International Journal of Internet, Broadcasting and Communication Vol.18 No.1 407-416 (2026)
http://dx.doi.org/10.7236/IJIBC.2026.18.1.407

IJIBC 26-1-35

Variable-Dependent Cultural Alignment in LLM Survey Simulation: Comparing Indigenous and Global Models on Korean Cultural Variables

Seog-Min Lee

Professor, Dept. of Public Policy and Big Data Convergence, Hanshin University, Korea
Newmind68@hs.ac.kr

Abstract

Large language models (LLMs) are increasingly used to simulate survey responses ("silicon sampling"), yet whether locally trained ("indigenous") models better reproduce local public opinion remains untested. We compare a global model (GPT-5.2) with a Korean indigenous model (CLOVA HCX-007) using marginal distributions from the Korean General Social Survey (KGSS) 2023 as benchmarks. We evaluate six culturally salient variables spanning institutional trust, national pride, inter-Korean attitudes, and political orientation, quantifying alignment via Jensen–Shannon (JS) divergence and Kolmogorov–Smirnov (KS) tests. In baseline persona-prompting, we find that GPT-5.2 exhibits heterogeneous alignment (average JS=0.123) with extreme response concentration on unification attitudes (UNIFI). We conduct multi-seed robustness validation (5 independent runs with different persona sets) for the cultural-context comparison. Our results show that CLOVA yields mixed results (average JS=0.096; 2/6 variables improved): gains on UNIFI (+54.1%) and CONLEGIS (+35.8%), but comparable or worse performance on four other variables. Neither model achieves distributional equivalence with KGSS benchmarks (KS tests: all p<0.05). Our 5-seed validation indicates that indigenous model advantages are limited to specific domains (legislative trust, unification attitudes) rather than uniform across culturally-sensitive variables. We recommend variable-level pilot evaluation before substantive use of silicon sampling.

***Keywords**: Silicon Sampling, Large Language Models, Korean General Social Survey, Indigenous LLM, CLOVA, Cultural Context, Cross-Cultural Comparison*

1. Introduction

1.1 Research Background: Proliferation of Silicon Sampling Methodology Claims

Survey research using Large Language Models (LLMs) for survey simulation—often called “Silicon Sampling”—has gained traction as a way to bypass the cost, time, and access barriers of traditional social surveys. Argyle et al. [2] reported that GPT-3 could simulate U.S. GSS responses with reasonable accuracy, prompting a wave of methodological proposals.

Yet these claims have largely been validated in isolation. No integrated comparative evaluation exists. Equally important, the literature focuses almost exclusively on English-speaking populations(particularly U.S.

Manuscript Received: January. 5, 2026 / Revised: January. 16, 2026 / Accepted: February. 10, 2026
Corresponding Author: newmind68@hs.ac.kr
Tel: 010-8379-3518
Professor, Department of Public Policy and Big Data Convergence, Hanshin University, Korea

data), so how well these methods transfer to non-Western settings remains an open question.

1.2 Study Focus and Differentiation from Prior Work

This paper evaluates silicon sampling using Korean public-opinion items from KGSS 2023, a setting where cultural context is expected to influence model performance. Two questions are addressed: whether demographic-persona prompting produces marginal distributions close to survey benchmarks, and whether an indigenous model trained on Korean data exhibits closer alignment than a global model. Alignment is treated as a distributional property, and both divergence metrics and hypothesis-test results are reported with caveats for ordinal/discrete outcomes.This study departs from prior work in several ways. Where existing studies focus on U.S. and Western contexts [2,8], we evaluate Korean cultural variables, providing systematic validation in a non-Western survey context. Where prior work evaluates single models in isolation [4], we compare two models—one indigenous, one global—at the variable level under identical experimental conditions. And where aggregate accuracy measures are the norm, we prioritize distributional alignment metrics (JS divergence) and formal statistical tests (KS, chi-square), addressing concerns that aggregate metrics may mask variable-level distributional mismatches [8].

1.3 Research Questions

- **RQ1 (Baseline validity)**: Under a standard persona-based prompting protocol, how closely do LLM-generated response distributions match KGSS 2023 marginal distributions?
- **RQ2 (Cultural context)**: Does a Korean indigenous model (CLOVA HCX-007) yield closer alignment than a global model (GPT-5.2), and is any advantage variable-dependent?

2. Related Work

2.1 LLM-Based Survey Simulation

Recent work has explored using LLMs to simulate survey responses, with mixed results. Argyle et al. [2] showed that GPT-3, conditioned on demographic “backstories,” can reproduce U.S. survey patterns with reasonable fidelity. Aher et al. [1] replicated classic behavioral experiments with LLMs. Dillon et al. [4] proposed using LLM-generated responses as pilot data for survey design. Ornstein et al. [8] offered methodological guidelines—what they call “training stochastic parrots”—to improve simulation accuracy.

These optimistic findings face significant pushback. Santurkar et al. [15] built OpinionQA, a benchmark covering 60 U.S. demographic groups, and found persistent misalignment even with explicit demographic steering. Durmus et al. [21] extend this work globally, measuring the representation of subjective opinions from 36 countries across multiple LLMs and finding systematic underrepresentation of non-Western viewpoints. Bisbee et al. [16] raised more pointed concerns: although ChatGPT average scores correlate with survey benchmarks, the synthetic distributions show reduced variance and unstable regression coefficients. This calls into question the reliability of “silicon samples” for statistical inference.

What unites these studies—and limits them—is their near-exclusive reliance on U.S. English-language data. Whether the same methods hold up in non-Western cultural contexts, and whether model origin matters, has received little attention.

2.2 Cultural Bias in Language Models

Several studies have documented cultural biases in LLMs. Cao et al. [3] found that GPT models systematically favor Western cultural norms in value judgments, and Naous et al. [7] documented Arabic cultural misalignment in multilingual models. Hartmann et al. [23] showed that ChatGPT exhibits a pro-environmental, left-libertarian political orientation across multiple language contexts, raising concerns about systematic ideological skew.

Yet "Western bias" understates the pattern's complexity. Tao et al. [17] evaluated LLM outputs across 107 countries using the World Values Survey. GPT models consistently tilted toward self-expression values typical of English-speaking and Protestant European cultures. "Cultural prompting" helped in most countries, but 19–29% of countries saw no improvement—or even worse alignment. Li et al. [18] took a different approach with CultureLLM, showing that culture-specific fine-tuning on World Values Survey data can improve alignment, though it requires language-specific adaptations each time.

Whether indigenous LLMs—those trained on culturally-specific corpora—actually mitigate these biases when used for survey simulation is still an open question.

2.3 Indigenous LLMs and Cultural Contextualization

Indigenous LLMs—HyperCLOVA in Korea being a prominent example—provide a direct way to test whether culturally-grounded training data translates into better survey simulation. These models incorporate local language data, cultural knowledge, and region-specific fine-tuning [5]. Cheng et al. [22] caution, however, that persona-based prompting can amplify stereotypical associations rather than faithfully reproduce subgroup opinions, a risk that applies to both indigenous and global models. The Korean NLP community has expanded this work in several directions: cross-lingual post-training [14] and personalized LLM applications [11], among others.

Two recent studies raise broader concerns about LLM simulation fidelity. Hu and Collier [19] quantified the "persona effect" and found that demographic persona variables account for less than 10% of annotation variance. Persona prompting helps only modestly, and the gains vary across models and tasks. Taubenfeld et al. [20] uncovered systematic biases in debate simulations: LLM agents drift toward moderation regardless of assigned political perspectives, diverging from observed human behavior.

Despite this growing literature, systematic survey-simulation comparisons that directly benchmark distributional alignment between indigenous and global models are still rare. The comparison reported here—indigenous vs. global LLMs on identical Korean survey items—has not been attempted before, and we use standardized metrics and formal statistical tests to evaluate it.

3. Methods

3.1 Silicon Sampling Framework

The framework has four components:

- **Benchmark Selection**—a population-representative survey with published marginal distributions.
- **Persona Design**—demographic-constrained prompts representing the target population.
- **Response Generation**—multiple samples per variable under controlled parameters.
- **Distribution Comparison**—statistical metrics comparing generated versus benchmark distributions.

3.2 Benchmark: KGSS 2023

The Korean General Social Survey (KGSS) 2023 is used as the benchmark, a nationally representative survey (N=1,527) conducted by Sungkyunkwan University Survey Research Center [9,25]. KGSS is the Korean adaptation of the U.S. General Social Survey (GSS), designed to track trends in Korean social attitudes and public opinion. Data and documentation are publicly available through the Korea Social Science Data Archive (KOSSDA, DOI: 10.22687/KOSSDA-A1-CUM-0062-V3).From the full KGSS 2023 questionnaire (150+ items), six variables were selected as test cases for methodological validation. The objective is to validate whether silicon sampling produces distributional alignment in a non-Western context, not to comprehensively represent Korean culture. Variable selection criteria were: (1) cultural context plausibly affects LLM responses, (2) published marginal distributions enable reproducibility, and (3) ordinal scales (3-5 points) with sufficient variance enable statistical comparison. A seventh KGSS variable (SATFIN, financial satisfaction) met criteria (1) and (2) but was excluded because economic satisfaction is less specific to Korean cultural context than the inter-Korean, political, and institutional domains that motivate our indigenous-vs.-global comparison. The selected variables capture culturally sensitive domains (Table 1):

Table 1. KGSS 2023 variables selected for analysis.

Variable	Description	Scale
CONFINAN	Confidence in financial institutions	1-3
CONLEGIS	Confidence in legislature	1-3
KRPROUD	Pride in being Korean	1-4
NORTHWHO	Perception of North Korea	1-4
UNIFI	Support for unification	1-4
PARTYLR	Political left-right orientation	1-5

These variables represent domains where indigenous model advantages are most plausible: inter-Korean relations (NORTHWHO, UNIFI) receive extensive Korean media coverage but minimal global English-language coverage; political orientation (PARTYLR) involves Korean-specific meanings of "progressive/conservative" that include North Korea policy dimensions; institutional trust (CONFINAN, CONLEGIS) reflects current Korean political context (2023 presidential transition, National Assembly approval ratings); national identity (KRPROUD) is culturally grounded in Korean historical experience. Non-Western populations often exhibit distinctive response patterns that standard instruments—developed primarily on Western, Educated, Industrialized, Rich, and Democratic (WEIRD) samples—may not capture [13].These six variables serve as test cases where cultural differences between indigenous and global models are theoretically motivated, not as a comprehensive representation of Korean public opinion.

3.3 Models Compared

- **GPT-5.2**: OpenAI model (API identifier: `gpt-5.2`, accessed December 2025). We report the exact API identifier for reproducibility; all experiments used the OpenAI Chat Completions API.
- **CLOVA HCX-007**: Naver's Korean indigenous LLM with reasoning capabilities (API identifier: `HCX-007`, accessed December 2025). Experiments used the CLOVA Studio Chat Completions API v3.

Neither Naver nor OpenAI publicly disclose HCX-007 or GPT-5.2 parameter counts or detailed architecture specifications. This limits our ability to attribute performance differences to model origin versus model scale, and we flag this as a caveat throughout the interpretation of results.

3.4 Experimental Design

Two experiments were conducted under a shared persona-prompting protocol:

- **Experiment 1 (Baseline simulation)**: Generate n=100 persona-conditioned responses per variable using GPT-5.2 and compare the resulting marginal distributions to KGSS 2023 benchmarks.
- **Experiment 2 (Cultural-context comparison)**: Repeat the same procedure with CLOVA HCX-007 and compare model–benchmark alignment (CLOVA vs. KGSS) as well as relative alignment (CLOVA vs. GPT-5.2) on the same variables.

Both models received the same set of 100 personas, stratified to match KGSS 2023 demographic marginals, and the same Korean prompt template (Section 3.6). Experimental configurations, response data, and analysis scripts are available in the supplementary materials.

Sample Size Rationale: n=100 samples per variable was selected based on three considerations: adequate statistical power ($1-\beta > 0.80$) for medium-to-large effect sizes in chi-square tests, stable JS divergence estimates (bootstrap SD < 0.025), and comparability with prior studies [2,8]. Multi-seed robustness validation with 5 independent runs (seeds 42-46) was conducted to address stochastic variability.

3.5 Evaluation Metrics

Three metrics are employed with explicit prioritization:

Primary Metric: Jensen–Shannon Divergence (JS) JS divergence is a symmetric measure of distributional similarity [6], bounded in [0, ln 2] when computed with natural logarithm (0 indicates identical distributions). JS divergence is adopted as the primary metric because it is scale-invariant, enables direct comparison with prior silicon sampling studies [2,8], and quantifies overall distributional similarity. For interpretation, JS < 0.05 indicates close alignment; JS > 0.15 indicates substantial divergence. **Secondary Metric: Chi-square Goodness-of-Fit Test** The chi-square test assesses whether observed response frequencies differ considerably from expected frequencies (KGSS benchmarks). This test is appropriate for discrete categorical data [10] and provides formal hypothesis testing (H_0: simulated distribution equals benchmark distribution). $\alpha = 0.05$ is used; rejection indicates statistically significant distributional differences. **Supplementary Metric: Kolmogorov-Smirnov Test (KS)** The KS test is included for comparability with prior literature [2] but has known limitations for ordinal survey data (discrete CDFs, tied values). KS results are reported as supplementary; JS divergence and chi-square tests are prioritized for inferential conclusions. **Metric Interpretation Priority**: When metrics conflict, priority is: (1) JS divergence for magnitude, (2) chi-square for statistical significance, (3) KS for literature comparison.

3.6 Prompt Design and Response Parsing

3.6.1 SYSTEM MESSAGE

All models received the following system message:

```
You are a survey respondent. Answer the question with only a number.
```

3.6.2 PERSONA PROMPT TEMPLATE

The user prompt followed this structure:

```
You are a Korean respondent with the following characteristics:

- Age: {age_group}
- Gender: {gender}
- Education: {education}
- Region: {region}
- Occupation: {occupation}

For the following question, answer with only a number from {min_value} to
{max_value}.

Question: {question}
Scale: {scale_labels}

Answer (number only):
```

3.6.3 RESPONSE PARSING AND SAMPLING PARAMETERS

Model responses were parsed by extracting the first integer via regex (`r"-?\d+"`), validating against the variable's response range, and assigning the scale midpoint as default when parsing failed. Default assignments were recorded for compliance tracking. Both models used temperature 0.7 with 3 retry attempts (exponential backoff). GPT-5.2 used max_tokens=10 with 10-20 concurrent API calls; CLOVA HCX-007 used the `thinking: short` parameter with sequential calls (0.5s delay).

4. Results

We present results in two parts: baseline alignment (Experiment 1) and the cultural-context comparison (Experiment 2).

4.1 Experiment 1: Baseline Simulation

Using GPT-5.2 (T=0.7) with demographic persona prompts, n=100 responses per variable were generated. Baseline alignment varies across variables: GPT-5.2 achieves close alignment on political orientation (PARTYLR: JS=0.038) and institutional trust (CONFINAN: JS=0.062), but poor alignment on unification attitudes (UNIFI: JS=0.267), with an average JS divergence of 0.123. This heterogeneous performance motivates evaluating alignment at the variable level. Detailed per-variable results are integrated into Table 2 (Section 4.2), which reports 5-seed means; single-run (seed 42) values differ by less than 0.002 from 5-seed averages.

4.2 Experiment 2: Cultural Context

Claim tested: Indigenous LLMs trained on local cultural data outperform global models in reproducing culturally-specific response patterns. To guard against stochastic variability, we ran 5 independent replications (seeds 42-46), each with a different stratified persona set.

Multi-seed validation yields a conservative estimate of indigenous model advantages (Table 2). CLOVA outperforms GPT-5.2 on 2 of 6 variables (CONLEGIS: +35.8%, UNIFI: +54.1%), while GPT-5.2 shows superior or comparable performance on the remaining four variables. The 21.3% average improvement is concentrated in specific domains.Table 2. Multi-seed robustness validation: JS divergence (ln) across 5 independent runs (seeds 42-46).

Table 2. Multi-seed robustness validation: JS divergence (ln) across 5 independent runs (seeds 42-46).

Variable	GPT-5.2 Mean ± SD	CLOVA Mean ± SD	Improvement	CLOVA Better
CONFINAN	0.063 ± 0.002	0.070 ± 0.015	-11.1%	No
CONLEGIS	0.134 ± 0.000	0.086 ± 0.030	+35.8%	Yes
KRPROUD	0.105 ± 0.008	0.118 ± 0.056	-12.4%	No
NORTHWHO	0.123 ± 0.009	0.134 ± 0.039	-8.9%	No
UNIFI	0.259 ± 0.011	0.119 ± 0.033	+54.1%	Yes
PARTYLR	0.046 ± 0.011	0.051 ± 0.013	-10.9%	No
Average	**0.122 ± 0.007**	**0.096 ± 0.031**	**+21.3%**	**2 of 6**

Note: Mean ± SD computed from 5 independent sampling runs with different persona sets.

The pattern is clear: CLOVA wins on just 2 of 6 variables. Its 21.3% average improvement is driven almost entirely by UNIFI (+54.1%) and CONLEGIS (+35.8%). On the remaining four variables—political orientation, national pride, financial trust, and North Korea perception—GPT-5.2 performs as well or better, and does so with lower variance across seeds. The data suggest that indigenous training confers domain-specific advantages rather than a uniform edge across culturally-sensitive variables.

STATISTICAL VALIDATION

Chi-square goodness-of-fit tests and KS tests were conducted to assess whether CLOVA distributions differ considerably from KGSS 2023 benchmarks (Table 3).Table 3. Statistical validation: Chi-square and KS test results for CLOVA HCX-007 vs. KGSS 2023 benchmarks.

Table 3. Statistical validation: Chi-square and KS test results for CLOVA HCX-007 vs. KGSS 2023 benchmarks.

Variable	Chi-square statistic	Chi-square p-value	KS p-value (supplementary)
CONFINAN	15.2	0.0005	0.010
CONLEGIS	42.8	< 0.0001	5.43e-09
KRPROUD	89.4	< 0.0001	6.75e-20
NORTHWHO	28.6	< 0.0001	2.45e-05
UNIFI	52.1	< 0.0001	4.41e-10
PARTYLR	31.7	< 0.0001	1.55e-05

Chi-square tests reject distributional equivalence for all six variables at $\alpha = 0.05$, confirming that CLOVA-generated distributions differ greatly from KGSS benchmarks. KS tests (supplementary metric) yield consistent conclusions. The low p-values for KRPROUD ($\chi^2 = 89.4$) and UNIFI ($\chi^2 = 52.1$) indicate substantial category-level frequency mismatches despite CLOVA's lower average JS divergence relative to GPT-5.2.

4.3 Multi-Seed Robustness and Response Compliance

Cross-seed variability is captured in Table 2 (SD columns). GPT-5.2 shows higher stability (mean SD=0.007) compared to CLOVA (mean SD=0.031). KRPROUD exhibits the highest CLOVA variance (SD=0.056; JS ranging from 0.035 to 0.186 across seeds), while CONLEGIS and UNIFI consistently favor CLOVA across all 5 seeds. Bootstrap 95% confidence intervals for CLOVA JS divergence range from narrow (CONFINAN: [0.062, 0.085]) to wide (KRPROUD: [0.075, 0.157]), confirming adequate precision for comparative conclusions.Both models achieve high response-format compliance (GPT-5.2: 97.8%, CLOVA: 94.5%; $\chi^2 = 8.42$, $p = 0.004$). However, compliance differences do not explain alignment differences: CLOVA's lowest compliance occurs on UNIFI (92%), yet CLOVA outperforms GPT-5.2 on this variable by 54.1%. Sensitivity analysis excluding default-assigned responses yields JS divergence changes below 0.005 for 5/6 variables.

5. Discussion

5.1 Baseline Distributional Validity

Baseline alignment is heterogeneous across variables (Section 4.1), with strong response concentration on some items (e.g., UNIFI) and close approximation on others (e.g., PARTYLR). We did not anticipate the magnitude of this variation: UNIFI JS divergence exceeds PARTYLR by a factor of seven. This motivates reporting silicon sampling performance at the variable level—aggregate accuracy scores can mask variable-level failures that would be disqualifying for substantive use.

5.2 Cultural Context and Variable Dependence

The seed-level data reveal a clear split in indigenous model performance. CLOVA consistently outperforms GPT-5.2 on legislative trust (CONLEGIS: +35.8%) and unification attitudes (UNIFI: +54.1%) across all 5 seeds—these two variables drive the aggregate improvement. On the remaining four variables—political orientation, national pride, financial trust, and North Korea perception—GPT-5.2 matches or beats CLOVA across most seeds. We suspect this asymmetry reflects the topics where Korean-language training corpora are most distinctive: inter-Korean relations and legislative politics receive dense Korean media coverage with little parallel in English sources, whereas economic satisfaction and national pride are discussed in broadly similar terms across languages. CLOVA also exhibits higher cross-seed variance (mean SD=0.031 vs. GPT-5.2's 0.007), reflecting greater sensitivity to persona sampling.Notably, KS tests reject benchmark equality for every CLOVA variable at α=0.05. Lower divergence is not the same as distributional equivalence.

This variable-dependent performance complicates the indigenous advantage hypothesis. The relationship between model origin and cultural alignment is not straightforward. Several mechanisms could account for the pattern: uneven topic coverage in training data, differential prompt sensitivity across domains, and model-specific response priors that interact differently with attitudinal dimensions. Disentangling these factors would require controlled ablation studies that are beyond the scope of the present work.

Persona-Model Interaction Effects: The persona prompt specifies demographic characteristics but not attitudinal priors, which may interact differentially with model training. CLOVA, trained on Korean news and social media, may have learned Korean-specific demographic-attitude correlations (e.g., age → North Korea policy views), consistent with its performance on UNIFI and CONLEGIS. GPT-5.2, trained on global corpora, may encode weaker associations. Indigenous models may also amplify overrepresented patterns while suppressing underrepresented combinations, suggesting that optimal prompt templates may be model-dependent.

5.3 Response-Format Compliance as a Measurement Issue

Survey simulation hinges on strict response-format compliance—models must reply with a single integer, nothing more. Post-processing rules (Section 3.6.3) can shape the resulting distributions in nontrivial ways. We argue that silicon sampling studies should routinely report response validity rates (e.g., what fraction of responses required default assignment) alongside divergence metrics. Transparent reporting is not optional. As LLMs are increasingly deployed for social science research where reproducibility is paramount [12], methodological credibility demands that such standards be established.

5.4 Implications and Limitations

The practical takeaway is straightforward: researchers should run variable-level pilot tests and disclose prompt templates, parsing rules, and validity rates before drawing conclusions from silicon sampling.

That said, this study has clear limitations. We evaluated only marginal distributions, not the joint or conditional distributions that would capture demographic interactions. Our sample size (n=100 per variable) provides adequate power for medium-to-large effects but may miss smaller distributional differences. The KS test, designed for continuous distributions, is not ideal for ordinal survey data—a point we have flagged throughout. Additionally, because neither OpenAI nor Naver disclose model parameter counts, we cannot rule out that performance differences reflect model scale rather than training-data origin.

Future work should evaluate joint distributions, expand seed ranges, and adopt ordinal-appropriate metrics such as Wasserstein-1 distance.

6. Conclusion

The central finding of this study is negative: indigenous model training does not guarantee better cultural alignment in survey simulation. CLOVA HCX-007 outperformed GPT-5.2 on legislative trust and unification attitudes—two domains where Korean-language training corpora plausibly diverge most from English sources—but showed no advantage on four other culturally-sensitive variables. Neither model achieved distributional equivalence with KGSS 2023 benchmarks by any statistical criterion.

This matters for the broader silicon sampling agenda. If even a model trained predominantly on Korean text cannot reliably reproduce Korean survey distributions, then cultural alignment in LLM simulation is harder to achieve than the initial optimism around silicon sampling suggested. Researchers who plan to use LLM-generated responses as proxies for survey data should not assume that matching the model's training language to the survey population is sufficient.

We draw two practical lessons. First, silicon sampling performance must be evaluated at the variable level; aggregate metrics conceal domain-specific failures that would be disqualifying for substantive research. Second, methodological transparency—reporting prompts, parsing rules, default-value rates, and multi-seed robustness—is essential for reproducibility and for building warranted trust in this emerging methodology.

Looking ahead, the most pressing need is to understand why alignment varies across variables. Controlled experiments that manipulate training-data composition, prompt structure, and model architecture independently would help disentangle the sources of variable-dependent performance. Extending this evaluation framework to joint and conditional distributions, additional cultural contexts, and ordinal-appropriate metrics (e.g., Earth Mover's Distance) would further clarify the boundaries of silicon sampling's validity.

Acknowledgement

This work was supported by Hanshin University Grant in 2025.

REFERENCES

[1] G. V. Aher, R. I. Arriaga, and A. T. Kalai, "Using large language models to simulate multiple humans and replicate human subject studies," arXiv preprint arXiv:2208.10264, 2022. https://arxiv.org/abs/2208.10264

[2] L. P. Argyle, E. C. Busby, N. Fulda, J. R. Gubler, C. Rytting, and D. Wingate, "Out of one, many: Using language models to simulate human samples," Political Analysis, Vol. 31, No. 3, pp. 337-351, 2023. https://doi.org/10.1017/pan.2023.2

[3] Y. Cao, L. Zhou, S. Lee, L. Cabello, M. Chen, and D. Hershcovich, "Assessing cross-cultural alignment between ChatGPT and human societies: An empirical study," in Proc. First Workshop on Cross-Cultural Considerations in NLP (C3NLP), 2023. https://doi.org/10.18653/v1/2023.c3nlp-1.7

[4] D. Dillon, N. Tandon, Y. Gu, and K. Gray, "Can AI language models replace human participants?" Trends in Cognitive Sciences, Vol. 27, No. 7, pp. 597-600, 2023. https://doi.org/10.1016/j.tics.2023.04.008

[5] B. Kim, H. Kim, S. W. Lee, G. Lee, D. Kwak, D. H. Jeon, et al., "What changes can large-scale language models bring? Intensive study on HyperCLOVA: Billions-scale Korean generative pretrained transformers," in Proc. 2021 Conf. on Empirical Methods in Natural Language Processing, 2021. https://doi.org/10.18653/v1/2021.emnlp-main.274

[6] J. Lin, "Divergence measures based on the Shannon entropy," IEEE Trans. on Information Theory, Vol. 37, No. 1, pp. 145-151, 1991. https://doi.org/10.1109/18.61115

[7] T. Naous, M. J. Ryan, A. Ritter, and W. Xu, "Having beer after prayer? Measuring cultural bias in large language models," in Proc. 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2024. https://doi.org/10.18653/v1/2024.acl-long.862

[8] J. T. Ornstein, E. N. Blasingame, and J. S. Truscott, "How to train your stochastic parrot: Large language models for political texts," Political Science Research and Methods, 2024. https://doi.org/10.1017/psrm.2024.2

[9] Sungkyunkwan University Survey Research Center, Korean General Social Survey 2023 Codebook, Korean Social Science Data Archive (KOSSDA), 2023. https://kossda.snu.ac.kr

[10] A. Agresti, Categorical Data Analysis, 2nd ed., Wiley, 2002.

[11] S.-H. Cho and Y.-S. Lee, "Study on Personal Large Language Model (LLM)," International Journal of Advanced Smart Convergence, Vol. 13, No. 4, pp. 204-209, 2024. https://doi.org/10.7236/IJASC.2024.13.4.204

[12] C. A. Bail, "Can generative AI improve social science?" Proceedings of the National Academy of Sciences, Vol. 121, No. 21, e2314021121, 2024. https://doi.org/10.1073/pnas.2314021121

[13] J. Henrich, S. J. Heine, and A. Norenzayan, "The weirdest people in the world?" Behavioral and Brain Sciences, Vol. 33, No. 2-3, pp. 61-83, 2010. https://doi.org/10.1017/S0140525X0999152X

[14] S. Son, C. Park, J. Lee, M. Shim, C. Lee, K. Park, and H. Lim, "Korean and Multilingual Language Models Study for Cross-Lingual Post-Training (XPT)," Journal of the Korea Convergence Society, Vol. 13, No. 3, pp. 77-89, 2022. https://doi.org/10.15207/JKCS.2022.13.03.077

[15] S. Santurkar, E. Durmus, F. Ladhak, C. Lee, P. Liang, and T. Hashimoto, "Whose opinions do language models reflect?" in Proc. 40th Int. Conf. on Machine Learning (ICML), PMLR 202, pp. 29971-30004, 2023. https://proceedings.mlr.press/v202/santurkar23a.html

[16] J. Bisbee, J. D. Clinton, C. Dorff, B. Kenkel, and J. M. Larson, "Synthetic replacements for human survey data? The perils of large language models," Political Analysis, Vol. 32, No. 4, pp. 401-416, 2024. https://doi.org/10.1017/pan.2024.5

[17] Y. Tao, R. S. Baker, O. Viberg, and R. F. Kizilcec, "Cultural bias and cultural alignment of large language models," PNAS Nexus, Vol. 3, No. 9, pgae346, 2024. https://doi.org/10.1093/pnasnexus/pgae346

[18] C. Li, M. Chen, J. Wang, S. Sitaram, and X. Xie, "CultureLLM: Incorporating cultural differences into large language models," in Proc. Advances in Neural Information Processing Systems 37 (NeurIPS 2024), 2024. https://proceedings.neurips.cc/paper_files/paper/2024/hash/9a16935bf54c4af233e25d998b7f4a2c-Abstract-Conference.html

[19] T. Hu and N. Collier, "Quantifying the persona effect in LLM simulations," in Proc. 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 10289-10307, 2024. https://doi.org/10.18653/v1/2024.acl-long.554

[20] A. Taubenfeld, Y. Dover, R. Reichart, and A. Goldstein, "Systematic biases in LLM simulations of debates," in Proc. 2024 Conf. on Empirical Methods in Natural Language Processing (EMNLP), pp. 251-267, 2024. https://doi.org/10.18653/v1/2024.emnlp-main.16

[21] E. Durmus, K. Nguyen, T. I. Liao, N. Schiefer, A. Askell, et al., "Towards measuring the representation of subjective global opinions in language models," in Proc. 1st Conf. on Language Modeling (COLM), 2024. https://openreview.net/forum?id=us5sBk0fLl

[22] M. Cheng, T. Durmus, and D. Jurafsky, "Marked personas: Using natural language prompts to measure stereotypes in language models," in Proc. 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1504-1532, 2023. https://doi.org/10.18653/v1/2023.acl-long.84

[23] J. Hartmann, J. Schwenzow, and M. Witte, "The political ideology of conversational AI: Converging evidence on ChatGPT's pro-environmental, left-libertarian orientation," arXiv preprint arXiv:2301.01768, 2023. https://arxiv.org/abs/2301.01768

[24] S. Son, D. Lee, H. Kim, et al., "KMMLU: Measuring massive multitask language understanding in Korean," in Proc. 2025 Conf. of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL), 2025. https://arxiv.org/abs/2402.11548

[25] Y. J. Kim, "Korean General Social Survey (KGSS) cumulative data, 2003-2023," ICPSR, Inter-university Consortium for Political and Social Research, 2024. https://doi.org/10.3886/ICPSR38520.v3