Evaluating the Efficacy of LLM-augmented Imputation in Longitudinal Surveys

Andrew C. Forrester Srijeeta Mitra Ujjayini Das

Joint Program in Survey Methodology
University of Maryland, College Park

May 15, 2025

Outline

Background

Data

Methodology & Results

Discussion

Background (1 of 2)

- Attrition in longitudinal surveys leads to non-response (Groves et al., 2012) affecting data quality
- Statistical measures combating nonresponse includes weighting (unit nonresponse) and imputation (item nonresponse) (Little & Rubin, 2019)
- Complex time-dependent structure of longitudinal data requires sophisticated techniques such as machine/deep learning algorithms (Random Forest (Stekhoven & Bühlmann, 2012), Recurrent Neural Network (Lipton et al., 2016))

Background (2 of 2)

- Recent surge in Gen AI opened new possibilities for survey research (Jansen et al., 2023)
- Generating synthetic survey responses using AI has gained attention
  - Imputing item missingness in surveys (Budhwani et al., 2023)
  - Generating human-like "personas" to mimic survey responses using LLMs (Khaokaew et al., 2025; Kim & Lee, 2024; Schuller et al., 2024; Von Der Heyde et al., 2025)
- Findings are well documented for:
  - Predicting **subjective** opinion and attitudes
  - **Cross-sectional** and **repeated cross-sectional** surveys

Research Questions

1. How well do LLM-augmented imputations work in the case of objective economic measures?
2. How well do LLMs can predict responses in subsequent waves of longitudinal surveys given the response from prior wave(s)?
3. How aligned are the predictions with those from traditional imputation methods?

Data

Survey of Income and Program Participation (SIPP) Data (2014):

- Longitudinal study conducted by U.S. Census Bureau
  - Assesses Americans' economic well-being and participation in federally-administered programs (e.g., income and assets, labor force participation, and government program usage etc.).
  - Provides flag variables to designate the values that are imputed and the method used to impute them.
- Collects *monthly* data from individuals over the course of the year with "waves" representing a calendar year:
  - Wave 1: January-December 2013
  - Wave 2: January-December 2014
  - Wave 3: January-December 2015
  - Wave 4: January-December 2016

Methodology

- **Goal:** Predicting participation in the Supplemental Nutrition Assistance Program (SNAP) based on data from the 2014 Survey of Income and Program Participation (SIPP) using Open AI's GPT-4o
  - **Outcome:** Did the respondent participate in SNAP over the course of a year? Yes/no
  - Use GPT-4o to produce *model-based* predictions of SNAP participation
- **Scenarios:**
  1. Cross-sectional Imputation: using only current wave information (RQ1)
  2. Longitudinal Imputation: incorporating current and previous respondent information (RQ2)

Persona Generation

- Synthetic personas generated using 2014 SIPP data focused on the voting-age population (18+ years).
- Sociodemographic, economic and geographic variables are used
- Personas mimic real survey respondents from the first wave of SIPP
- Inspired by role-play prompting; effective for eliciting realistic human decision-making and improving zero-shot reasoning in LLMs.

Role-play Prompting: Imagine you are a `age` year old `marital status` `race and ethnicity` `sex` with a `education` degree residing in a `metropolitan status` area of `state of residence`. You are `nativity` and have `presence of children`.

Contextual Variables

Demographic	**Economic**	**Geographic**
Age	Employment Status	State of Residence
Sex	Monthly Income	Metro/non-metro
Race/Ethnicity	Education	
Nativity		
Marital Status		
Presence of Children		

Variables Used for Persona Generation

Cross Sectional Imputations Using LLM

Training the LLM:

- Developed a training data set on SNAP participation, from the first wave of 2014 SIPP which included a random 80% subset of the respondents from that wave who had non-missing values for SNAP participation.
- Fed this to ChatGPT and ensured that ChatGPT learns the survey responses to the SNAP participation for the generated personas as part of its training process.

Prompt for training: `Imagine you are a Survey Respondent. You will be provided with a data set that contains information on socio-demographic variables of people and whether they have participated in the Supplemental Nutrition Assistance Program (SNAP) run by the United States Federal Government. Refer to this as ‘learning data’.`

Cross Sectional Imputations Using LLM

Testing LLM Model:

- Use the 20% holdout set (without response) and ask ChatGPT to predict the SNAP participation of each of the records based on what it learned from the training data that includes the structured personas.
- **Prompt for testing:** `Based on what you have learned from the ‘learning data’ regarding how different personas are related to participation in SNAP, predict the same for the data you will be given now. Please output a CSV with the predictions.`

Wave 1 Holdout Set Predictions

False positive rate relatively high when predicting holdout non-responses in Wave 1 alone

- Comparisons to “ground truth” responses not imputed by the Census Bureau

Actual / Predicted	**No**	**Yes**
No	7,670 (0.720)	1,766 (0.166)
Yes	304 (0.0285)	915 (0.0859)

Confusion Matrix on Wave 1 20% Test Dataset

Longitudinal Imputations Using LLM

- Assessing whether incorporating prior-wave responses improves the LLM's ability to predict SNAP participation in future waves.
- Restricted to cases that have SNAP participation reported (non-imputed) across waves

Predictions in Subsequent Waves

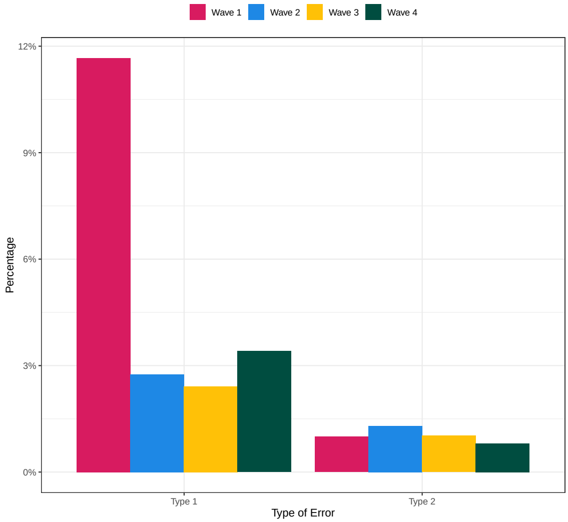
Model predictive power lower with a single wave data and improves as it receives additional waves

- Increased true negatives over waves
- Decreased false positives and false negatives; even if not significantly
- No improvement of true positive rates over waves

		Actual No		**Actual Yes**	
Trained on	**Predicted**	No	Yes	No	Yes
No prior info	**Wave 1**	31309 (73.5%)	6453 (15.1%)	550 (1.3%)	4308 (10.1%)
Wave 1	**Wave 2**	27394 (85.8%)	915 (2.9%)	430 (1.3%)	3179 (10.0%)
Waves 1-2	**Wave 3**	19078 (87.2%)	532 (2.4%)	226 (1.0%)	2051 (9.4%)
Waves 1-3	**Wave 4**	13617 (86.6%)	537 (3.4%)	127 (0.8%)	1437 (9.1%)

Confusion Matrices with Percentages for SNAP Participation Predictions

Types of Error by Wave



Predicting Imputed Responses

- Fed the imputed subset of SNAP participants in Wave 1 and ask ChatGPT to predict the participation value.
- No ground truth available— compared alignment of ChatGPT's predictions to SIPP's imputed values using Spearman's rank correlation coefficient.

Comparing LLM Predictions to Census Bureau Imputations

- Among respondents with imputed values in Wave 1, LLM predictions aligned most with imputed "No's"
- High non-match rate at 26.2% with imputed "No's"
- Low association between the LLM and Census Bureau imputations ($\Phi = \mathbf{0.12}$)

Imputed / Predicted	**No**	**Yes**
No	1,125 (0.546)	540 (0.262)
Yes	209 (0.101)	186 (0.0903)

Confusion Matrix for Respondents with Imputed Values

Discussion and Next Steps

1. **Using LLMs for imputation provides another viable method for imputing item non-response.**
    - Predicting SNAP coverage in SIPP using GPT-4o on 80% subsampled training data provided results consistent with ground truth data
    - Predictive ability improves when using additional longitudinal context
2. **Still unclear whether LLMs outperform traditional multiple imputation methods.**
    - Weak agreement with Census Bureau imputation methods highlights “black box” nature of LLMs
    - LLMs rely on fundamentally different training data (internet text v. researcher experience)
3. **Next steps:** Improve model prompting with additional SNAP-specific context and previous prediction errors

References I

Budhwani, A., Lin, T., Feng, D., & Bachmann, C. (2023).Assessing and Comparing Data Imputation Techniques for Item Nonresponse in Household Travel Surveys [Publisher: SAGE Publications Inc]. *Transportation Research Record*, *2677*(1), 1404–1417.

Groves, R. M., Couper, M., & Couper, M. P. (2012). *Nonresponse in household interview surveys*. Wiley.

Jansen, B. J., Jung, S.-g., & Salminen, J. (2023).Employing large language models in survey research. *Natural Language Processing Journal*, *4*, 100020.

Khaokaew, Y., Salim, F. D., Züfle, A., Xue, H., Anderson, T., MacIntyre, C. R., Scotch, M., & Heslop, D. J. (2025, April). Evaluating the Bias in LLMs for Surveying Opinion and Decision Making in Healthcare [arXiv:2504.08260 [cs]].

Kim, J., & Lee, B. (2024, April). AI-Augmented Surveys: Leveraging Large Language Models and Surveys for Opinion Prediction [arXiv:2305.09620 [cs]].

Lipton, Z. C., Kale, D. C., & Wetzel, R. (2016, November). Modeling Missing Data in Clinical Time Series with RNNs.

References II

Little, R., & Rubin, D. (2019, April). *Statistical Analysis with Missing Data, Third Edition* (1st ed.). Wiley.

Schuller, A., Janssen, D., Blumenröther, J., Probst, T. M., Schmidt, M., & Kumar, C. (2024).Generating personas using LLMs and assessing their viability. *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 1–7.

Stekhoven, D. J., & Bühlmann, P. (2012).MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, *28*(1), 112–118.

Von Der Heyde, L., Haensch, A.-C., & Wenz, A. (2025).Vox Populi, Vox AI? Using Large Language Models to Estimate German Vote Choice [Publisher: SAGE Publications]. *Social Science Computer Review*.

Thank you!

smitra98@umd.edu

ujstat@umd.edu