Harnessing Large Language Models for Market Research: A Data-augmentation Approach

Mengxin Wang[1], Dennis Zhang[2], Heng Zhang[3]

[1] Naveen Jindal School of Management, The University of Texas at Dallas, [2] Olin School of Business, Washington University in St. Louis, St. Louis, [3] W. P. Carey School of Business, Arizona State University

Problem Setting: Conjoint Analysis

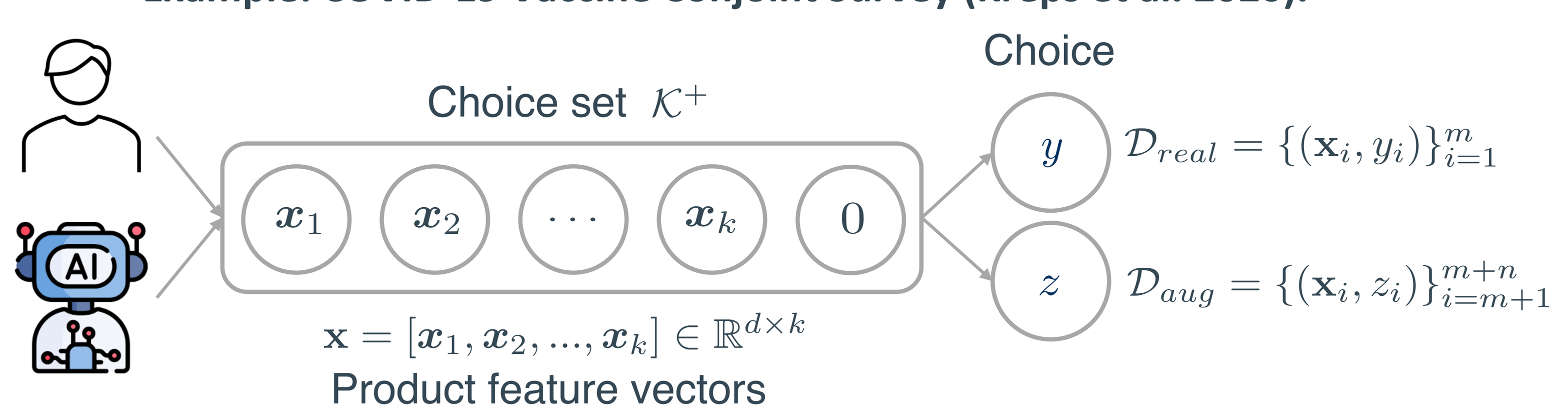
Conjoint analysis is one of the most important market research tool to understand consumer preferences. It relies on **choice-based surveys**, wherein responses are tasked with indicating their preferences among several products distinguished by various attributes.

Exhaustive survey requires substantial costs and resources. Reducing these costs has been a long-standing problem.

Generating Choice Survey Data Using LLMs

	Efficacy	*Protective Duration*	*Major Side effect*	*Minor Side Effect*	*Authorization*	*Origin*	*Endorsement*
A	*50%*	*1 year*	*1 in 10,000*	*1 in 30*	*Approved and licensed by US FDA*	*China*	*President Donald Trump*
B	*70%*	*5 years*	*1 in 1,000,000*	*1 in 30*	*Emergence use authorization from the US FDA*	*UK*	*Vice President Joe Biden*

Example: COVID-19 Vaccine Conjoint Survey (Kreps et al. 2020).

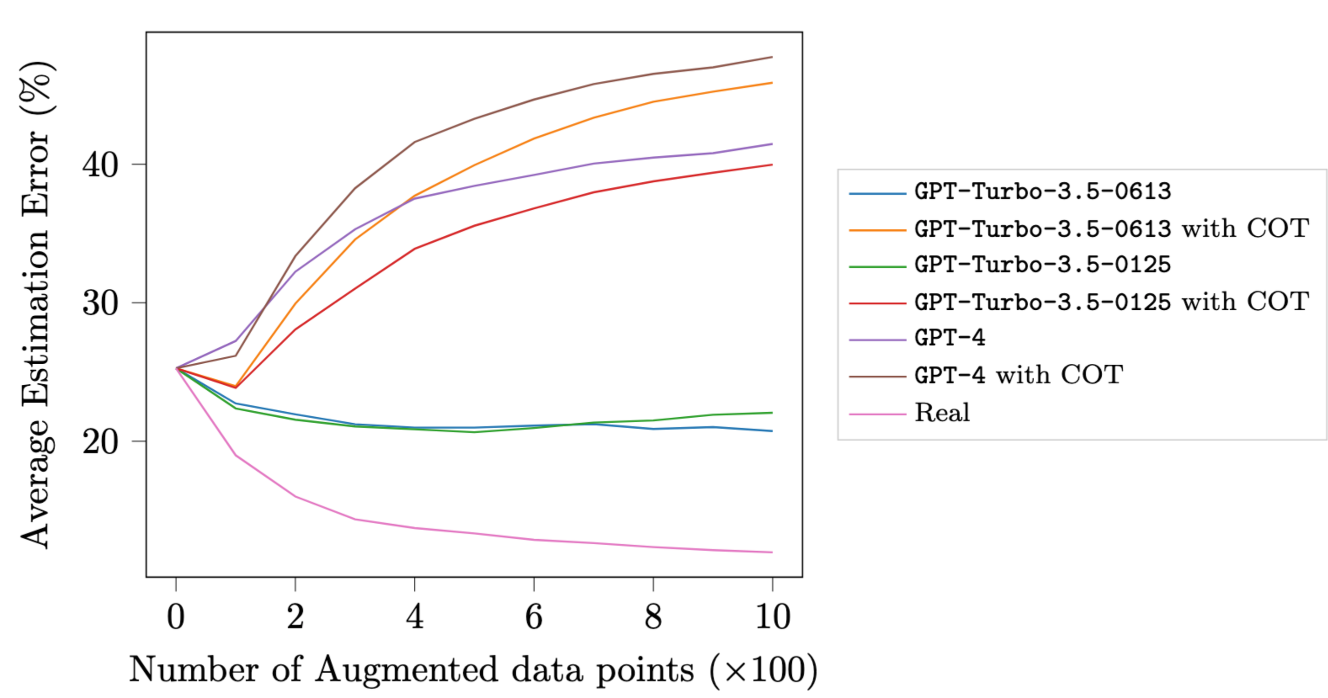


Misalignment of LLM and Human Choice data

Naïve augmentation: Use $D_{real} + D_{aug}$ with standard MLE to estimate the choice preference parameters:

$$\hat{\beta}^{\mathsf{Naive}} = \underset{\beta\in\mathbb{R}^d}{\arg\max} \left\{ \frac{1}{m+n} \sum_{i=1}^{m+n} \sum_{j\in\mathcal{K}^+} \log \sigma_j(\mathbf{x}_i;\beta) \right\}.$$

Naïve augmentation with LLM-generated choice data leads to inaccurate estimation:



- The gap persists with higher version of LLM models or advanced prompting
- Such misalignment is widely observed in surveys (Santurkar et al. 2023)
- Finding the right prompt & LLM can become a wild-goose chase

Contributions

- *Despite this imperfect alignment, we* develop a statistical data augmentation approach for **extracting value from LLM-generated choice data**
- Large-scale empirical studies show that **our method consistently reduces estimation error** with various LLMs, which saves 22%-82% of market research costs
- **Performance improves with better data**, showing the potential with larger models/more advanced prompting methods
- We provide **theoretical performance guarantees** for our method

Our Approach: AI-Augmented Estimation (AAE)

- Primary set: $\mathcal{D}^{\mathsf{P}} = \{(\mathbf{x}_i^{\mathsf{P}}, y_i^{\mathsf{P}}, z_i^{\mathsf{P}})\}_{i=1}^m$
- Auxiliary set $\mathcal{D}^{\mathsf{A}} = \{(\mathbf{x}_i^{\mathsf{A}}, z_i^{\mathsf{A}})\}_{i=1}^n$
- Assumption: $\mathbb{P}(y = j \mid \mathbf{x}, z) = g_j(\mathbf{x}, z; \boldsymbol{\theta}^*), \forall j \in \mathcal{K}^+$

AI-Augmented Estimation (AAE)

Step 1: Obtain an estimator of θ^* using the primary set

$$\{X_i^{\mathsf{P}}, z_i^{\mathsf{P}}\}_{i=1}^m \xrightarrow{\hat{\theta}} (y_i^{\mathsf{P}})_{i=1}^m$$

Step 2: Using the auxiliary data, obtain

$$\hat{\beta}^{\mathsf{AAE}} = \underset{\beta\in\mathbb{R}^{kd}}{\arg\max} \left\{ \widehat{Q}(\hat{\theta};\beta) = \frac{1}{n}\sum_{i=1}^n \sum_{j\in\mathcal{K}^+} g_j(\mathbf{x}_i^{\mathsf{A}}, z_i^{\mathsf{A}}; \hat{\theta}) \log \sigma_j(\mathbf{x}_i^{\mathsf{A}};\beta) \right\}.$$

Theoretical Performance Guarantee

Asymptotic Consistency and Normality of AAE

$$\boldsymbol{\beta}^* \in \underset{\boldsymbol{\beta}\in\mathbb{R}^d}{\arg\min} \left\{ \mathbb{E}_{\mathbf{x}}\Big[\mathsf{KL}\big(\mathbb{P}(y|\mathbf{x}) \,\big|\, \sigma_y(\mathbf{x},\boldsymbol{\beta})\big)\Big] = \mathbb{E}_{\mathbf{x}}\left[\sum_{j\in\mathcal{K}+} \mathbb{P}\big(y=j|\mathbf{x}\big) \log\left(\frac{\mathbb{P}\big(y=j|\mathbf{x}\big)}{\sigma_j\big(\mathbf{x};\boldsymbol{\beta}\big)}\right)\right]\right\}. \quad (1)$$

Theorem 1 (Consistency and Asymptotic Normality of AI-augmented Estimator)

(i) Under certain regularity assumption, the optimizer $\boldsymbol{\beta}^*$ defined in (1) is unique and the AAE satisfies $\hat{\boldsymbol{\beta}}^{\mathsf{AAE}} \xrightarrow{\mathrm{P}} \boldsymbol{\beta}^*$, when $m, n \to \infty$.

(ii) Under certain regularity assumptions, it holds that

$$\sqrt{n}(\hat{\boldsymbol{\beta}}^{\mathsf{AAE}} - \boldsymbol{\beta}^*) = \boldsymbol{\Omega}^{-1} \times \left(\frac{1}{n}\sum_{i=1}^n (\mathbf{p}(\mathbf{x}_i, z_i) - \boldsymbol{\sigma}(\mathbf{x}_i, \boldsymbol{\beta}^*)) \otimes \mathbf{x}_i + \sqrt{\frac{n}{m}}\boldsymbol{\Gamma} \times \sqrt{m}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\right) + o_{\mathrm{P}}(1)$$
$$\rightsquigarrow N\Big(\mathbf{0}, \boldsymbol{\Omega}^{-1}(\mathbf{J} + \rho \times \boldsymbol{\Gamma\Lambda\Gamma}^\top)\boldsymbol{\Omega}^{-1}\Big).$$

Value of AAE: Variance Reduction

Proposition 1 (Dominance of $\mathsf{Var}^{\mathsf{AAE}}$.) *Assume certain regularity assumptions hold and*

$$\boldsymbol{\Lambda} = \mathbb{E}_{\mathbf{x},y,z}\left[\nabla_\theta \log g_y(\mathbf{x}, z, \boldsymbol{\theta}^*) \nabla_\theta \log g_y(\mathbf{x}, z, \boldsymbol{\theta}^*)^\top\right]^{-1}.$$

(i) It holds that $\check{\mathbf{J}} \succeq \boldsymbol{\Gamma\Lambda\Gamma}^\top$. Therefore, for any $\delta > 0$ and any m, $\mathsf{Var}^{\mathsf{AAE}} \prec \mathsf{Var}^{\mathsf{P}} + \delta\mathbf{I}$ for all n sufficiently large.

(ii) If $\check{\mathbf{J}} \succ \boldsymbol{\Gamma\Lambda\Gamma}^\top$, for any m, $\mathsf{Var}^{\mathsf{AAE}} \prec \mathsf{Var}^{\mathsf{P}}$ for all n sufficiently large.

Empirical Results

We examine the performance of the AAE based on a real choice-based conjoint dataset for COVID-19 vaccines (Kreps et al. 2020). A total of 1,971 US adults responded to the survey, each expressing preferences for a series of hypothetical vaccines.

Estimation Error Reduction

		$m=50$			$m=100$			$m=150$			$m=200$		
Model	**Prompt**	A	Naive	AAE	A	Naive	AAE	A	Naive	AAE	A	Naive	AAE
GPT-3.5-Turbo-0613	Basic	-5.45	-10.80	-13.72	1.30	-4.53	-6.90	6.90	-2.06	-2.09	7.56	-1.79	-0.96
	CoT	20.18	16.81	-14.79	26.93	20.62	-8.09	32.53	23.06	-3.12	33.19	21.10	-1.82
GPT-3.5-Turbo-0125	Basic	-8.91	-9.92	-13.29	-2.16	-3.21	-6.50	3.43	0.86	-1.78	4.10	1.60	-0.76
	CoT	15.67	10.95	-14.84	22.43	14.71	-7.72	28.02	17.72	-2.74	28.69	16.16	-1.81
GPT-4	Basic	14.81	12.39	-15.70	21.56	16.20	-8.04	27.16	20.03	-3.12	27.82	19.20	-2.04
	CoT	21.77	18.12	-15.80	28.53	22.49	-8.30	34.12	26.33	-3.37	34.79	24.10	-2.27
GPT-4o	Basic	15.70	13.05	-15.55	22.46	17.34	-8.06	28.05	20.90	-3.07	28.72	19.65	-1.84
	CoT	20.61	16.50	-15.74	27.37	20.49	-8.06	32.96	23.65	-3.31	33.63	21.28	-2.25
	FS	12.46	9.71	-16.16	19.22	14.56	-8.26	24.81	18.41	-3.44	25.48	17.50	-2.10
GPT-4o Fine-tuned	Basic	4.83	3.18	-16.66	11.59	7.96	-9.58	17.18	12.46	-4.81	17.85	11.31	-3.36

Table 1: Change in MAPE per Feature (%)

Data and Cost Saving

Model	Prompt	$m=50$	$m=100$	$m=150$	$m=200$
GPT-3.5-Turbo-0613	Basic	74.5	48.5	31.4	2.2
	CoT	78.6	57.7	43.1	15.2
GPT-3.5-Turbo-0125	Basic	72.5	44.7	28.3	-0.8
	CoT	78.8	54.5	39.9	15.0
GPT-4	Basic	81.6	57.2	43.2	19.4
	CoT	81.9	59.6	45.0	22.2
GPT-4o	Basic	81.3	57.4	42.8	15.6
	CoT	81.8	57.4	44.7	22.0
	FS	82.7	59.4	45.6	20.3
GPT-4o Finetuned	Basic	83.8	66.2	54.0	32.7

Table 2: Percentage of Saving in Data Size (%)

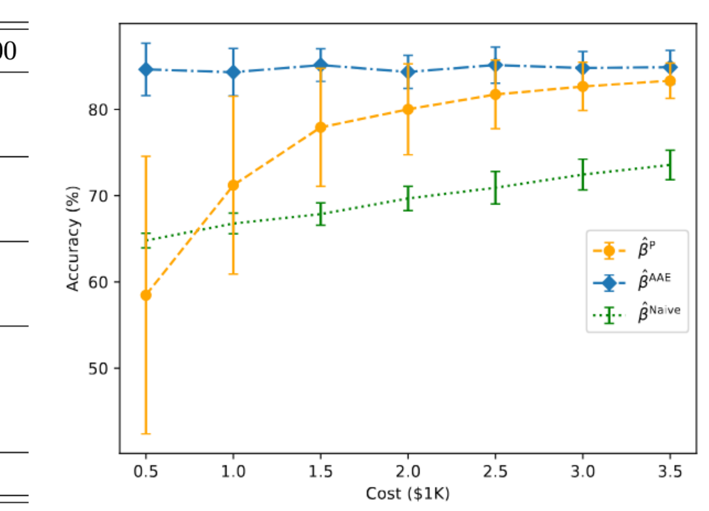
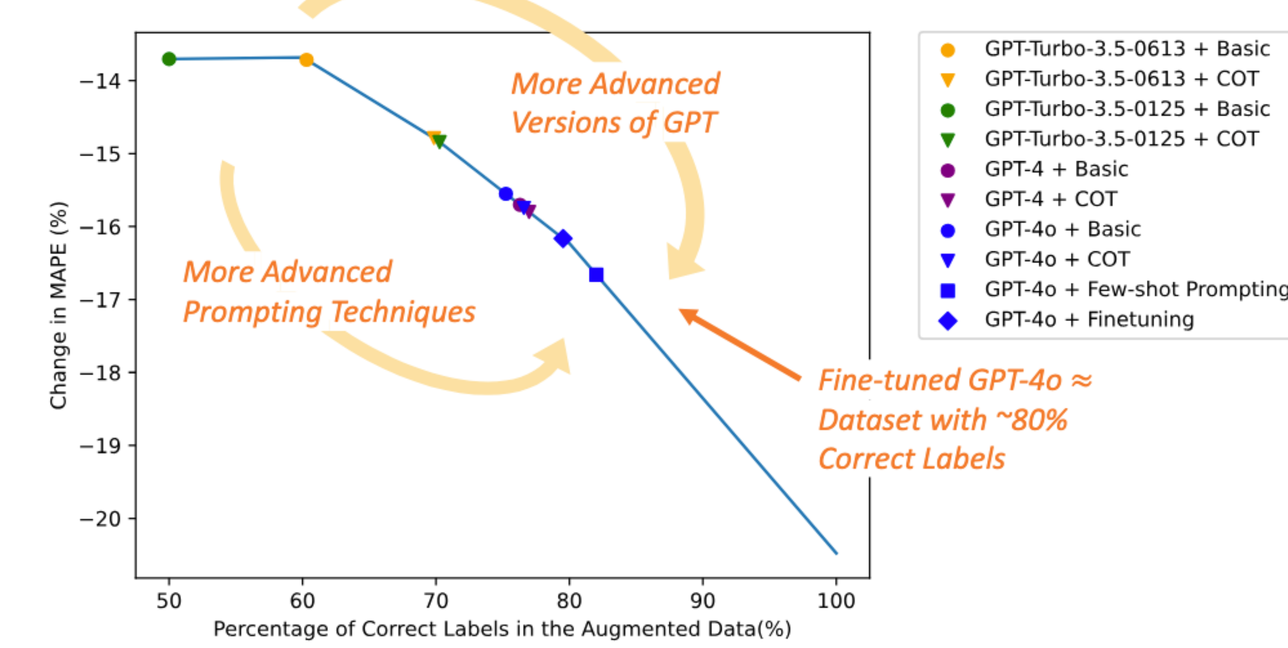


Figure 1: Estimation Accuracy vs. Market Research Costs

Value of LLM-generated Choice Data



References

Kreps, Sarah, et al. "Factors associated with US adults' likelihood of accepting COVID-19 vaccination." *JAMA network open* 3.10 (2020): e2025594-e2025594.

Santurkar, Shibani, et al. "Whose opinions do language models reflect?." *International Conference on Machine Learning*. PMLR, 2023.