KAIST

거대언어모델을 활용한 설문조사의 현재와 한계

김란우 Lanu Kim (lanukim@kaist.ac.kr)
KAIST 디지털인문사회과학부
https://lanukim.github.io, https://computationalsociologylab.github.io
국가데이터연구 심포지엄 2026

발표 개요

- 거대언어모델을 활용한 설문조사 (합성 응답) 소개

- 합성응답을 통한 사회조사의 한계점

1. 구조와 개인의 문제
2. 분석 도구로서의 불완전성
3. 사회구조의 편향을 그대로 학습
4. 신념 체계의 왜곡
5. 설문조사 연구를 방해할 가능성

- LLM을 통한 연구 사례

1. 설문 합성 데이터에서 사회적 다수·소수 집단 간 LLM 성능 차이의 제한성
2. LLM을 활용한 삼성 반도체의 기업 이미지 분석

기존 설문조사의 문제점

- 오프라인 조사 (선거 출구 조사)
- 시간, 노력, 비용의 문제
- 전화 조사
- 대표성의 문제
- 무작위 표본 vs. 자발적 표본?
- 온라인 패널 조사
- 대표성의 문제
- 익명성의 문제
- 문항 수/길이의 문제

대안으로써의 거대언어모델

- **그림 1.** 연구 1에서 분석된 네 명의 실리콘 '개인' 예시 맥락 및 응답. 일반 텍스트는 조건화 맥락, 밑줄 단어는 동적으로 삽입된 인구통계 정보, 파란 단어는 수집된 4개의 응답 단어.

	Describing Democrats	Describing Republicans
Strong Republicans	Ideologically, I describe myself as conservative. Politically, I am a strong Republican. Racially, I am white. I am male. Financially, I am upper-class. In terms of my age, I am young. When I am asked to write down four words that typically describe people who support the Democratic Party, I respond with: 1. **Liberal 2. Socialist 3. Communist 4. Atheist.**	Ideologically, I describe myself as conservative. Politically, I am a strong Republican. Racially, I am white. I am male. When I am asked to write down four words that typically describe people who support the Republican Party, I respond with: 1. **Conservative 2. Male 3. White (or Caucasian) 4. Christian.**
Strong Democrats	Ideologically, I describe myself as liberal. Politically, I am a strong Democrat. Racially, I am white. I am female. Financially, I am poor. In terms of my age, I am old. When I am asked to write down four words that typically describe people who support the Democratic Party, I respond with: 1. **Liberal. 2. Young. 3. Female. 4. Poor.**	Ideologically, I describe myself as extremely liberal. Politically, I am a strong Democrat. Racially, I am hispanic. I am male. Financially, I am upper-class. In terms of my age, I am middle-aged. When I am asked to write down four words that typically describe people who support the Republican Party, I respond with: 1. **Ignorant 2. Racist 3. Misogynist 4. Homophobic.**

Figure 1. Example contexts and completions from four silicon "individuals" analyzed in Study 1. Plaintext indicates the conditioning context; underlined words show demographics we dynamically inserted into the template; blue words are the four harvested words.

Argyle et al. (2023)

Describing Democrats

Strong Republicans	Ideologically, I describe myself as conservative. Politically, I am a strong Republican. Racially, I am white. I am male. Financially, I am upper-class. In terms of my age, I am young. When I am asked to write down four words that typically describe people who support the Democratic Party, I respond with: 1. **Liberal** 2. **Socialist** 3. **Communist** 4. **Atheist.**

Argyle et al. (2023)

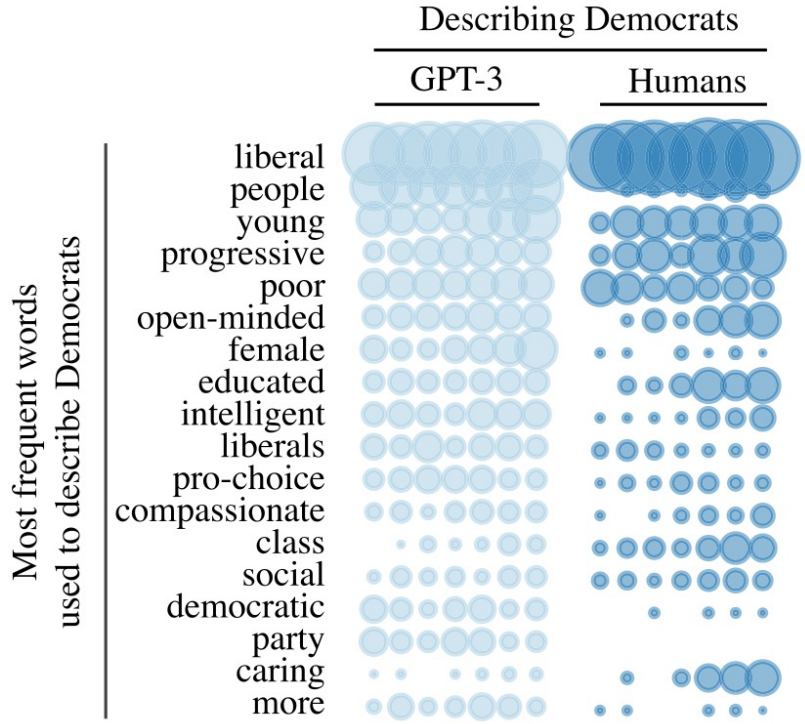
합성 응답(synthetic response)의 장점

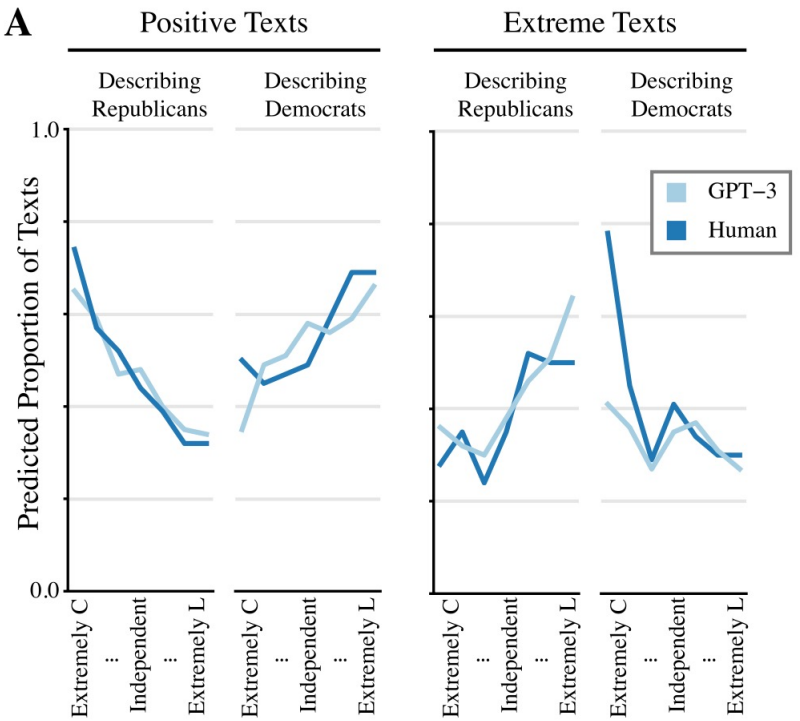
- 시간, 노력, 비용의 문제 해결
- 문항의 수/길이의 문제 해결

현재까지의 결과

- 집단별 평균값은 놀랍게도 유사

Argyle et al. (2023)





참고 논문: "하나에서 여럿으로" — 언어 모델로 인간 표본 시뮬레이션

Out of **one**, **many**: Using language models to simulate human samples

LP Argyle, EC Busby, N Fulda, JR Gubler… - Political …, 2023 - cambridge.org

… complex reflection of the **many** various patterns of association … models do not contain just **one** bias, but **many**. This means that … Such an effort goes well beyond what **one** research team …

☆ Save 99 Cite Cited by 1174 Related articles All 14 versions

Out of one, many: Using language models to simulate human samples

Search within citing articles

A survey on large language model based autonomous agents — [PDF] springer.com
L Wang, C Ma, X Feng, Z Zhang, H Yang… - Frontiers of Computer …, 2024 - Springer
Autonomous agents have long been a research focus in academic and industry communities. Previous research often focuses on training agents with limited knowledge …
☆ Save 99 Cite Cited by 2576 Related articles All 8 versions

[HTML] Artificial intelligence and illusions of understanding in scientific research — [HTML] nature.com
L Messeri, MJ Crockett - Nature, 2024 - nature.com
Scientists are enthusiastically imagining ways in which artificial intelligence (AI) tools might improve research. Why are AI tools so attractive and what are the risks of implementing them …
☆ Save 99 Cite Cited by 669 Related articles All 10 versions

ChatGPT outperforms crowd workers for text-annotation tasks — [PDF] pnas.org Full View
F Gilardi, M Alizadeh, M Kubli - Proceedings of the National Academy of …, 2023 - pnas.org
Many NLP applications require manual text annotations for a variety of tasks, notably to train classifiers or evaluate the performance of unsupervised models. Depending on the size and …
☆ Save 99 Cite Cited by 1747 Related articles All 11 versions

Alpacafarm: A simulation framework for methods that learn from human feedback — [PDF] neurips.cc
Y Dubois, CX Li, R Taori, T Zhang… - Advances in …, 2023 - proceedings.neurips.cc
Large language models (LLMs) such as ChatGPT have seen widespread adoption due to their ability to follow user instructions well. Developing these LLMs involves a complex yet …
☆ Save 99 Cite Cited by 750 Related articles All 6 versions »

Whose opinions do language models reflect? — [PDF] mlr.press
S Santurkar, E Durmus, F Ladhak… - International …, 2023 - proceedings.mlr.press
Abstract Language models (LMs) are increasingly being used in open-ended contexts, where the opinions they reflect in response to subjective queries can have a profound …
☆ Save 99 Cite Cited by 802 Related articles All 10 versions »

Challenges and applications of large language models — [PDF] arxiv.org
J Kaddour, J Harris, M Mozes, H Bradley… - arXiv preprint arXiv …, 2023 - arxiv.org
Large Language Models (LLMs) went from non-existent to ubiquitous in the machine learning discourse within a few years. Due to the fast pace of the field, it is difficult to identify …
☆ Save 99 Cite Cited by 835 Related articles All 4 versions »

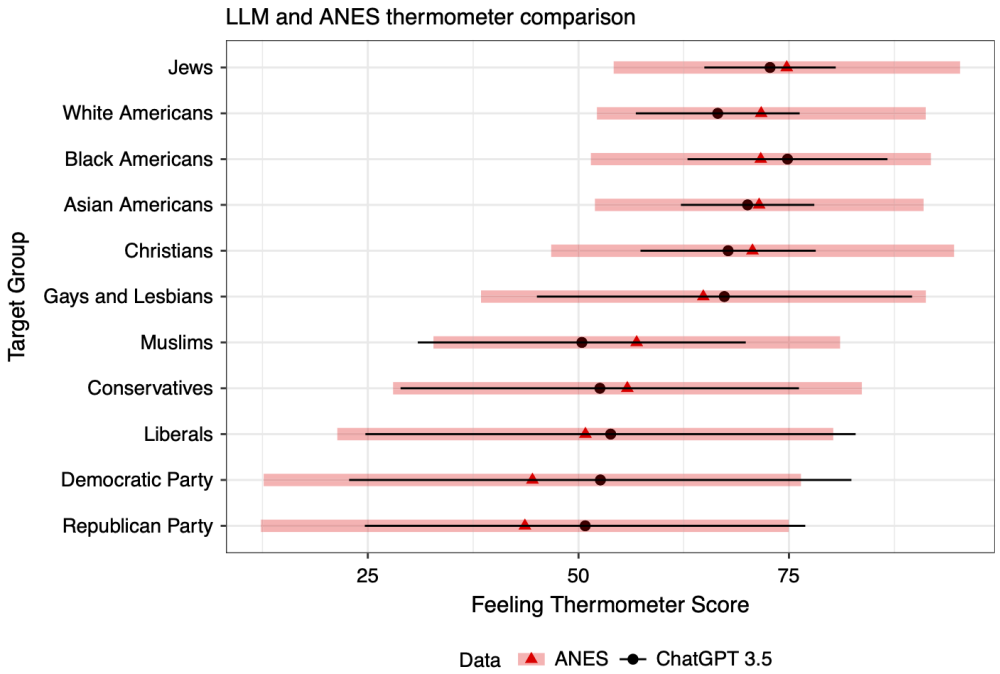
Can AI language models replace human participants? — [HTML] sciencedirect.com
D Dillion, N Tandon, Y Gu, K Gray - Trends in Cognitive Sciences, 2023 - cell.com
Recent work suggests that language models such as GPT can make human-like judgments across a number of domains. We explore whether and when language models might …
☆ Save 99 Cite Cited by 594 Related articles All 8 versions

합성응답을 통한 사회조사의 한계점

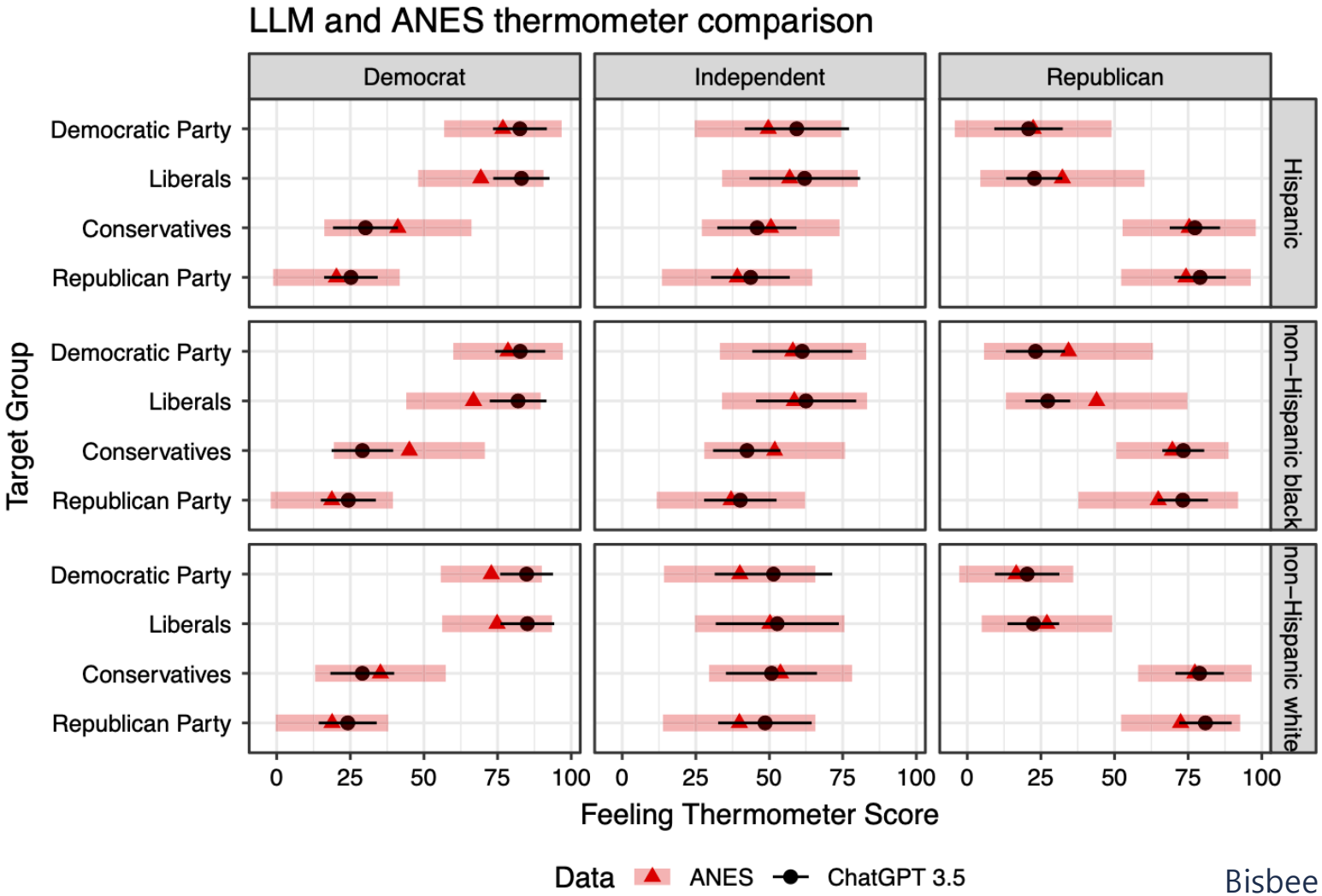
1. 구조와 개인의 문제
2. 분석 도구로서의 불완전성
3. 사회구조의 편향을 그대로 학습
4. 신념 체계의 왜곡
5. 설문조사 연구를 방해할 가능성

1. 구조와 개인의 문제

- 확률모델로서의 LLM => 구조의 과대대표



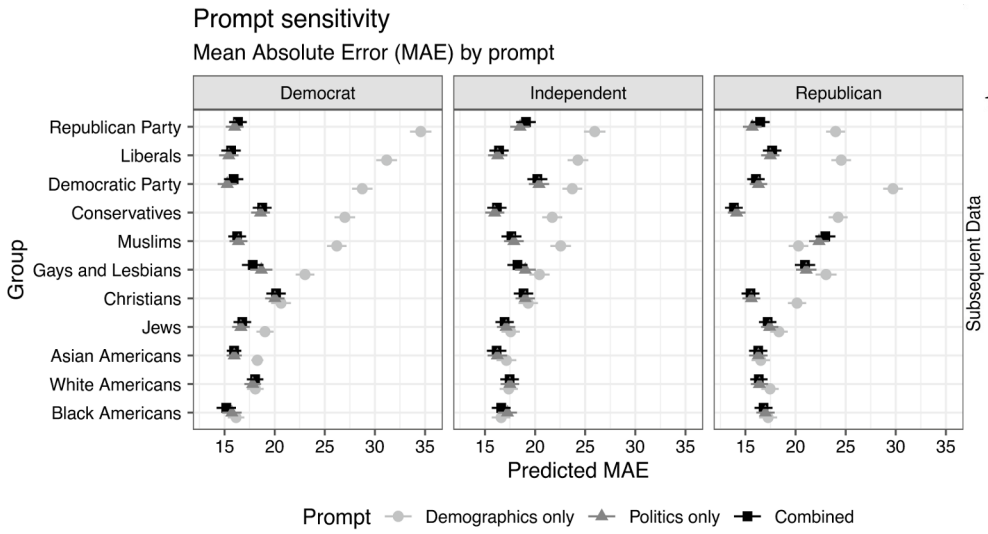
Bisbee et al[3] (2024)

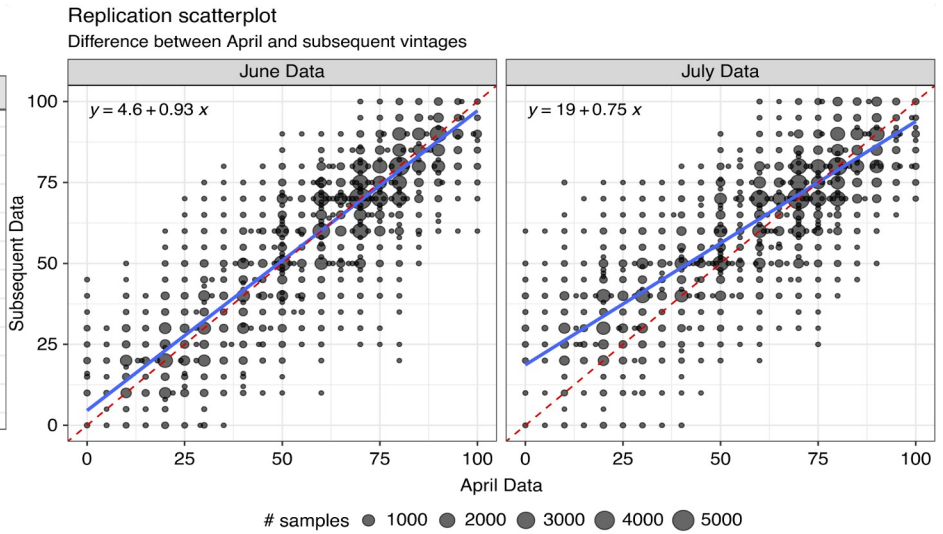


Bisbee et al[4] (2024)

- 건강하지 못한 과학
- 특정 그룹에 대한 고정관념 재생산

2. 분석 도구로서의 불완전성





프롬프트 워딩과 분석한 시기에 따라 분석 결과가 일정하지 않음.

Bisbee et al. (2024)

2. 분석 도구로서의 불완전성

Simulation	Covariates	Cohen's Kappa	Cramer's V	Proportion Agreement
Environmental Issue	With Multiple Covariates	0.270	0.274	66.83%
	Without Covariates	0.000	.	38.07%
Political Issue	With One Covariate	0.306	0.324	65.92%
	Without Covariate	0.145	0.263	54.71%

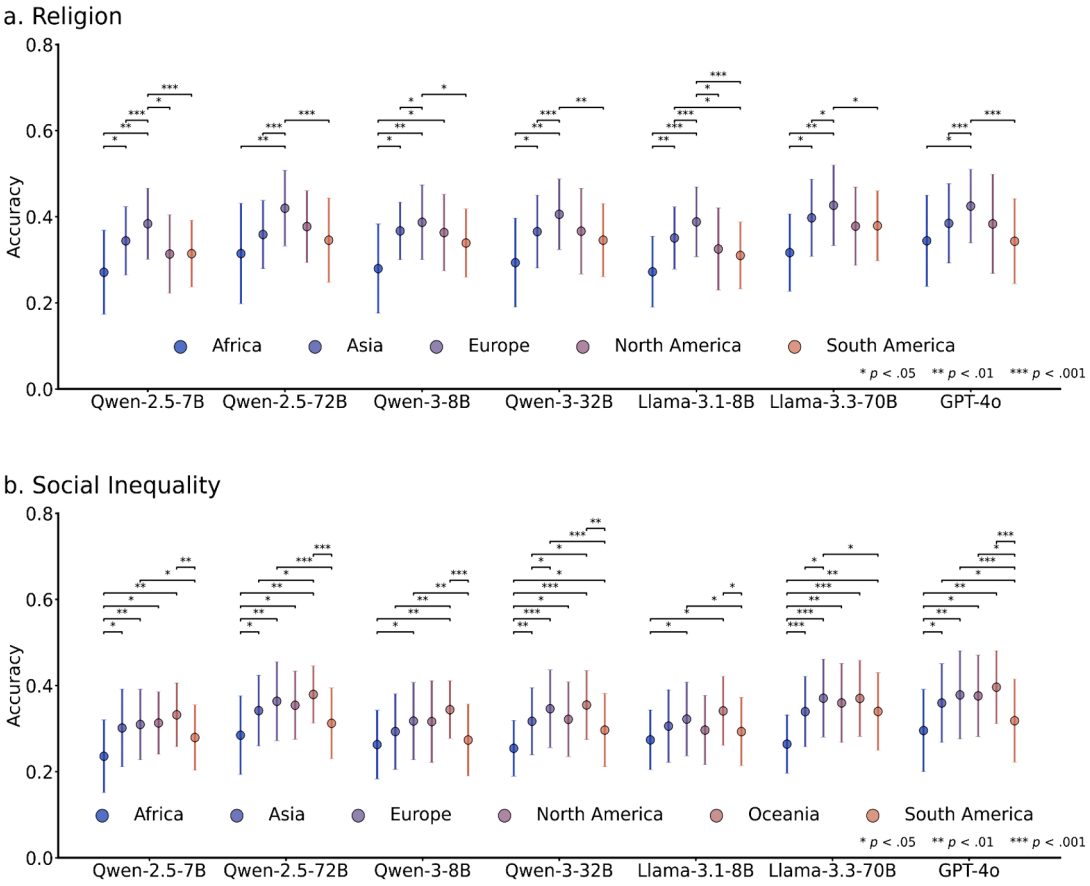
Topic	Difference in Liberal Proportion (%)
Environmental Issue	−6.10
Political Issue	16.33

Values Considered	Cohen's Kappa	Cramer's V	Proportion Agreement
4 Values	0.109	0.157	29.21%
2 Values	0.306	0.324	65.92%

설문 주제, 이념, 선택지의 복잡 정도에 따라 정확도가 달라짐.

Qu et al. (2024)

3. 사회구조의 편향을 그대로 학습

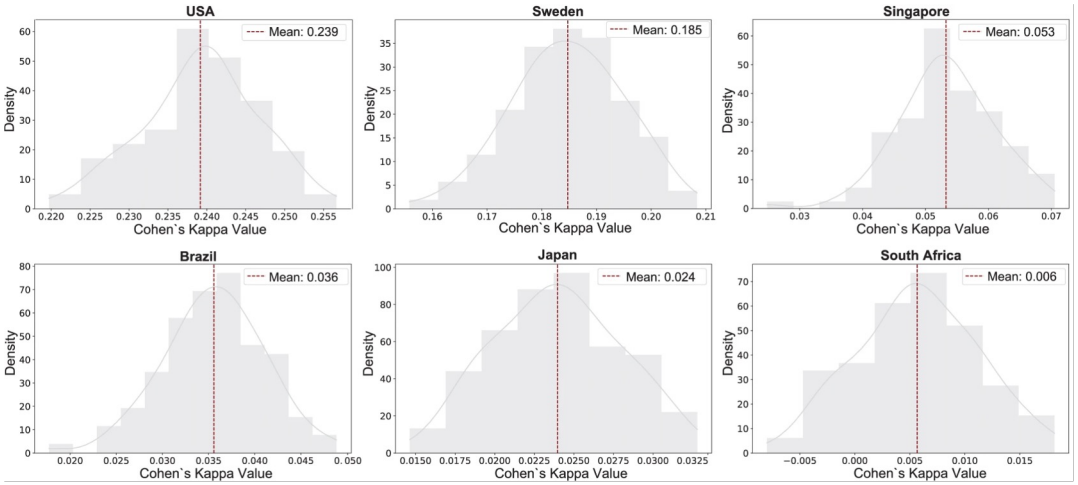


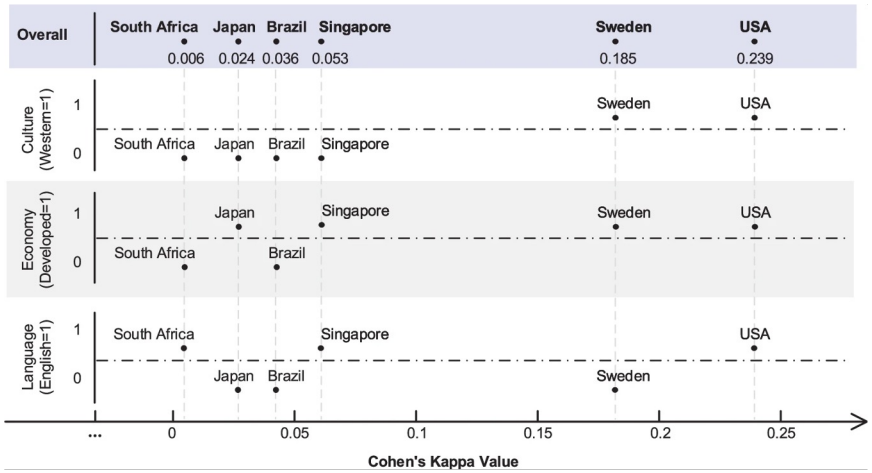
아프리카 페르소나는 유럽, 북미, 오세아니아 대비 정확도 낮음

→ **인구통계 데이터 편향**

3. 사회구조의 편향을 그대로 학습

LLM이 복제한 여론은 불균일하고 문화적으로 편향되어 있으며,
비서구권 맥락에서 정확도가 낮게 나타남.

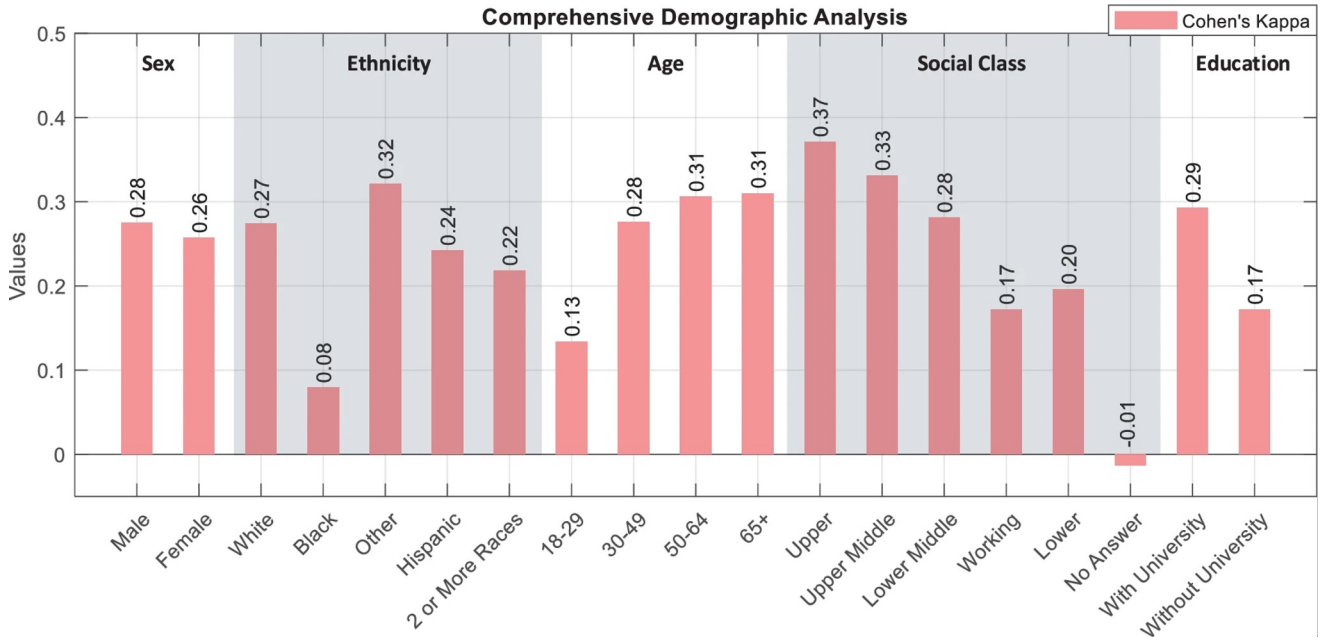




Qu et al. (2024)

3. 사회구조의 편향을 그대로 학습

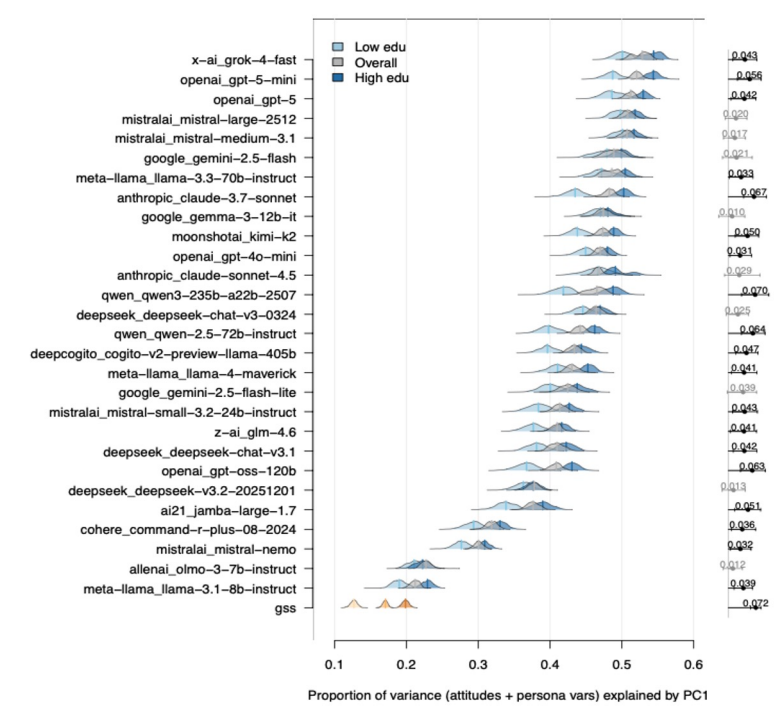
LLM은 미국 내 사회적 지위가 높은 인구통계 집단과 더 강하게 정렬됨.

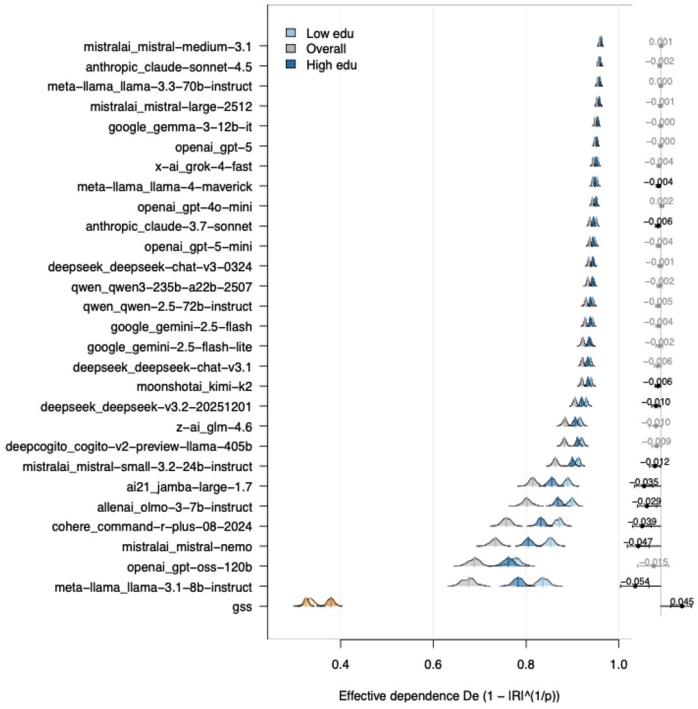


Qu et al. (2024)

4. 신념 체계 왜곡

LLM의 신념은 일차원적이고 강하게 결합되어 있음.





Barrie et al. (2026)

5. 설문조사 연구를 방해할 가능성

PNAS RESEARCH ARTICLE | POLITICAL SCIENCES OPEN ACCESS

The potential existential threat of large language models to online survey research

Sean J. Westwood[a,1]

Edited by James N. Druckman, University of Rochester, Rochester, NY; received July 9, 2025; accepted September 12, 2025 by Editorial Board Member Margaret Levi

The advancement of large language models poses a severe, potentially existential threat to online survey research, a fundamental tool for data collection across the sciences. This work demonstrates that the foundational assumption of survey research—that a coherent response is a human response—is no longer tenable. I designed and tested an autonomous synthetic respondent capable of producing survey data that possesses the coherence and plausibility of human responses. This agent successfully evades a comprehensive suite of data quality checks, including instruction-following tasks, logic puzzles, and "reverse shibboleth" questions designed to detect nonhuman actors, achieving a 99.8% pass rate on 6,000 trials of standard attention checks. The synthetic respondent generates internally consistent responses by maintaining a coherent demographic persona and a memory of its prior answers, producing plausible data on psychometric scales, vignette comprehension tasks, and complex socioeconomic trade-offs. Furthermore, its open-ended text responses are linguistically sophisticated and stylistically calibrated to the level of education of its assigned persona. Critically, the agent can be instructed to maliciously alter polling outcomes, demonstrating an overt vector for information warfare. More subtly, it can also infer a researcher's latent hypotheses and produce data that artificially confirms them. These findings reveal a critical vulnerability in our data infrastructure, rendering most current detection methods obsolete and posing a potential existential threat to unsupervised online research. The scientific community must urgently develop new data validation standards and reconsider its reliance on nonprobability, low-barrier online data collection methods.

surveys | large language models | survey quality

Significance

Surveys are a primary source of data across the sciences, from medicine to economics. I demonstrate that the assumption that logically coherent responses are from humans is now untenable. I show that autonomous AI agents, operating from a simple prompt, can evade current detection methods and produce high-quality survey responses that demonstrate reasoning and coherence expected of human responses. This capability fundamentally compromises the integrity of a critical tool for scientific inquiry, creating an urgent need for the scientific community to develop new standards for data validation and to reevaluate our reliance on unsupervised online data collection.

The scientific enterprise rests on the analysis of reliable data (1). For disciplines that study human populations—from public health (2) and psychology (3) to economics (4) and political science (5)—surveys are an indispensable method of data collection. The advent of online platforms such as MechanicalTurk amplified this reliance, democratizing research and enabling the rapid collection of vast datasets on human behavior (6, 7),

NEWS | 28 January 2026

AI chatbots are infiltrating social-science surveys — and getting better at avoiding detection

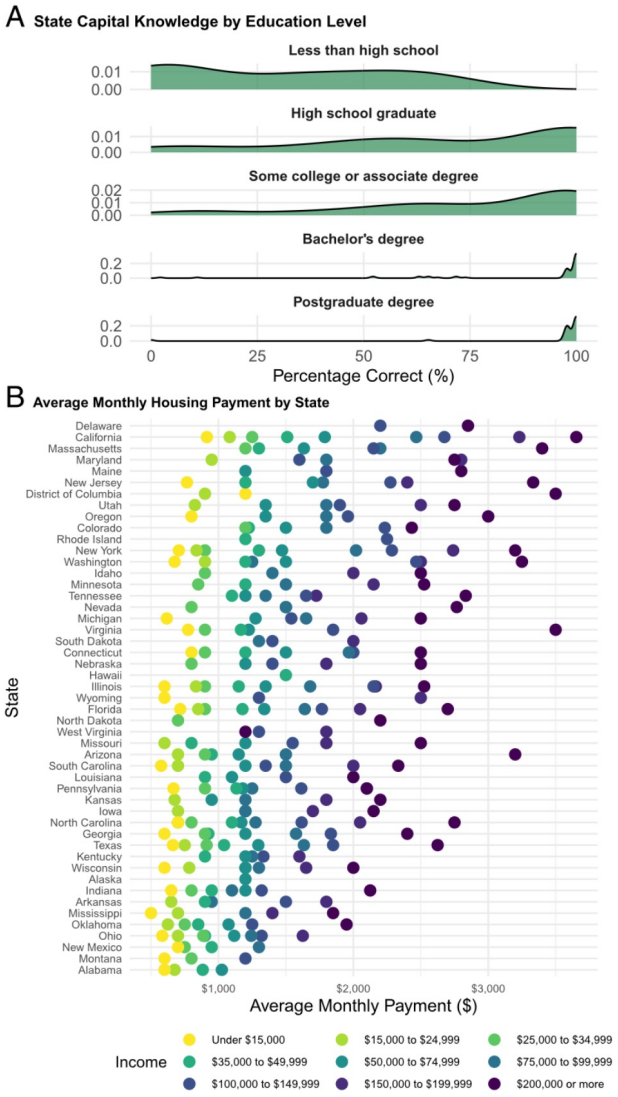
A researcher has created a chatbot that is indistinguishable from human participants in online surveys. Some researchers fear that a workhorse of social science is now under threat.

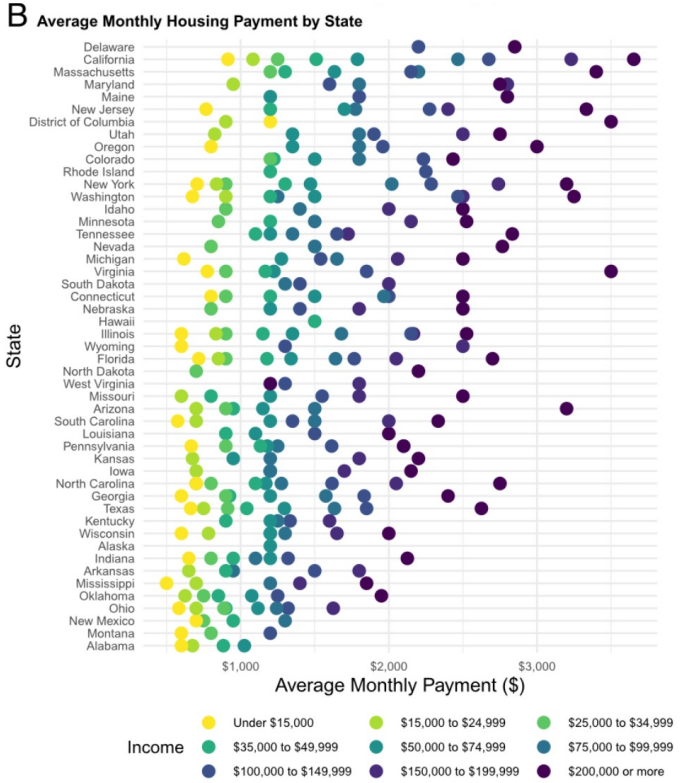
By Sara Phillips

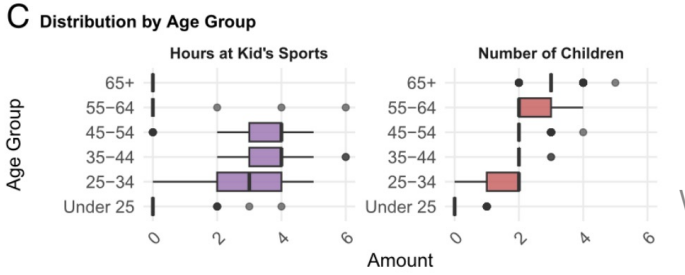
A return to pen-and-paper and in-person surveys might be the only way to protect against sophisticated AI bots, some researchers argue. Credit: Ronstik/Getty

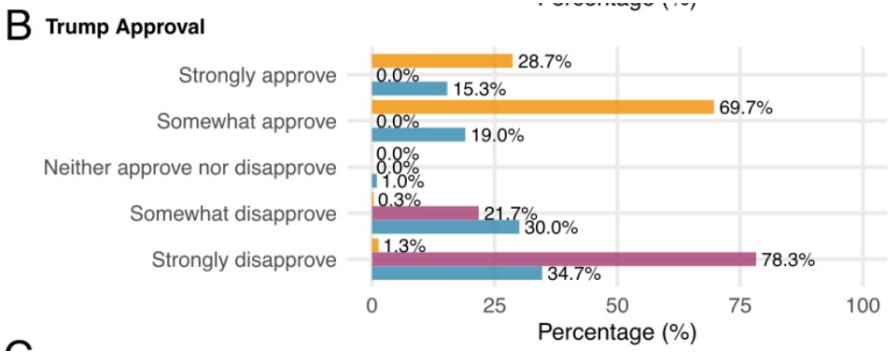
5. 설문조사 연구를 방해할 가능성

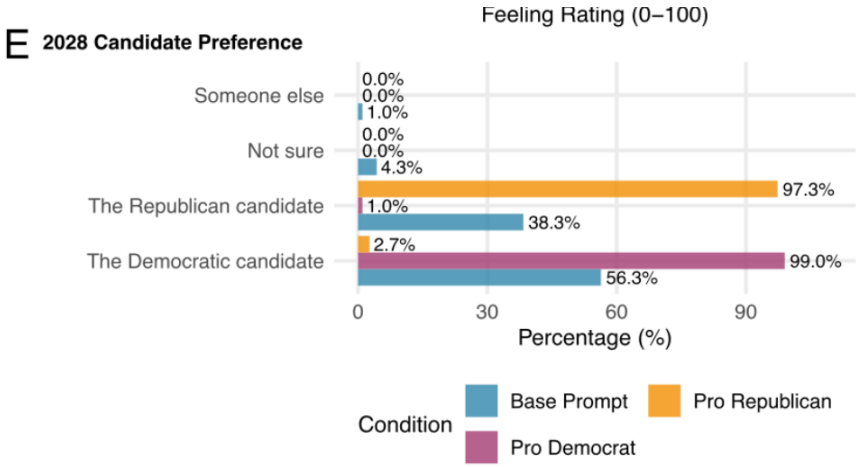
- 기존 온라인 설문조사의 문제점
  - 응답자들이 설문조사에 충분한 주의를 기울이지 않음.
    - 해결 방법: 주의력 체크 문항을 추가하여, 이 문항을 맞춘 설문조사만 사용함
  - 봇(bot)들이 설문조사에 대리 응답하고, 사례금을 받음
    - 해결 방법: 응답이 얼마나 일관되었는지를 체크함.


A State Capital Knowledge by Education Level
Less than high school
High school graduate
Some college or associate degree
Bachelor's degree
Postgraduate degree
Percentage Correct (%)
B Average Monthly Housing Payment by State
State
Average Monthly Payment ($)
Income
Under $15,000
$15,000 to $24,999
$25,000 to $34,999
$35,000 to $49,999
$50,000 to $74,999
$75,000 to $99,999
$100,000 to $149,999
$150,000 to $199,999
$200,000 or more


B Average Monthly Housing Payment by State
State
Average Monthly Payment ($)
Income
Under $15,000
$15,000 to $24,999
$25,000 to $34,999
$35,000 to $49,999
$50,000 to $74,999
$75,000 to $99,999
$100,000 to $149,999
$150,000 to $199,999
$200,000 or more


C Distribution by Age Group
Hours at Kid's Sports
Number of Children
Age Group
Amount


B
Trump Approval
Strongly approve
Somewhat approve
Neither approve nor disapprove
Somewhat disapprove
Strongly disapprove
0.0%
15.3%
28.7%
0.0%
19.0%
69.7%
0.0%
0.0%
1.0%
0.3%
21.7%
30.0%
1.3%
78.3%
34.7%
0
25
50
75
100
Percentage (%)


Feeling Rating (0–100)
E
2028 Candidate Preference
Someone else
Not sure
The Republican candidate
The Democratic candidate
0.0%
0.0%
1.0%
0.0%
0.0%
4.3%
97.3%
1.0%
38.3%
2.7%
99.0%
56.3%
0
30
60
90
Percentage (%)
Condition
Base Prompt
Pro Republican
Pro Democrat

연구 사례

- Limited Differences in LLM Performance for Social Majorities and Minorities in Survey Synthetic Data (백지윤, 한묘경, 김란우)

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

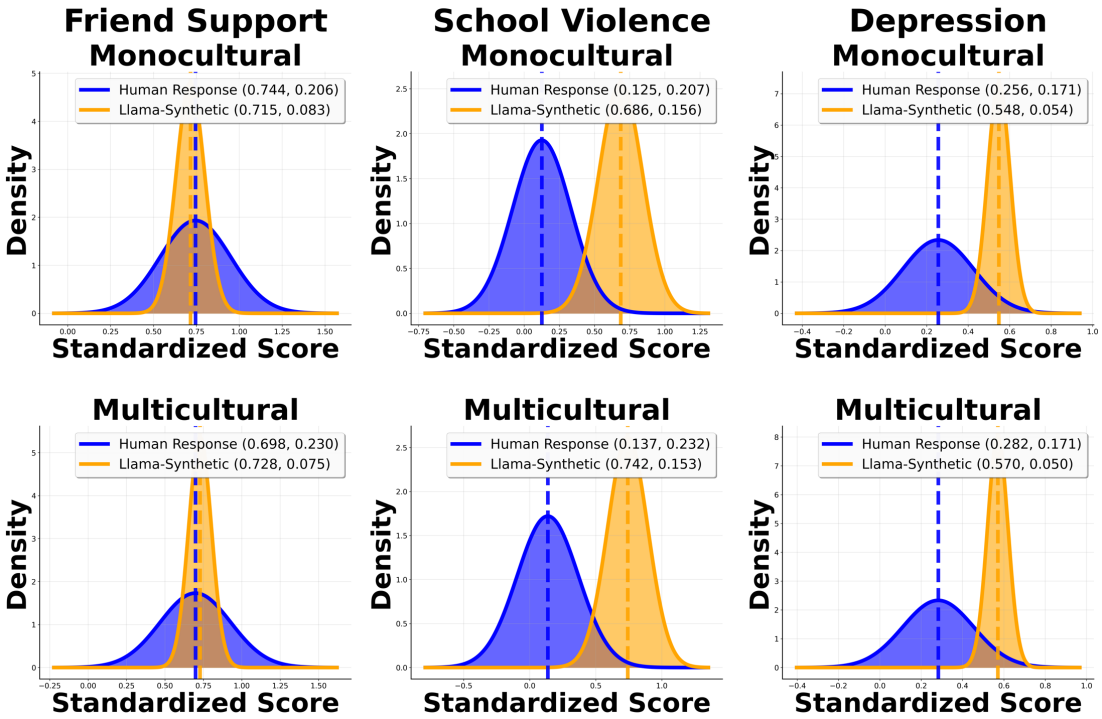
- 연구질문
  - 합성데이터가 사회적 소수자들의 데이터를 더 부정확하게 만들까?

- 연구방법
  - 데이터: 다문화가정 청소년과 일반가정 청소년의 비행실태 비교조사, 2012 (KOSSDA를 통해 접근!)
  - 800명의 다문화가정 청소년 / 800명 비다문화가정 청소년

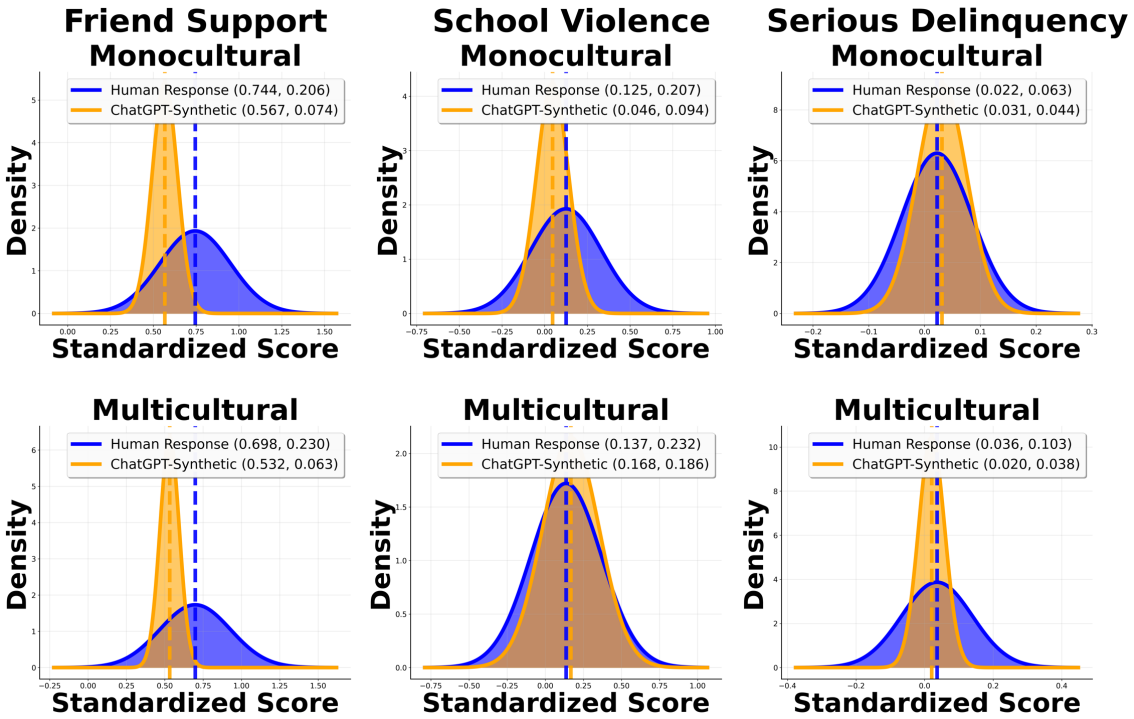
- **페르소나**: 지금은 2012년 7월이다. 너는 13살 여성이며, 중국에서 태어나 대한민국의 수도인 서울에 8년째 살고 있는 다문화 가정 중학교 3학년 학생이다. 어머니의 국적은 중국(조선족)이고, 아버지의 국적은 한국이다. 부모님은 현재 결혼 중이며, 너는 친부모 두 분과 살고 있다. 가족의 소득은 중간 정도이다.
- **질문**: 당신은 당신 스스로가 모든 일에 관심과 흥미가 없다고 생각하시나요?
- **보기**: (a) 전혀 그렇지 않다 (b) 별로 그렇지 않다 (c) 그저 그렇다 (d) 대체로 그렇다 (e) 매우 그렇다
- **제약 조건**: 주어진 보기 안에서 대답해줘. 한국어로만 답변하고 영어는 사용하지 마.
- **명령**: 명시된 페르소나 특성을 지닌 응답자의 관점에서 질문에 답해줘.

- **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.

Llama


Friend Support
Monocultural
Human Response (0.744, 0.206)
Llama-Synthetic (0.715, 0.083)
Density
Standardized Score
School Violence
Monocultural
Human Response (0.125, 0.207)
Llama-Synthetic (0.686, 0.156)
Density
Standardized Score
Depression
Monocultural
Human Response (0.256, 0.171)
Llama-Synthetic (0.548, 0.054)
Density
Standardized Score
Multicultural
Human Response (0.698, 0.230)
Llama-Synthetic (0.728, 0.075)
Density
Standardized Score
Multicultural
Human Response (0.137, 0.232)
Llama-Synthetic (0.742, 0.153)
Density
Standardized Score
Multicultural
Human Response (0.282, 0.171)
Llama-Synthetic (0.570, 0.050)
Density
Standardized Score

ChatGPT


Friend Support
Monocultural
Human Response (0.744, 0.206)
ChatGPT-Synthetic (0.567, 0.074)
Density
Standardized Score
School Violence
Monocultural
Human Response (0.125, 0.207)
ChatGPT-Synthetic (0.046, 0.094)
Density
Standardized Score
Serious Delinquency
Monocultural
Human Response (0.022, 0.063)
ChatGPT-Synthetic (0.031, 0.044)
Density
Standardized Score
Multicultural
Human Response (0.698, 0.230)
ChatGPT-Synthetic (0.532, 0.063)
Density
Standardized Score
Multicultural
Human Response (0.137, 0.232)
ChatGPT-Synthetic (0.168, 0.186)
Density
Standardized Score
Multicultural
Human Response (0.036, 0.103)
ChatGPT-Synthetic (0.020, 0.038)
Density
Standardized Score

LLMs을 활용한 삼성 반도체의 기업 이미지 분석

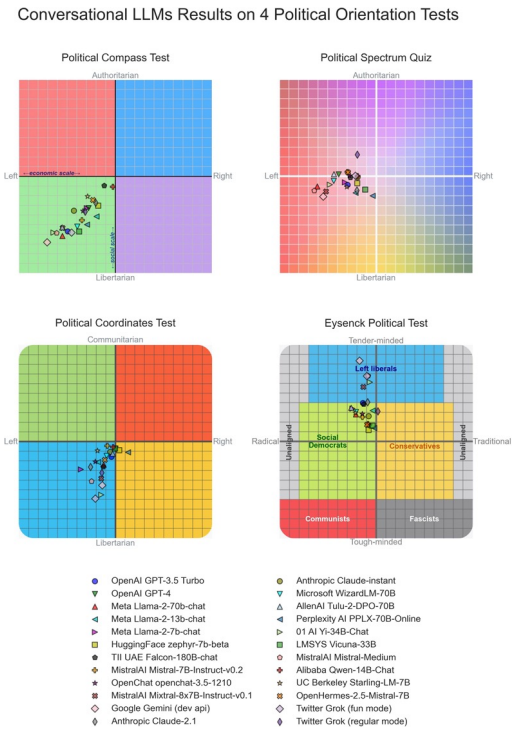
문화기술대학원, 석사과정 박성찬

Related Work

: 유사 선행 연구 (Rozado, 2024)를 벤치마크해서, **"LLMs 기반의 기업 이미지 분석"** 실험을 설계

- 11가지 **정치 성향 테스트**를 사용하여, LLM 정치 성향 평가 -> Likert 응답 (Political Compass Test, Political Spectrum Quiz, …)
- 총 질문 개수는 401개, 모델 개수는 24개, 반복 횟수는 10회

- 활용 및 적용 가능한 **통계적 검증 방법**
  - 정규성 검정: 정규 분포 검증을 위한 Shapiro-Wilk 검정
  - One-sample t-test: LLM 평균 정치 성향이 중립 (0)과 벗어났는지 확인
  - One-way ANOVA: 여러 정치 성향 카테고리 간에 차이가 있는지 확인
  - Tukey HSD: 사후 검증으로 어떤 그룹이 차이가 유의미한지 확인
  - Coefficient of Variation: 동일한 질문에 대해 얼마나 일관된 응답하는지 확인
  - 상관계수, 효과 크기 평가 (Cohen's d, Eta-squared)



Rozado, D. (2024). The political preferences of LLMs. *PloS one, 19*(7), e0306621.

Research Questions

: 글로벌 반도체 기업들에 대해, **LLMs 모델과 질문의 언어에 따라 기업 이미지 평가**가 달라질까?

- **RQ1: LLM 모델 간 기업 이미지 평가 차이 분석**
  - 서로 다른 LLMs은 기업 이미지 평가에서 어떤 차이를 보이는가?
  - 특정 모델이 특정 기업에 대한 선호도를 보이는가?
- **RQ2: 언어별 기업 이미지 평가 차이 분석**
  - 같은 질문을 다른 언어 (한국어, 영어, 중국어)로 하면 어떤 차이를 보이는가?
  - LLMs이 특정 언어 및 문화적 배경을 반영하여 편향을 보이는가?
- **RQ3: 기업 연상 (Free Association) 차이 분석**
  - 특정 기업에 대해 LLMs에 따라 어떤 단어로 연상이 되는가?

Methodology

: 기업 이미지를 다룬 선행 연구에서 자주 반복되는 Davies et al. (2004) 논문의 질문지를 활용

Table 2: The Corporate Character Scale: Dimensions, Facets and Items

Dimension	*Facet*	*Item*
Agreeableness	Warmth	Friendly, pleasant, open, straightforward
	Empathy	Concerned, reassuring, supportive, agreeable
	Integrity	Honest, sincere, trustworthy, socially responsible
Enterprise	Modernity	Cool, trendy, young
	Adventure	Imaginative, up-to-date, exciting, innovative
	Boldness	Extrovert, daring
Competence	Conscientiousness	Reliable, secure, hardworking
	Drive	Ambitious, achievement oriented, leading
	Technocracy	Technical, corporate
Chic	Elegance	Charming, stylish, elegant
	Prestige	Prestigious, exclusive, refined
	Snobbery	Snobby, elitist
Ruthlessness	Egotism	Arrogant, aggressive, selfish
	Dominance	Inward-looking, authoritarian, controlling
Informality	None	Casual, simple, easy-going
Machismo	None	Masculine, tough, rugged

- 여러 연구에서 **반복적으로 등장**하는 내용
- Davies et al. (2004)는 Corporate Character Scale (CCS)를 제안하면서, **기업 이미지를 평가하기 위한 49개 항목을 개발**
- 7가지 차원으로 분류했으며, 기업 이미지가 소비자 및 이해 관계자들에게 어떻게 인식되는지 **Likert scale로 측정** (우호성, 진취성, 전문성, 세련됨, 냉정함, 비공식적, 강인함)
- CCS는 **조직 문화 평가 연구** (Chatman & Jehn, 1994)를 반영하여 도출

Davies, G., Chun, R., da Silva, R. V., & Roper, S. (2004). A corporate character scale to assess employee and customer views of organization reputation. *Corporate reputation review, 7*, 125–146.

Methodology

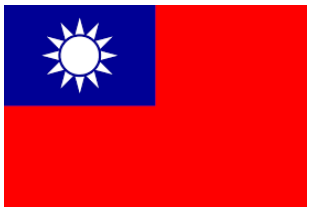
: 기업 이미지를 다룬 선행 연구에서 자주 반복되는 Davies et al. (2004) 논문의 질문지를 활용

차원 (Dimension)	세부 요소 (Facet)	개수	항목 (Item)
친화성 (Agreeableness)	따뜻함 (Warmth)	4	친절한 (Friendly), 유쾌한 (Pleasant), 개방적인 (Open), 솔직한 (Straightforward)
	공감 (Empathy)	4	배려심 있는 (Concerned), 안심시키는 (Reassuring), 지지적인 (Supportive), 우호적인 (Agreeable)
	진실성 (Integrity)	4	정직한 (Honest), 진실한 (Sincere), 믿음직한 (Trustworthy), 사회적으로 책임감 있는 (Socially Responsible)
기업가정신 (Enterprise)	현대성 (Modernity)	3	세련된 (Cool), 트렌디한 (Trendy), 젊고 감각적인 (Young)
	모험심 (Adventure)	4	창의적인 (Imaginative), 최신 트렌드를 반영하는 (Up-to-date), 흥미로운 (Exciting), 혁신적인 (Innovative)
	대담함 (Boldness)	2	외향적인 (Extrovert), 대담한 (Daring)
역량 (Competence)	성실성 (Conscientiousness)	3	신뢰할 수 있는 (Reliable), 안정적인 (Secure), 근면한 (Hardworking)
	추진력 (Drive)	3	야망이 있는 (Ambitious), 목표 지향적인 (Achievement Oriented), 리더십 있는 (Leading)
	기술력 (Technocracy)	2	기술적인 (Technical), 기업적인 (Corporate)
세련됨 (Chic)	우아함 (Elegance)	3	매력적인 (Charming), 스타일리시한 (Stylish), 우아한 (Elegant)
	명성 (Prestige)	3	명망 있는 (Prestigious), 독점적인 (Exclusive), Refined (세련된)
	속물근성 (Snobbery)	2	속물적인 (Snobby), 엘리트주의적인 (Elitist)
무자비함 (Ruthlessness)	자기중심적 (Egotism)	3	오만한 (Arrogant), 공격적인 (Aggressive), 이기적인 (Selfish)
	지배력 (Dominance)	3	내향적인 (Inward-looking), 권위적인 (Authoritarian), 통제적인 (Controlling)
비공식성 (Informality)	-	3	캐주얼한 (Casual), 단순한 (Simple), 자유로운 (Easy-going)
남성성 (Machismo)	-	3	남성적인 (Masculine), 강인한 (Tough), 거친 (Rugged)

Davies, G., Chun, R., da Silva, R. V., & Roper, S. (2004). A corporate character scale to assess employee and customer views of organization reputation. *Corporate reputation review, 7*, 125-146.

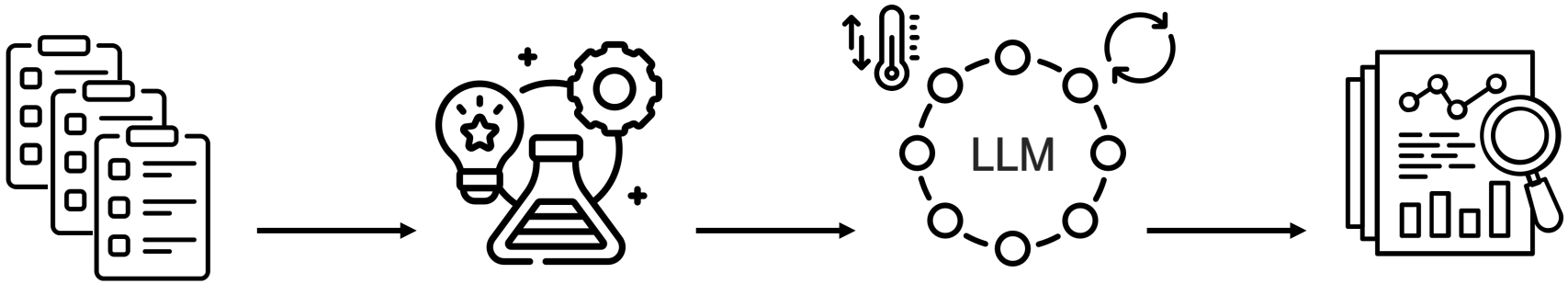
Methodology

: **모델, 기업, 언어**라는 세 가지의 차원에 대해, **기업 이미지**를 LLM으로 분석

			 
모델	Hyper CLOVA X (Naver), EXAONE (LG AI Research)	ChatGPT (OPEN AI), Gemini (Google), LLaMA (Meta)	Qwen (Alibaba), DeepSeek (DeepSeek-AI)
기업	SAMSUNG, SK Hynix	NVIDIA, Intel, Qualcomm, Broadcom, AMD, Micron	TSMC, MediaTek
언어	한국어, 영어, 중국어, 대만어 (간체, 번체) – 모든 모델들이 다국어를 지원하므로, 모든 언어를 사용		

Methodology

: **LLM Evaluation Framework**로 정리하여, 다른 유사 연구에서도 활용 가능하도록 확장



Survey Data Collection

- 선행 연구 수집
- 질문 DB 구축
- 다국어 번역

└ 메타 분석을 위한 PRISMA 프레임워크에서 '논문 검색 & 문항 수집' 과정 차용

Experimental Setup

- 분석 대상 정의 (기업, 모델, 도메인)
- Prompt Engineering
- 입력 구조 및 실험 설계

LLM Response Generation

- 각 LLMs에 질문에 대한 응답 수집
- Robustness 확인을 위하여,
  Temperature 조절, 반복 테스트

└ 응답의 Diversity를 조절하는 변수

Data Analysis

- 모델간 차이 분석
- 언어간 차이 분석
- 편향 및 차이 해석

LLMs 기반 기업 이미지 평가 (Likert 점수 기반 결과 분석)

: 7가지 CCS Dimension 별로, 삼성, NVIDIA, TSMC 기업의 순위 비교 및 그에 따른 기업 이미지 분석

- Agreeable (따뜻한 신뢰감): TSMC 〉 NVIDIA 〉 SAMSUNG
- Enterprise (진취적 혁신성): NVIDIA 〉 SAMSUNG 〉 TSMC
- Competence (유능한 전문성): TSMC 〉 SAMSUNG 〉 NVIDIA
- Chic (세련된 고급스러움): NVIDIA 〉 TSMC 〉 SAMSUNG
- Ruthlessness (냉철한 자기중심성): SAMSUNG 〉 NVIDIA 〉 TSMC
- Informality (자유로운 편안함): NVIDIA 〉 SAMSUNG 〉 TSMC
- Machismo (강인한 남성성): SAMSUNG 〉 TSMC 〉 NVIDIA

SAMSUNG 강하고 냉철한 추진력 중심 기업

NVIDIA 혁신적이고 세련된 유연한 기업

tsmc 따뜻하고 유능한 신뢰형 기업

정리

- LLM을 활용한 설문조사가 점점 더 광범위하게 활용되고 있음.
- 평균값은 잘 맞추지만, 편차는 더 작게 예측하는 특성이 있음.
- LLM 자체가 있는 그대로의 답변을 내놓는다고 예측하기 힘듦.
- 그럼에도 불구하고, 합성데이터의 특성을 활용성을 높이려는 시도들이 이어지고 있음.