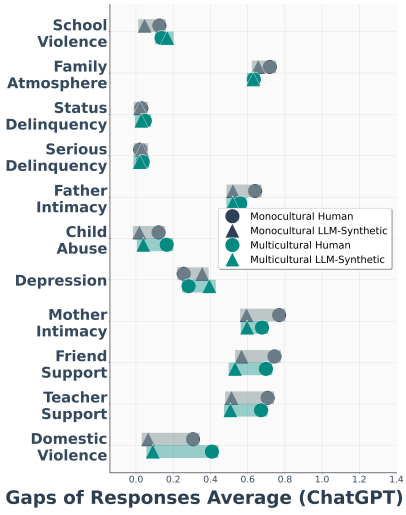


Figure D.5: Gaps between the average of human responses and ChatGPT-synthetic data. The gray lines indicate the gaps of monocultural group and the green lines indicate the gaps of multicultural adolescents.

E The Distribution Graphs of Scores Across the Domain Categories

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

E.1.1 Non-stigmatized Categories

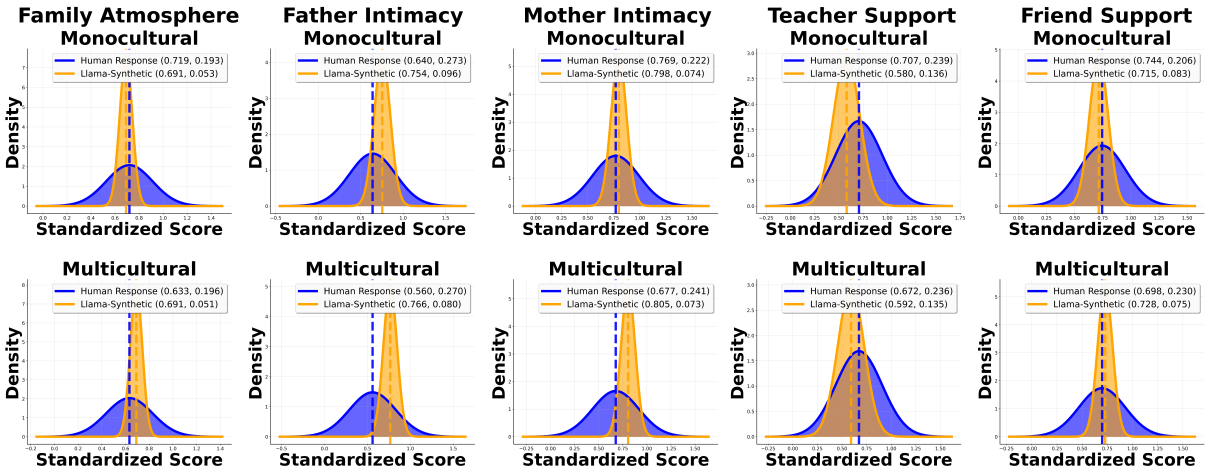


Figure E.6: Distribution of scores in socially non-stigmatized domains (Llama-synthetic data).

E.1.2 stigmatized Categories

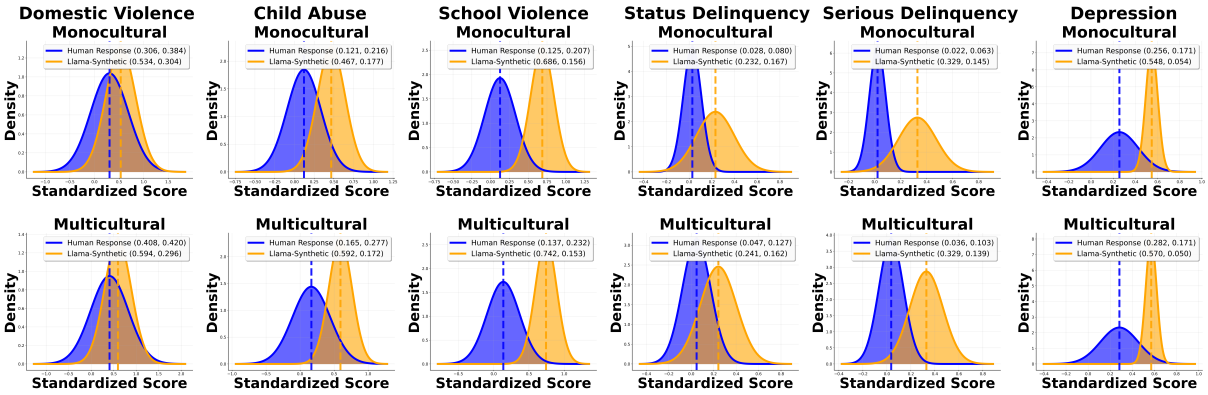


Figure E.7: Distribution of scores in socially stigmatized domains (Llama-synthetic data).

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

E.2.1 Non-stigmatized Categories

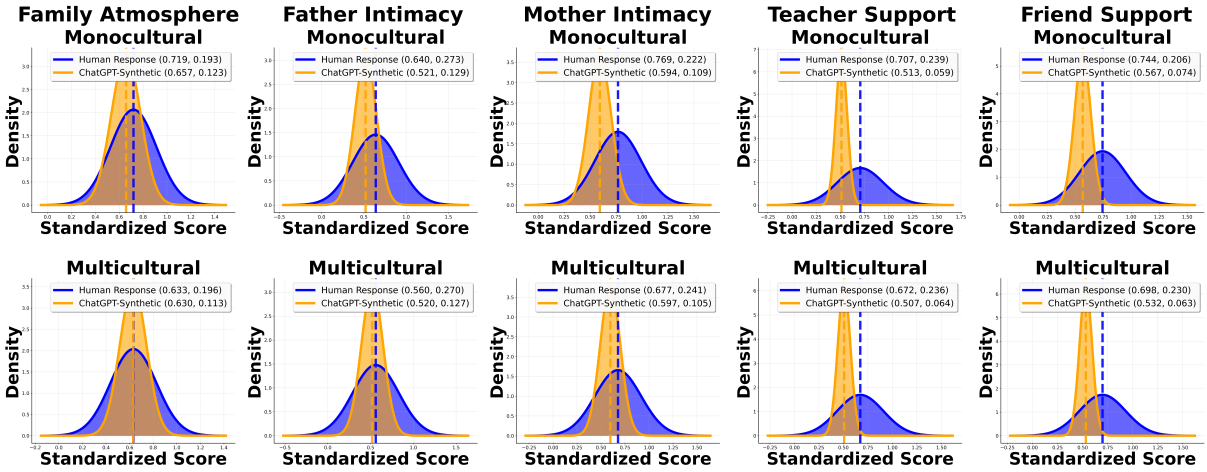


Figure E.8: Distribution of scores in socially non-stigmatized domains (ChatGPT-synthetic data).

E.2.2 stigmatized Categories

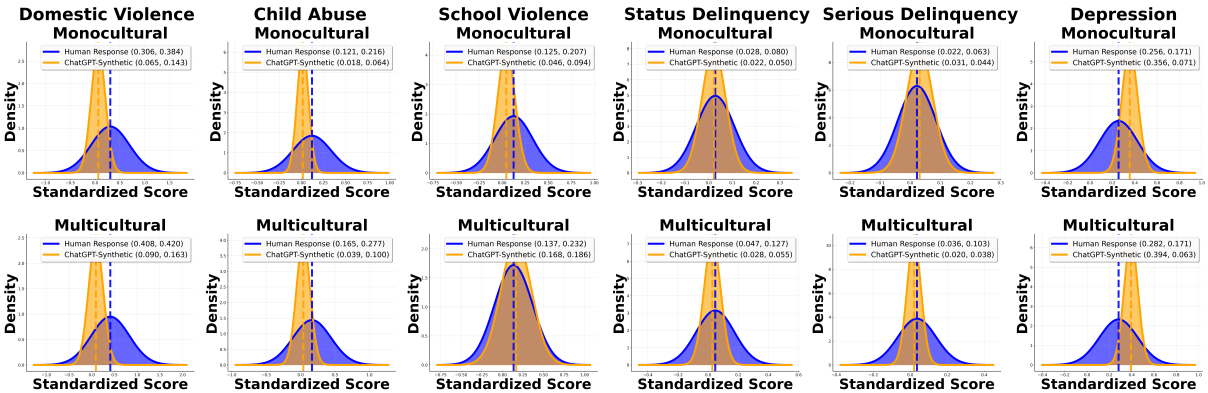


Figure E.9: Distribution of scores in socially stigmatized domains (ChatGPT-synthetic data).

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 conversations 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 treats every student 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 ask 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 heard swear words or experienced harassment from your family (parents, grandparents, 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 for 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 for 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 (Standardized)

H.1 Monocultural Adolescents

Table H.6: Descriptive Statistics of Human Responses and Llama-Synthetic Data for Monocultural Adolescents (Standardized) (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 for Multicultural Adolescents (Standardized) (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 for 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 for 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 (Standardized)

J.1 Monocultural Adolescents

Table J.10: Descriptive Statistics of Human Responses and ChatGPT-Synthetic Data for Monocultural Adolescents (Standardized) (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 for Multicultural Adolescents (Standardized) (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. The study was 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 `gpt-4o` (OpenAI) for all synthetic data generation. The models were 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 123,200 (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 (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.