NYU

Assessing the Representativeness of LLM-Generated Item Responses Using Latent Class Analysis

Yining Lu, Yuan Huang, Klint Kanopka

Agenda

- **Background & Research Questions**
- **Data Generation**
- **Results**
  a. Research Question 1
  b. Research Question 2
  c. Research Question 3
- **Discussion**

NYU

PART 01

Background

The use of Large Language Models (LLMs) in psychometric research provides a potentially powerful tool for researchers and item developers.

Previous Research:

1. Simulated respondents for item evaluation and test development (Liu et al., 2024)
2. Help address challenges related to question-wording and response bias in survey research (Jansen et al., 2023)
3. Fine-tuned LLMs on repeated cross-sectional surveys enables accurate retrodiction of missing responses and unasked opinion prediction (Kim & Lee, 2024)

PART 01

Background

Despite being trained on data written by humans, some researchers have found deficiencies in LLMs' ability to reproduce human-like response behaviors.

Previous Research:

1. When prompted with specific demographic profiles, both GPT-3.5 and GPT-4 produce responses with poor psychometric properties (Petrov et al., 2024)
2. Popular open and commercial LLMs generally fail to reflect human-like response biases (Tjuatja et al., 2024)

PART 02

Research Questions:

1. Are LLM-generated responses to social survey items consistent and stable?
2. How well do LLMs reproduce observed response distributions for social survey items?
3. Do LLM-generated responses to social survey items reproduce the latent class from observed responses?

PART 03

Data Generation

Data

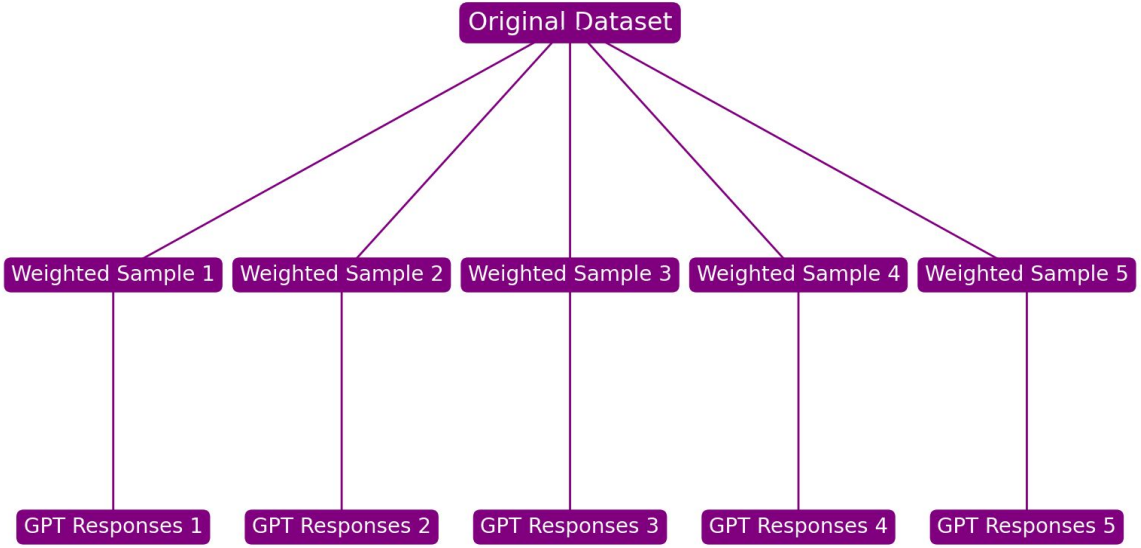
Original Data (3,616 respondents):

2022 cross-sectional portion of General Social Survey (GSS), focusing on government spending attitudes and associated demographics (e.g., age, sex, region, education level, race, income).

All Data we used:

From 3,616 respondents, we used the GSS-provided sampling weight variable (WTSSNRPSAS) to create five sample dataset of 25,000 respondents. Using GPT-4o, we generate synthetic responses via zero-shot prompts based on each respondent's demographic profile.

Data


Original Dataset
Weighted Sample 1
Weighted Sample 2
Weighted Sample 3
Weighted Sample 4
Weighted Sample 5
GPT Responses 1
GPT Responses 2
GPT Responses 3
GPT Responses 4
GPT Responses 5

Questionnaire

NORC at the University of Chicago and the National Science Foundation have done the General Social Survey (GSS) for more than 50 years to keep a historical record of the concerns, experiences, attitudes and practices of Americans. Your household was randomly selected from addresses across the nation for participation. The information collected is used by policy makers, scientific researchers, government officials and students to better understand Americans and better meet their changing needs. The questions we'll ask have to do with your opinions and knowledge on a variety of important topics like work, family, government, science, and health. Some topics may be sensitive for you, and you can decline to answer any question. Your participation does not involve any risks other than what you would encounter in daily life. Most participants find the survey to be interesting with a chance to think about things that matter to them. Which questions are asked depends partly on chance, and partly on your answers to other questions. The interview length varies for different households; for most people, it usually takes about 90 minutes. You may also be contacted in the future to participate in a future round of the General Social Survey or another study established by the National Science Foundation. Participation in this round of the survey does not obligate you to participate in any future rounds of the General Social Survey or any other surveys. Any future participation on your part will again require us to gain your consent.

Let's begin with some things people think about today. We are faced with many problems in this country, none of which can be solved easily or inexpensively. For each of the following problems, please indicate whether you think we're spending too much money on it, too little money, or about the right amount.

... are we spending too much, too little, or about the right amount on The space exploration program?
Same as following: Improving and protecting the environment? Improving and protecting the nation's health? Improving the nation's education system? Welfare? Social Security? Highways and bridges? Assistance for childcare? Supporting scientific research?

Prompting

Prompting Sample 1:

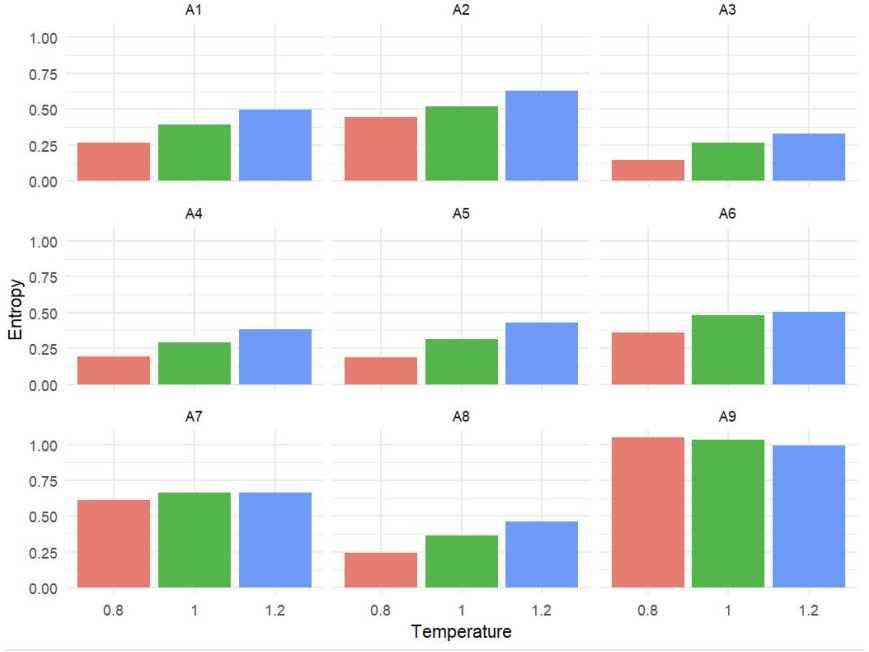
"Respond to the following survey questions as if you are a 38 year old in the year of 2022, and you are a Never married Black male who lives in A suburb of a large central city in the West South Central. You are in Good health and you are Protestant. Your highest degree is High school, you work in the Other telecommunications services industry, and earn $25,000 or more dollars per year. You identify as Independent political party and think of yourself as a Moderate, middle of the road. For each question, only answer with one of the following options: [too much, too little, about the right amount]. You must answer all questions, and no options outside of those are valid answers (No NA, refusals to answer, etc...). No preambles, and answer like 1. <answer> 2. <answer> 3. <answer>..."

Prompting Sample 2:

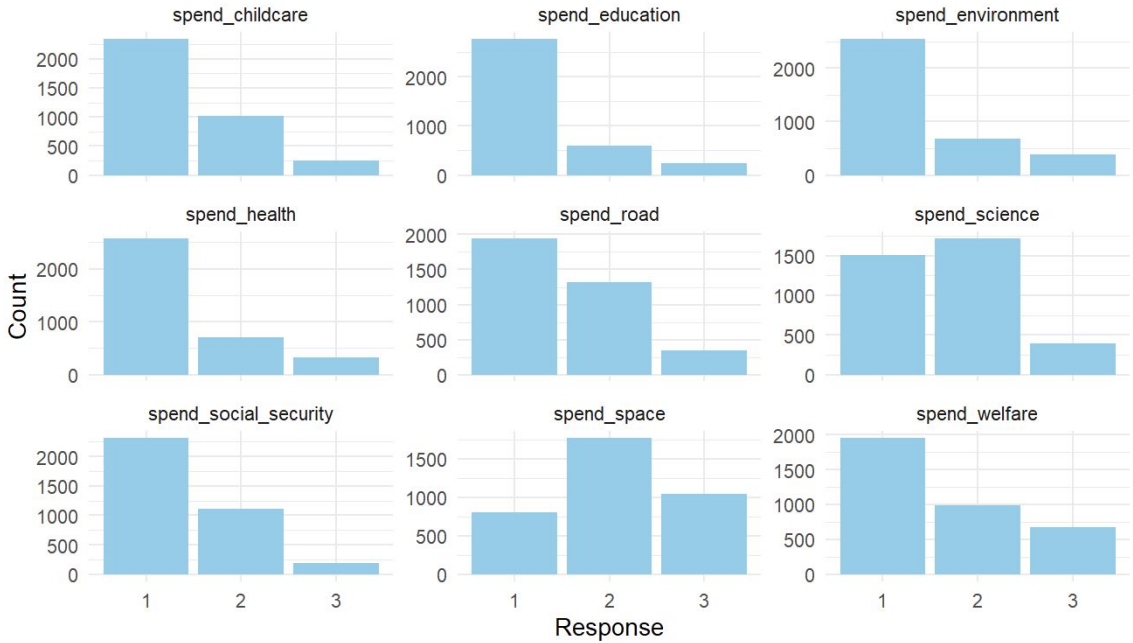
"Respond to the following survey questions as if you are a 68 year old in the year of 2022, and you are a Married White female who lives in a small city in the East North Central. You are in Good health and you are Protestant. Your highest degree is High school, you work in the Banking and related activities industry, and earn $25,000 or more dollars per year. You identify as Strong democrat political party and think of yourself as a Liberal. For each question, only answer with one of the following options: [too much, too little, about the right amount]. You must answer all questions, and no options outside of those are valid answers (No NA, refusals to answer, etc...). No preambles, and answer like 1. <answer> 2. <answer> 3. <answer>..."

Temperature Selection

entropy of responses by temperature



Real Data Response Prevalence by Questions



PART 03

Research Question 1

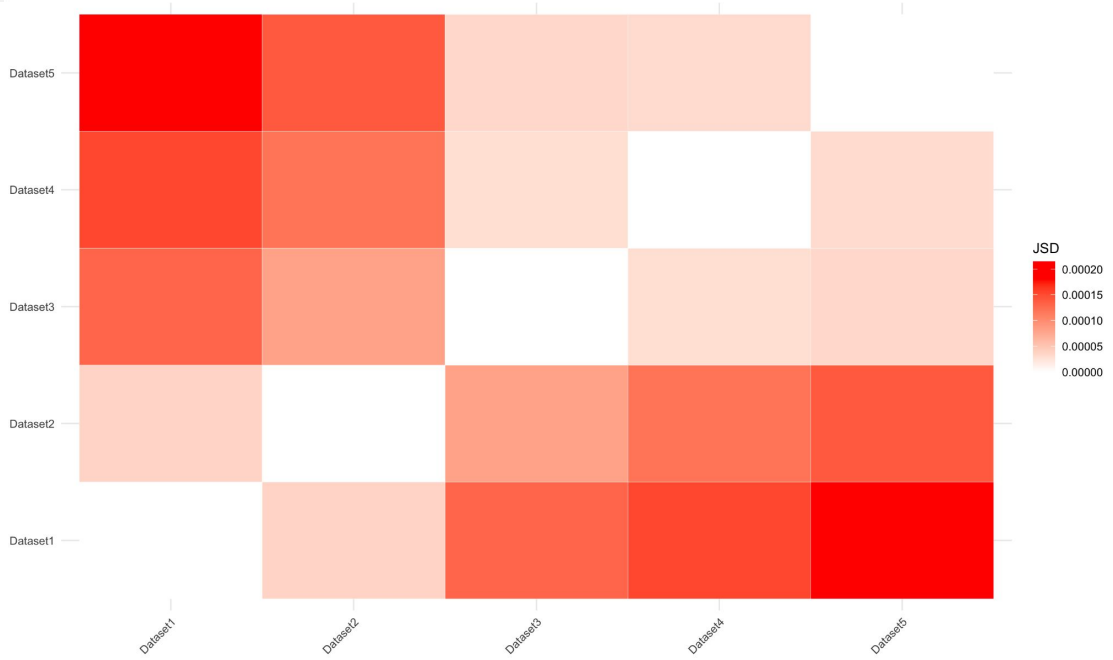
Are LLM-generated responses to social survey items consistent and stable?

Methods

- **Jenson-Shannon Divergence**: compare the overall response distributions, with 0 indicating identical distributions
- **Entropy**: a measure of response diversity or unpredictability within a sample
- **Latent structure**: latent class analysis (LCA) for class count and class alignment

Consistent and Stable Responses across All Samples

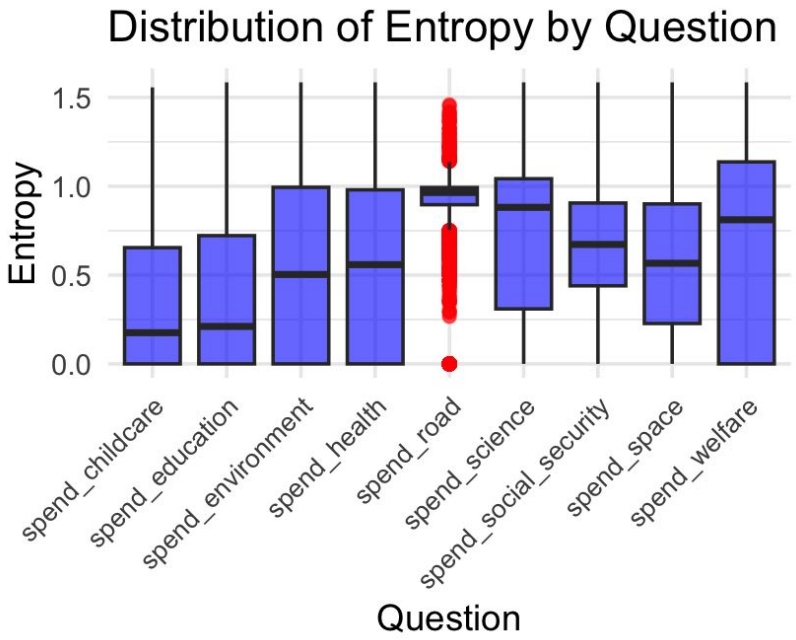
Jensen-Shannon Divergence Heatmap



NYU

Consistent and Stable Responses across All Samples

Distribution of Entropy by Question



Class alignment between GPT samples after LCA

Sample 1(25,000) vs Sample 2(25,000)

	Too little	About the Right	Too Much
Space	0.2438	0.0832	**0.673**
Environment	**0.9938**	0.0062	0
Health	**0.975**	0.025	0
Education	**0.9971**	0.0029	0
Welfare	0.4214	**0.5001**	0.0785
Social Security	0.415	**0.5847**	0.0003
Road	0.4767	**0.522**	0.0013
Childcare	**0.9951**	0.0049	0
Science	**0.9007**	0.099	0.0003

Class 1 in Sample 1

	Too little	About the Right	Too Much
Space	0.2466	0.0842	**0.6692**
Environment	**0.9918**	0.0082	0
Health	**0.9754**	0.0246	0
Education	**0.9976**	0.0022	0.0001
Welfare	0.4214	**0.5005**	0.0781
Social Security	0.4202	**0.5794**	0.0004
Road	0.4618	**0.537**	0.0012
Childcare	**0.9966**	0.0034	0
Science	**0.9016**	0.0972	0.0012

Class 2 in Sample 2

Class alignment between GPT samples after LCA

Sample 1(25,000) vs Sample 2(25,000)

	Too little	About the Right Amount	Too Much
Space	0.0146	0.0959	**0.8895**
Environment	**0.6941**	0.2909	0.015
Health	**0.7018**	0.2919	0.0064
Education	**0.8786**	0.1214	0
Welfare	0.0031	0.1378	**0.8591**
Social Security	0.0717	**0.9189**	0.0094
Road	0.4909	**0.5086**	0.0004
Childcare	**0.7992**	0.104	0.0968
Science	0.3132	**0.6083**	0.0784

Class 2 in Sample 1

	Too little	About the Right Amount	Too Much
Space	0.0134	0.0957	**0.8908**
Environment	**0.7011**	0.2868	0.0122
Health	**0.7052**	0.2891	0.0057
Education	**0.884**	0.116	0
Welfare	0.0035	0.1434	**0.8531**
Social Security	0.0733	**0.918**	0.0086
Road	**0.5049**	0.4946	0.0005
Childcare	**0.7976**	0.1101	0.0923
Science	0.3245	**0.597**	0.0785

Class 1 in Sample 2

Class alignment between GPT samples after LCA

Sample 1(25,000) vs Sample 2(25,000)

	Too little	About the Right Amount	Too Much
Space	0.0018	0.0358	**0.9623**
Environment	0.2016	0.2341	**0.5642**
Health	0.2721	**0.3745**	0.3534
Education	**0.4493**	0.3106	0.2401
Welfare	0.0004	0.0004	**0.9991**
Social Security	0.1026	**0.7942**	0.1033
Road	0.3946	**0.5976**	0.0078
Childcare	0.164	0.0339	**0.8022**
Science	0.1013	0.1746	**0.7241**

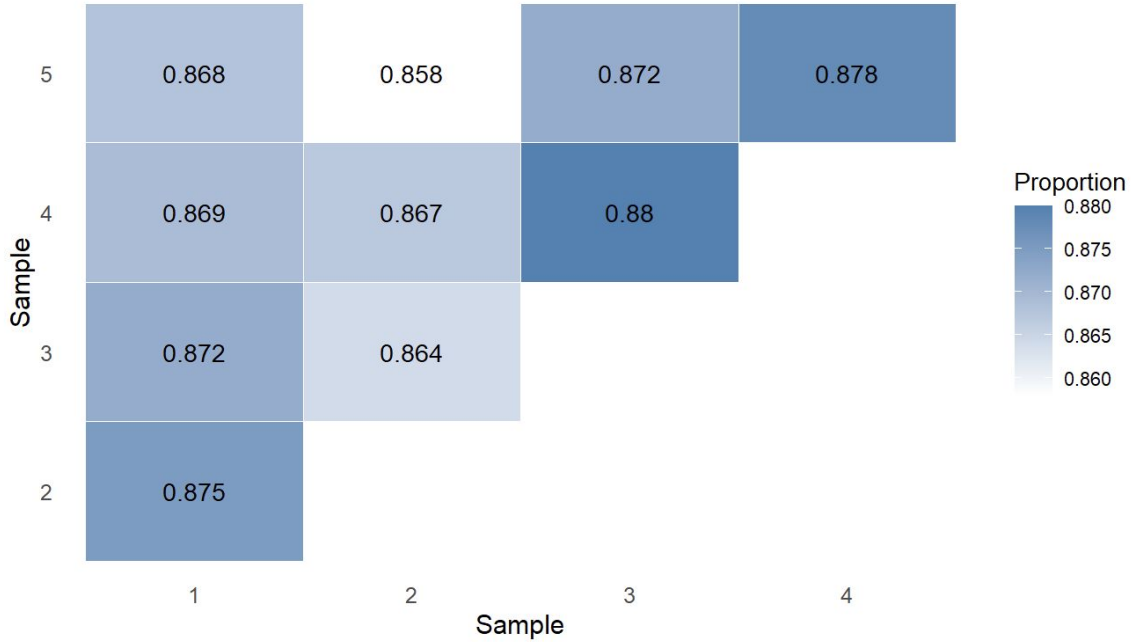
Class 3 in Sample 1

	Too little	About the Right Amount	Too Much
Space	0.0023	0.0333	**0.9644**
Environment	0.1999	0.2348	**0.5653**
Health	0.2704	**0.3695**	0.3601
Education	**0.4667**	0.3088	0.2245
Welfare	0.0003	0.0003	**0.9994**
Social Security	0.0953	**0.8147**	0.09
Road	0.3929	**0.6015**	0.0056
Childcare	0.156	0.0364	**0.8076**
Science	0.099	0.1627	**0.7383**

Class 3 in Sample 2

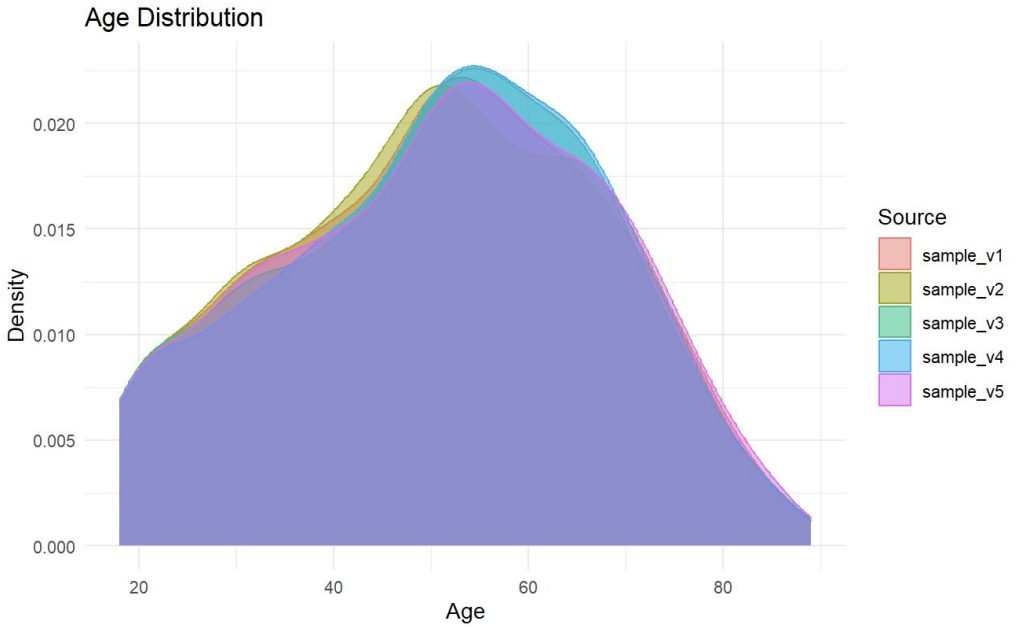
Class alignment across GPT samples

Aligned Class Match Proportion Heatmap



Are the same class formed by the same individuals across GPT samples?

One example of demographics



PART 04

Research Question 2

How well do LLMs reproduce observed response distributions for social survey items?

Methods

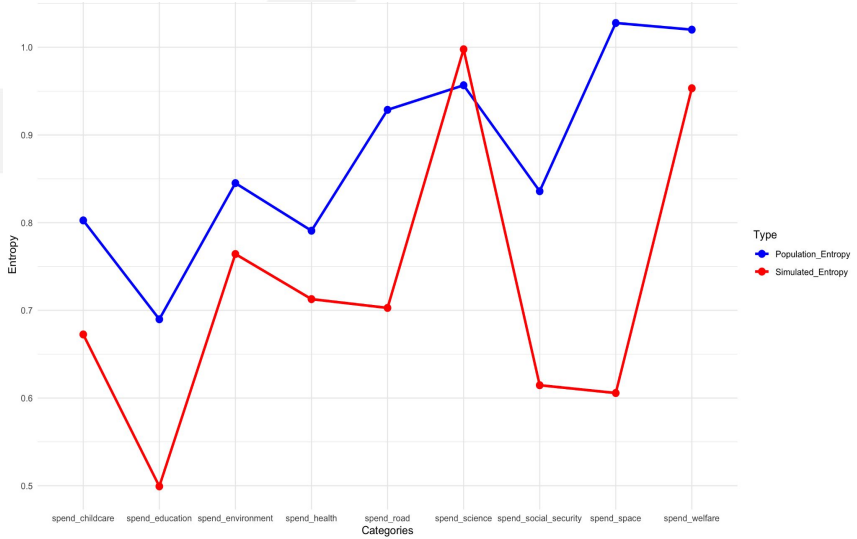
- **Response similarity**:
  Polychoric correlation: assess the strength of association between ordinal response patterns
  Cohen's Kappa and Quadratic Weighted Kappa: quantify agreement between real and synthetic categorical responses

- **Jenson-Shannon Divergence**: compare the overall response distributions between simulated responses and real responses, with 0 indicating identical distributions

PART 04

Alignment of GPT simulated response and real response

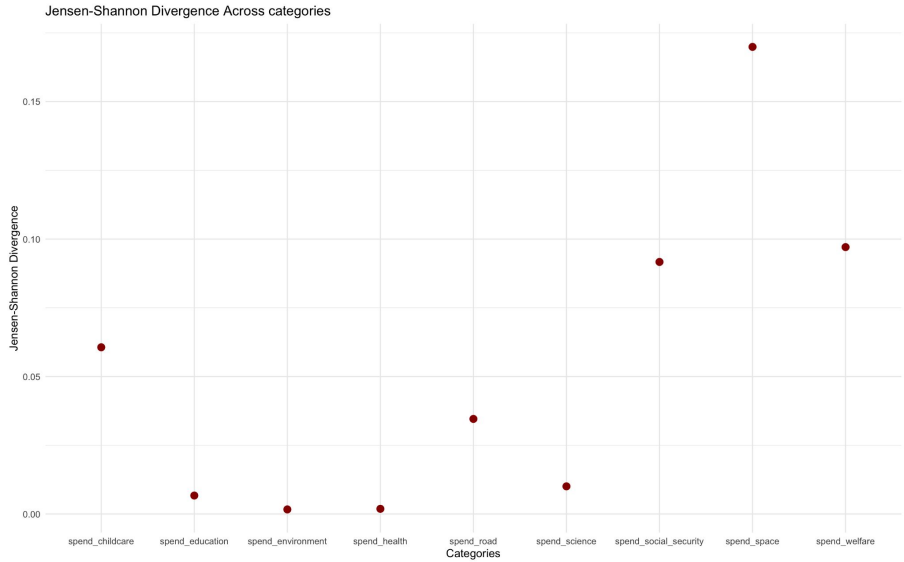
Group simulated responses as a whole dataset and compare with real dataset



NYU

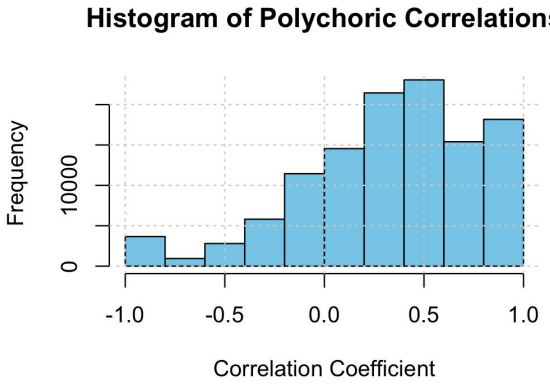
Comparison of the Entropy

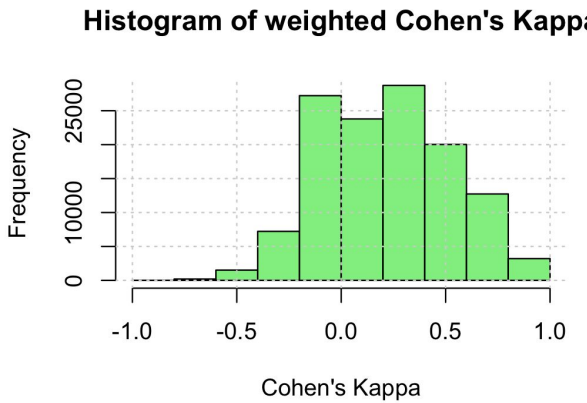
Jensen-Shannon Divergence Across Categories

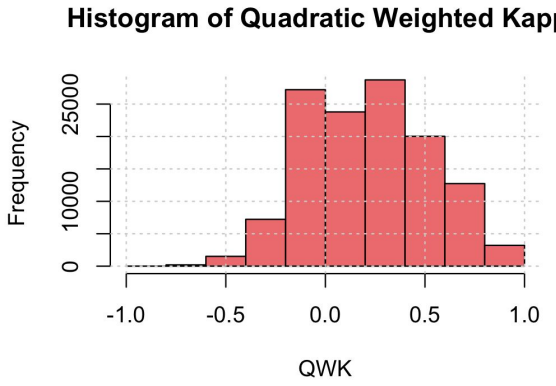


Alignment of GPT simulated response and real response

Group simulated responses as a whole dataset and compare with real dataset



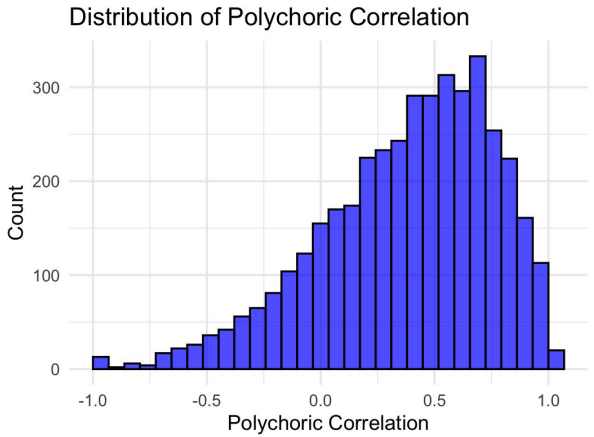


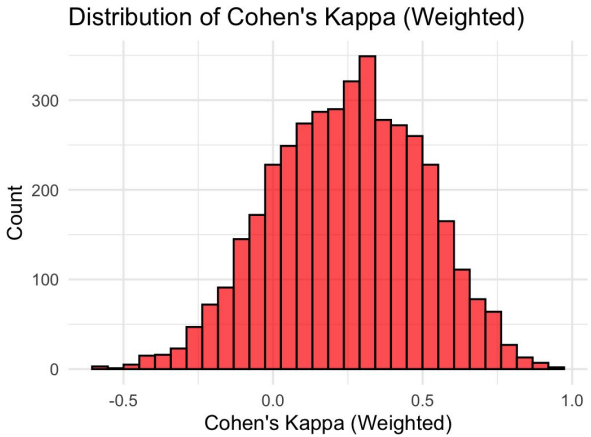


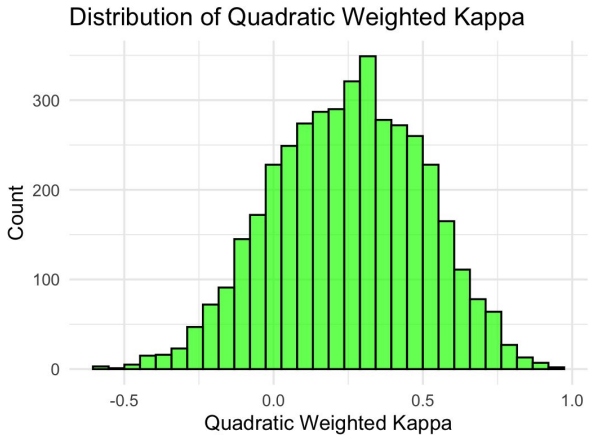
NYU

Alignment of GPT simulated response and real response

By person, across all simulated respondents, but drawn from the same person, one per question







NYU

PART 05

Research Question 3

Do LLM-generated responses to social survey items reproduce the latent class from observed responses?

Methods

Latent structure:

Compare the results of latent class analysis (LCA) for real data and GPT simulated data

Real Data LCA classes

We identify Three classes in 25,000 weighted Samples

Class 1 (extreme class): Always choose “too much” or “too little” on spending

Class 2 (moderate class): Mostly choose “about the right amount” on spending

Class 3 (conservative class): Mostly choose “too much” on spending

GPT-simulated Data LCA classes

We identify Three classes in GPT simulated responses

Class 1 (moderate class): Mostly choose “too little” on education, childcare, health, and science, but more balanced on other spending areas

Class 2 (social support class): Strongly support too little spending on environment, social security, childcare, and science

Class 3 (science and education class): Strongly favor more spending on education, health, childcare, and science, but less emphasis on environment

Class alignment between population and GPT

Real(25,000) vs Sample 1(25,000)

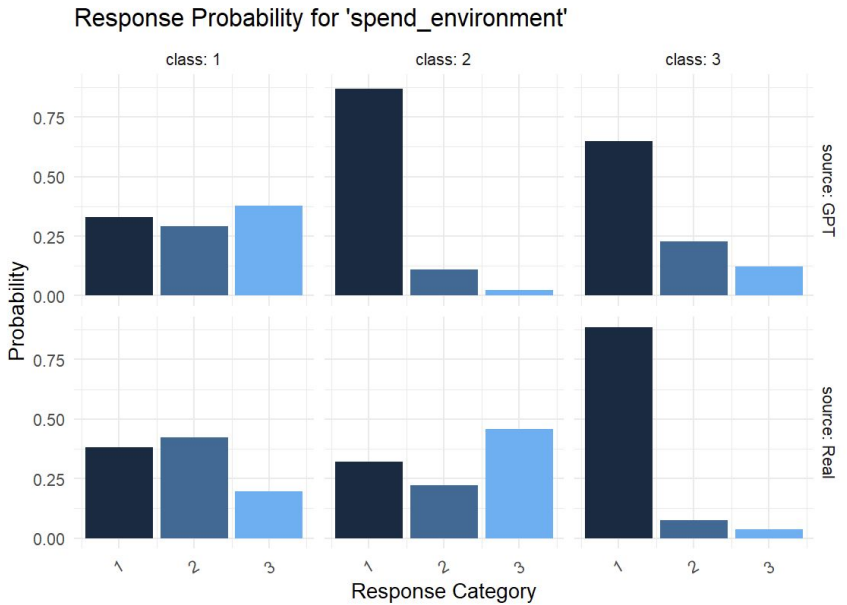
	GPT_class_1	GPT_class_2	GPT_class_3
Real_class_1	1671	1521	4408
Real_class_2	839	285	1253
Real_class_3	974	6925	7124

Class alignment between population and GPT

Real(25,000) vs Sample 1(25,000)

One example question

No aligned class



NYU

PART 06

Discussion

1. **(RQ1) Consistency in LLM-Generated Data**
   - Across multiple samples, GPT produced stable and internally consistent responses for our survey items.
   - Although we set the temperature to 1 which should have variability, the GPT-generated responses remain remarkably consistent, which may reveal that GPT has a stable internal logic when answering social survey items.

2. **(RQ2) Individual-Level Variations**
   - Despite overall consistency, there is still individual-level variability.
   - This variance indicates that while the sample-level distribution is stable, responses for the same hypothetical person can shift between model queries.

3. (RQ3) Latent Class Stability

- LCA on GPT-generated samples consistently returned three classes, mirroring the number of classes found in the real-world data.
- However, the profiles of these latent classes differ from those of actual human respondents.

4. (RQ3) Misalignment with Real Population

- GPT's latent class patterns do not match real population classes.
- This mismatch highlights that although GPT can replicate a plausible distribution of answers, it does not necessarily capture underlying population structure.

5. Implications and Cautions

- While LLMs may serve as convenient synthetic data generators, their output should be interpreted with caution—especially where realistic population inferences are critical.
- Piloting and validation against real data are essential before deploying LLM-simulated responses for high-stakes or policy-relevant decisions.

6. Future Directions

- Refine prompts or use few-/multi-shot prompting to improve alignment with human data.
- Investigate whether fine-tuning can bridge the gap between GPT-generated classes and real population classes.
- Examine external validity by comparing GPT's performance on different questions.

References

- Jansen, B. J., Jung, S. G., & Salminen, J. (2023). Employing large language models in survey research. *Natural Language Processing Journal*, *4*, 100020.
- Kim, J., & Lee, B. (2024). Ai-augmented surveys: Leveraging large language models and surveys for opinion prediction. https://arxiv.org/abs/2305.09620
- Liu, Y., Bhandari, S., & Pardos, Z. A. (2024). Leveraging LLM-Respondents for Item Evaluation: a Psychometric Analysis. *arXiv preprint arXiv:2407.10899*.
- Petrov, N. B., Serapio-García, G., & Rentfrow, J. (2024). Limited Ability of LLMs to Simulate Human Psychological Behaviours: a Psychometric Analysis. *arXiv preprint arXiv:2405.07248*.
- Tjuatja, L., Chen, V., Wu, T., Talwalkwar, A., & Neubig, G. (2024). Do llms exhibit human-like response biases? a case study in survey design. *Transactions of the Association for Computational Linguistics*, *12*, 1011-1026.

Thank you!

Yining Lu: yl11897@nyu.edu

Yuan Huang: yh2741@nyu.edu

Klint Kanopka: klint.kanopka@nyu.edu

NYU

Appendix

Prompting

You will respond to the survey questions by giving only a single answer to each question. The information collected is used by policy makers, scientific researchers, government officials and students to better understand Americans and better meet their changing needs. The questions we'll ask have to do with your opinions and knowledge on a variety of important topics like work, family, government, science, and health. Some topics may be sensitive for you, and you can decline to answer any question

Let's begin with some things people think about today. We are faced with many problems in this country, none of which can be solved easily or inexpensively. For each of the following problems, please indicate whether you think we're spending too much money on it, too little money, or about the right amount.

1. are we spending too much, too little, or about the right amount on Space exploration?
2. are we spending too much, too little, or about the right amount on Improving and protecting the environment?
3. are we spending too much, too little, or about the right amount on Improving and protecting the nation's health?
4. are we spending too much, too little, or about the right amount on Improving the nation's education system?
5. are we spending too much, too little, or about the right amount on Welfare?
6. are we spending too much, too little, or about the right amount on Social Security?
7. are we spending too much, too little, or about the right amount on Highways and bridges?
8. are we spending too much, too little, or about the right amount on Assistance for childcare?
9. are we spending too much, too little, or about the right amount on Supporting scientific research?"

"**Respond to the following survey questions as if you are a",**

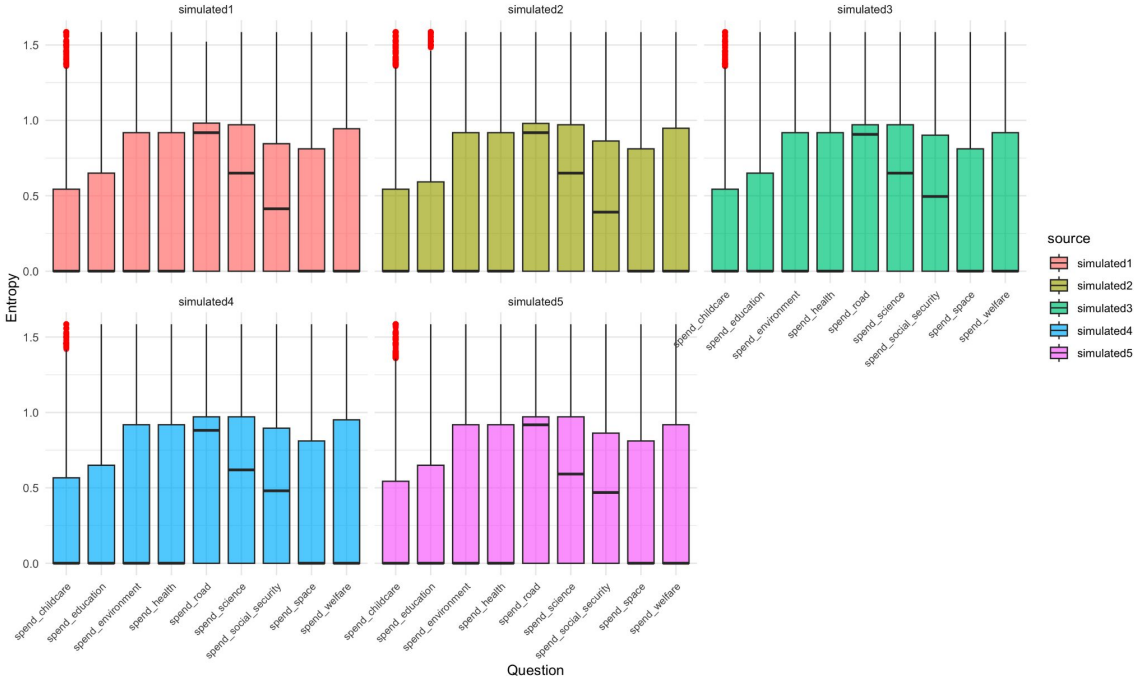
age_part, marital_part, race_part, gender_part, location_part, ".", health_part, religious_part, ".", degree_part, ",", industry_part, ",", income_part, ".", political_part, polviews_part, ".",

"For each question, only answer with one of the following options: [too much, too little, about the right amount]. You must answer all questions, and no options outside of those are valid answers (No NA, refusals to answer, etc...). No preambles, and answer like 1. <answer> 2. <answer> 3. <answer> ..."

Appendix

Consistent and Stable responses across all samples

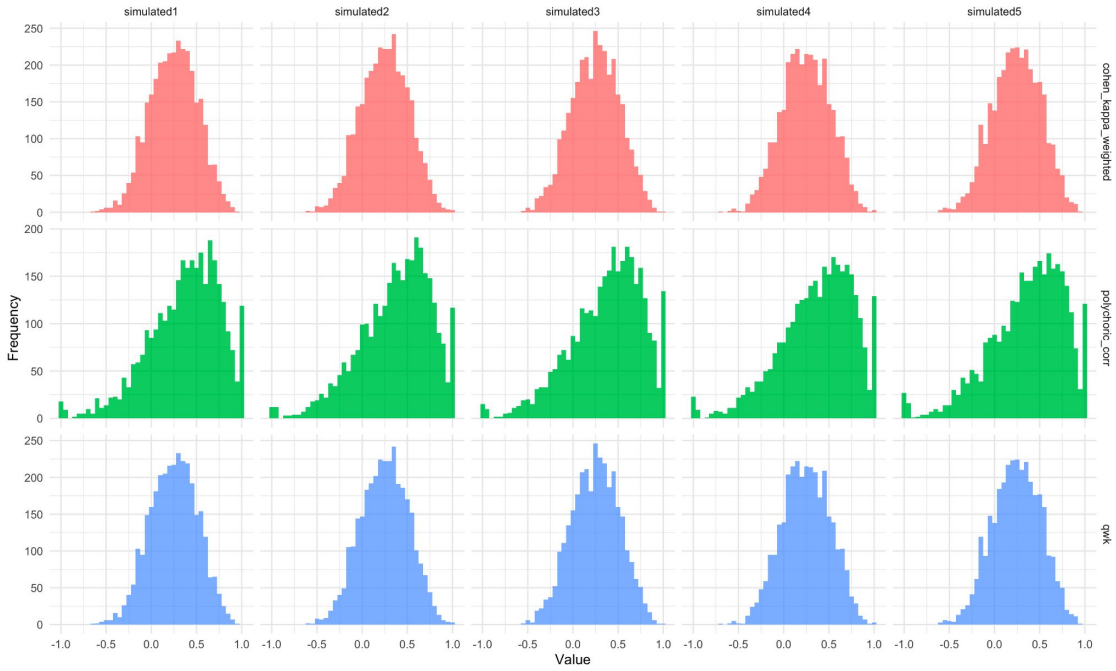
Distribution of Entropy by Question and Source



Appendix

Consistent and Stable responses across all samples

Histogram of Metrics by Source

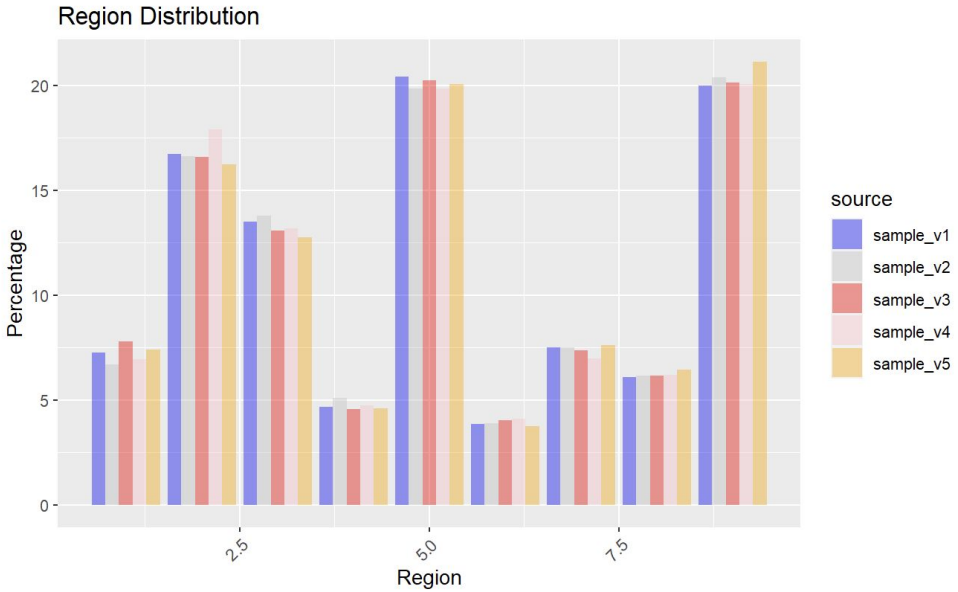


Appendix
Real Data Latent Classes

real data (25000)			
	class 1	class 2	class 3 (radical)
space	3 (53%)	2 (72%)	2 (45%)
environment	3 (41% and 38% chosse 1)	1 (42% and 39% choose 2)	1 (87%)
health	1 (47%)	1 (44%, almost equal with 2)	1 (88%)
education	1 (56%)	1 (50%)	1 (few choose 3)
welfare	3 (55%)	2 (51%)	1 (72%)
social security	1 (58%)	2 (53%)	1 (79%)
road	1 (53%)	2 (52%, few choose 3)	1 (57%)
childcare	1 (54%)	2 (59%)	1 (84%)
science	3 (49%)	2 (76%)	1 (56%)

Appendix

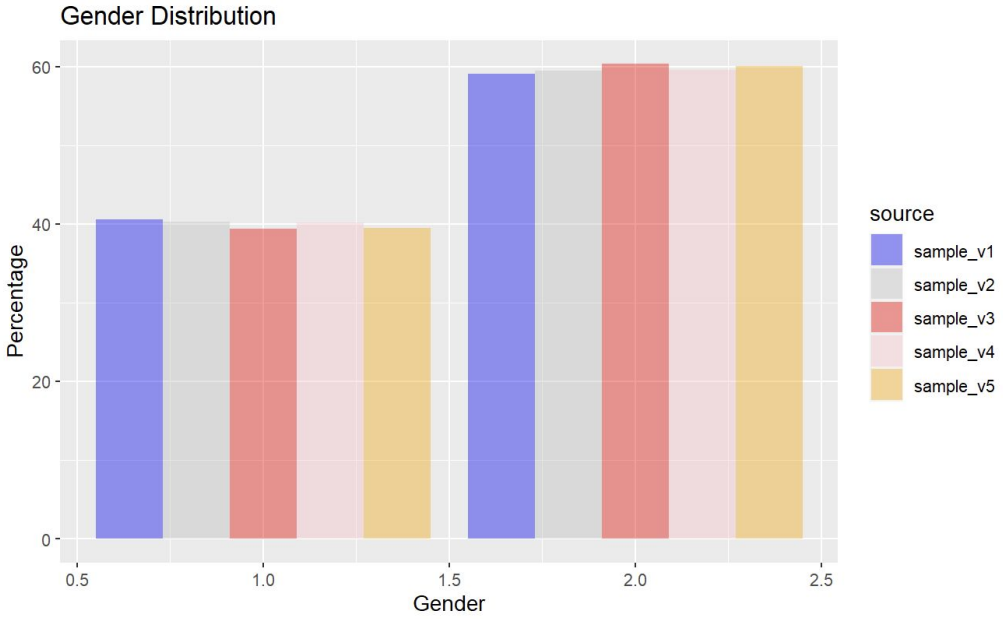
Are the same class formed by the same individuals across GPT samples?



NYU

Appendix

Are the same class formed by the same individuals across GPT samples?



NYU