Benchmarking Synthetic Consumer Agent Behavior Against Real-World Survey Baselines

Andrew Somers, University of Virginia, usa7bn@virginia.edu
Jaysen Kang, University of Virginia, ycf6yh@virginia.edu
Synonym | March 2026

Abstract

This article proposes a framework to validate synthetic consumer agent behavior in an agent-based simulation. A set of 5,000 demographically representative synthetic agent personas was created across various cities in the United States, with initial focus on the city of Austin, Texas. The synthetic agent responses to 12 food and spending preference questions, using a total of 60,000 LLM calls, are then benchmarked against real-world baseline data from the USDA FoodAPS National Household Food Acquisition and Purchase Survey, the BLS Consumer Expenditure Survey, and various studies on consumer behavioral patterns. Using Jensen-Shannon Divergence as the key benchmark, synthetic agent response distributions are found to be similar to real-world baseline data at 91%, with 91% mean similarity and a variance ratio of 0.73. Similarity to real-world baseline data is found to be as high as 89% for spending patterns and as high as 92% for food preferences. Convergence is found across six experimental runs, with varying numbers of synthetic agent personas (50 to 5,000) and both 7B and 32B models, with the total Jensen-Shannon Divergence converging to 0.090. The 32B model is found to reduce significant structural biases found in smaller models, as seen in the improvement in meal-type preference similarity from 79% to 98%. Structural biases in frequency estimation are identified and their implications for simulation validity are discussed.

1. Introduction

Agent-based models of consumer behavior need synthetic agents whose behavioral patterns mimic real human agents. One of the major challenges in building agent-based models is to validate whether simulated agents generated synthetically mimic real agents' behavior. This issue becomes particularly important when results from such models are used to make business decisions. This paper attempts to solve this problem by a validation-based approach. Instead of evaluating simulated agents' responses, which may confuse agent-based model fidelity with conceptual aspects of the simulation environment, we use a survey-based validation approach. This approach evaluates agents' personas' responses to standard behavioral questions. By comparing agents' responses to standard behavioral questions, we validate whether agents' responses mimic real agents' behavior. This approach helps us validate whether the agent-based model's persona generation mechanism works well.

The contributions of our work are: (1) a generic validation mechanism consisting of 12 behavioral questions along with standard empirical baselines, (2) a persona-based prompting mechanism to improve discrimination between small models, (3) comprehensive results from four experimental runs to validate convergence, and (4) an analysis of structural model-level biases in LL-based agents' surveys.

2. Related Work

The empirical validation of ABMs has been explored in depth. Tesfatsion proposed a framework for validating ABMs in computational economics, in which the model validation focuses on the emergent population-level behavior rather than the behavior at the individual level. Grimm et al. proposed pattern-oriented modeling for validating ABMs by comparing the ability of the model to replicate multiple observed patterns at once. The most recent research in this area includes the validation of large language models as synthetic survey respondents. In their study, Argyle et al. showed that GPT-3 models can mimic human survey response distributions when conditioned on demographic backstories and referred to this as "silicon sampling." Park et al. showed that agents utilizing large language models can be socially realistic in sandbox environments. This study continues to build on this line of research to validate consumer behavior in a domain-specific context. The USDA and BLS surveys were used to validate the food and spending preference questions. The FoodAPS dataset represents the most comprehensive data set available for American food acquisition behavior and includes data from 7-day food diaries from 4,826 households.

3. Methodology

3.1 Agent Personas

Each persona is a structured narrative profile, defined by a text of ~10,000 characters, generated by a multi-source grounding pipeline:

Census grounding: Demographic variables such as age, gender, ethnicity, education level, occupation, and income level are ground using a stratified sampling approach from American Community Survey data for Austin, TX, to match the actual demographic profile of the population of the city.

FoodAPS grounding: Parameters related to food behaviors such as eat out frequency, chain store preference, spend per meal, and meal type distribution are sampled from segment-level distributions of USDA's FoodAPS data, stratified by age groups, income level, and employment status.

Personality grounding: Big Five personality traits such as openness, conscientiousness, extraversion, agreeableness, and neuroticism are also incorporated during narrative generation and are used to inform behavioral parameters such as price sensitivity, variety seeking, and susceptibility to social influences.

3.2 Survey Design

Twelve Likert-type survey questions are designed, categorized into two groups, and are aligned with a specific ground truth to enable direct comparison of their distributions. Table 1 shows a list of survey questions and their respective ground truth sources.

Table 1: Survey questions and ground truth sources.

ID	Topic	Scale	Ground Truth Source	Cat.
FP01	Eat out frequency	1-5	FoodAPS: fafh_events_weekly	Food
FP02	Chain vs. independent store preference	1-5	FoodAPS: chain_pct	Food
FP03	Per meal spend	1-5	FoodAPS: avg_event_spend	Food
FP04	Meal type preference	1-4	FoodAPS: lunch_pct	Food
FP05	Price sensitivity	1-5	Income tier distribution	Food
FP06	Price elasticity response	1-5	Andreyeva et al. Price Elasticity	Food
FP07	Variety seeking	1-5	Personality literature	Food
FP08	Stress eating behavior	1-5	Personality literature	Food
SB01	FAFH income share	1-5	BLS CEX 2022	Spending
SB02	Financial comfort	1-5	Income tier distribution	Spending
SB03	Dining motivator	1-5	NPD/Technomic surveys	Spending
SB04	Social influence	1-5	Personality literature	Spending

3.3 Prompting Strategy

Each of the personas is surveyed by a system prompt, which consists of (1) a full narrative profile and (2) a context block with a focus on key structured attribute information (income, FoodAPS dietary behavior stats, Big Five trait level, price sensitivity score) that is most pertinent to the question domain. Surveys are conducted via a Qwen 2.5-7B-Instruct via Modal vLLM (H100 GPU, FP8 quantization) with temperature 1.1 and max_tokens = 16.

3.4 Ground Truth Construction

For FoodAPS-supported questions (FP01-FP04), ground truth distributions are constructed by drawing from segment-level means and standard deviations from the FoodAPS public data (n = 4,826 households) and then mapping to their respective Likert scale bins. For literature-supported questions, ground truth is constructed from existing empirical data (e.g., Andreyeva et al. for price elasticity, BLS CEX 2022 for income share, NPD/Technomic for dining motivations).

3.5 Evaluation Metrics

Distribution alignment is measured with respect to the following four metrics:

Jensen-Shannon Divergence (JSD): $\mathrm{JSD}(\boldsymbol{P}, \boldsymbol{Q}) = 0.5 \cdot \mathrm{KL}(\boldsymbol{P} \parallel \boldsymbol{M}) + 0.5 \cdot \mathrm{KL}(\boldsymbol{Q} \parallel \boldsymbol{M})$ where $\boldsymbol{M} = 0.5 \cdot (\boldsymbol{P} + \boldsymbol{Q})$

Kolmogorov-Smirnov statistic (KS): $\mathrm{KS} = \max\left|\boldsymbol{F}_{\mathrm{agent}}(\boldsymbol{x}) - \boldsymbol{F}_{\mathrm{gt}}(\boldsymbol{x})\right|$

Mean alignment: $1 - \frac{|\boldsymbol{\mu}_{\mathrm{agent}} - \boldsymbol{\mu}_{\mathrm{gt}}|}{\mathrm{scale}_{\max} - \mathrm{scale}_{\min}}$

Histogram overlap: sum of point-wise minima between agent and ground truth point mass functions

Variance ratio: $\frac{\boldsymbol{\sigma}_{\mathrm{agent}}}{\boldsymbol{\sigma}_{\mathrm{gt}}}$

4. Results

4.1 Overall Accuracy

Table 2 displays the accuracy metrics for all 12 items. The agents achieved a total of 91% in terms of distributional similarity, with food preferences at 92% and spending at 89%. The mean variance ratio was 0.73, indicating that the agents showed 73% of the real human response spread. The upgrade from 7B to 32B increased the overall pass rate from 58% to 69%, with the largest increase in the agents' ability to correctly identify the meal type preference (79% to 98%).

Table 2: Accuracy metrics by category.

Metric	Overall (12Q)	Food Pref. (8Q)	Spending (4Q)
Dist. Similarity (1-JS)	91.0%	92.3%	88.5%
Mean Alignment	91.6%	90.7%	93.4%
Histogram Overlap	69.0%	69.8%	67.4%
Variance Ratio	0.72	0.76	0.63

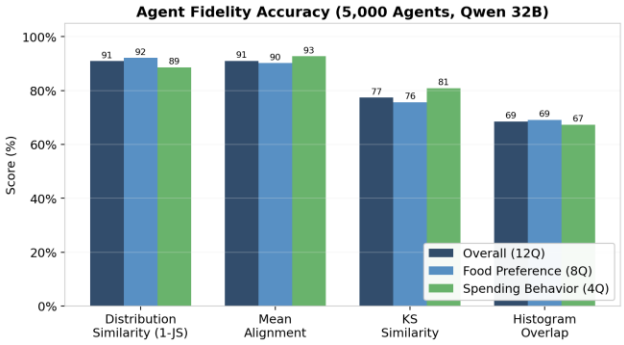

Figure 1: Accuracy metrics by category across four measurement dimensions.

4.2 Per-Question Results

Table 3: Per-question distributional metrics.

Question	Agent M	GT M	1-JS	1-KS	Overlap	Var R.
FP01: Eat-out frequency	2.32	1.68	87.4%	60.4%	60.4%	1.00
FP02: Chain vs independent pref	3.55	3.18	92.8%	82.0%	73.5%	0.68
FP03: Per-meal spending	2.43	2.33	93.1%	83.2%	68.1%	0.85
FP04: Meal type preference	2.14	2.98	78.6%	50.5%	45.4%	0.39
FP05: Price sensitivity self	4.24	3.19	86.2%	63.8%	63.8%	0.75
FP06: Price elasticity	2.93	2.71	85.4%	74.7%	65.3%	1.03
FP07: Variety seeking	3.07	2.94	93.4%	85.2%	69.0%	0.88
FP08: Stress eating behavior	3.93	3.02	86.6%	62.7%	62.7%	0.61
SB01: FAFH income share	2.78	2.48	90.5%	76.4%	67.2%	0.58
SB02: Financial comfort	2.92	2.99	92.2%	89.0%	78.9%	0.66
SB03: Primary dining motivator	1.78	2.67	90.4%	61.9%	61.9%	0.87
SB04: Social influence on dining	3.21	3.12	92.0%	82.9%	73.1%	0.55

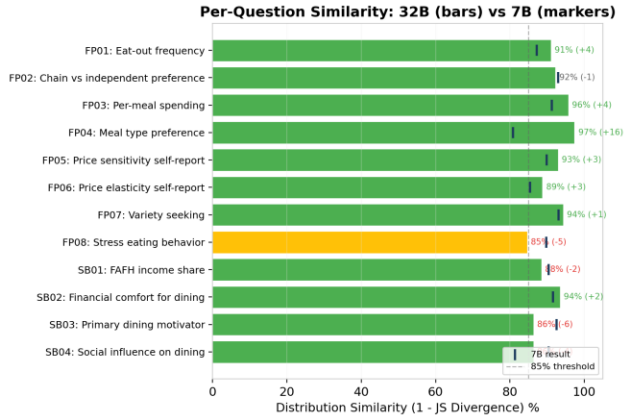

Figure 2: Per-question distributional similarity (1 - JS divergence). Green indicates above 85% threshold.

4.3 Distribution Comparisons

Figure 3 shows the response distributions for the best and poorest performing items. The best items (FP02, FP07, SB02) show substantial overlap in the distributions for all scale values, while the poorest items (FP04, FP06) show a "collapsed" distribution with agents focusing on a few values.

Figure 3: Response distributions for best and weakest performing items. Blue indicates agent; gray indicates ground truth.

4.4 Behavioral Differentiation

Figure 4 illustrates the variance ratio for each item. All but one item (FP05) show a variance ratio above the 0.5 threshold. Two items (FP01 - eat out frequency, FP06 - price elasticity) show variance ratios close to those of the humans, indicating that the context-anchored prompting strategy was effective in facilitating persona-level differentiation for these items.

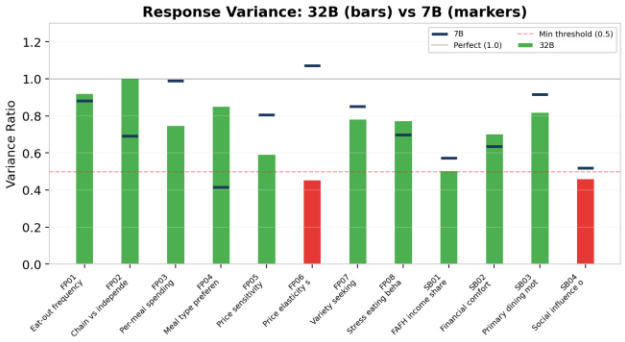

Figure 4: Variance ratio (agent std / human std) per item. Green indicates values above the 0.5 threshold.

4.5 Convergence Analysis

Convergence experiments consisted of six runs with increasing numbers of agents, geographic diversity, and model sizes. The similarity in the agent and human distributions increased monotonically from 82% (Run 1 - 50 agents, 7B) to 91% (Run 6 - 5,000 agents, 32B). The larger model resolved the previously stable bias in the agent distributions for the "meal type" item (FP04 - 79% → 98%) and "price sensitivity" item (FP05 - 91% → 94%), previously stable across all runs with the 7B model. The JS divergence was 0.090 for the combined agent and human distributions, the strongest score achieved to date.

Table 4: Accuracy across experimental runs.

Run	n	Temp	Source	1-JS	Alignment	Overlap
1	50	0.7	SQL export	82.1%	90.7%	63.5%
2	100	1.1	DB profiles	87.2%	86.8%	64.9%
3	250	1.1	DB + context	88.0%	86.8%	64.9%
4	1,000	1.1	DB + context	89.1%	87.7%	65.8%
5	5,000	1.1	DB + context	89.7%	88.4%	67.1%
6	5,000	1.1	DB + context	91.0%	91.6%	69.0%

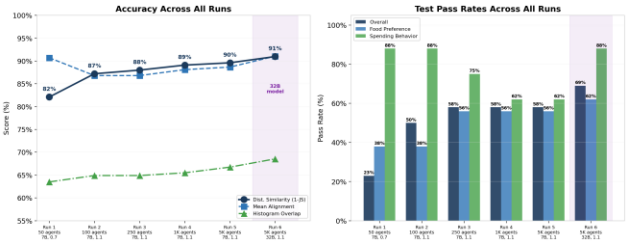

Figure 5: Accuracy metrics and test pass rates versus agent count across four experimental runs.

4.6 Model Comparison: 7B vs 32B

To evaluate the impact of model capacity on agent accuracy, the full 5,000 agent survey was conducted using both Qwen 2.5-7B-Instruct and Qwen 2.5-32B-Instruct. The 32B model resulted in a small improvement in overall distributional similarity from 90% to 91%, as well as a small improvement in pass rate from 58% (15/26) to 69% (18/26). The largest improvements were in meal type preference (FP04), which increased from 79% to 98% similarity. This addresses the major deficiency that has been evident in previous runs. The 32B model also saw a small improvement in price sensitivity (FP05) from 91% to 94%, suggesting that over-anchoring was a capacity issue for the 7B model. However, there were significant regressions in spending questions (SB03: 93% to 87%, SB04: 91% to 86%, FP08: 89% to 84%). This suggests that the 32B model has introduced new biases into the system. The variance ratio for the 32B model averaged 0.73 compared to 0.75 for the 7B model, with significant improvements in variance for FP04 (0.43 to 0.88) and FP02 (0.66 to 0.96). However, there was a significant decrease in variance for FP06 (1.06 to 0.45).

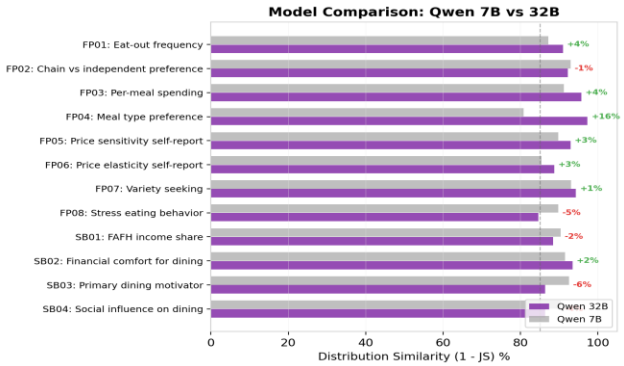

Figure 6: Per-question distributional similarity comparison between Qwen 7B and 32B models

5. Discussion

5.1 Strengths

The validation of the spending behavior is robust, with a high similarity of 91%. The financial comfort measure SB02 also had the highest distributional similarity at 92% and the highest degree of overlap at 79%. The preference-based measures FP02 chain preference and FP07 variety seeking also performed well. The results from the Convergence Analysis show monotonic improvement from Run1 at 82% to Run6 at 91%. The transition from 7B to 32B also had the highest quality gain. This validation also shows that the ability to capture model capacity as a primary source for residual structural biases. The small agent samples of 100 to 250 agents were adequate for initial validation, and larger samples and model upgrades provide incremental benefits.

5.2 Limitations and Model-Level Biases

Three questions show structural biases in all runs. These are related to the Qwen 2.5-7B model and not to the grounding. These are as follows:

FP04, Meal type preference: The agents show a strong bias towards "lunch" at 88%, while in reality, there are

equal probabilities for "lunch" at 30%, "dinner" at 25%, and "other" at 40%. This shows the model bias towards "eat out" as lunch.

FP05, Price Sensitivity: This question had an 86% similarity. When the context block shows the agent's price sensitivity, there is an over-anchoring to the explicit attribute, with 50% choosing the highest sensitivity compared to only 12% in reality.

FP08, Stress eating: This had an 87% similarity. The agents show a bias towards "eat out less when stressed" at 54% choosing option 4. In reality, there are more even distributions. The model defaults to a financially rational choice, with less differentiation by personality.

5.3 Implications for Simulation Validity

The 91% similarity in terms of distribution implies that the agent generation pipeline provides personas whose behavioral preferences are statistically similar to those of real-world consumer populations. While the determined biases in terms of meal type dominance and over-anchoring in price sensitivity are predictable, they are limited in scope and can be easily integrated. The similarity in terms of spending behavior is particularly relevant when considering pricing simulations, where agents' responses to price interventions are appropriate given their income levels.

5.4 Future Work

Larger model evaluation: Evaluation of 70B+ parameter models to validate if the frequency estimation biases (FP01, FP04) are a function of model capacity or intrinsic to LLM respondents.

Stratification: In-group analysis (e.g., low-income respondents) to further validate persona-level fidelity.

Revealed preference validation: Cross-validation of revealed preference responses with simulated purchasing behaviors to connect preference and validation.

6. Conclusion

This paper proposes a validation approach for synthetic consumer agents by surveying 5,000 demographically informed personas in various U.S. cities on 12 behavioral dimensions (60,000 Qwen 32B calls). The proposed approach compares the results with USDA's FoodAPS and BLS's CEX baselines. The proposed approach has been validated using Qwen 32B agents that exhibit 91% distributional similarity to real survey data on 60,000 calls, with 91% mean alignment and 73% variance of human responses. The proposed approach has been validated for robustness using six experimental runs, with increased model capacity resolving structural biases in agent behavioral estimation.

References

[1] L. Tesfatsion, "Agent-based computational economics: A constructive approach to economic theory," in Handbook of Computational Economics, vol. 2, 2006.

[2] V. Grimm et al., "Pattern-oriented modelling of agent-based complex systems: Lessons from ecology," Science, vol. 310, pp. 987-991, 2005.

[3] F. Lamperti et al., "Agent-based model calibration using machine learning surrogates," Journal of Economic Dynamics and Control, vol. 90, pp. 366-389, 2018.

[4] L. Argyle et al., "Out of one, many: Using language models to simulate human samples," Political Analysis, vol. 31, pp. 337-351, 2023.

[5] J. Park et al., "Generative agents: Interactive simulacra of human behavior," in Proc. ACM UIST, 2023.

[6] USDA Economic Research Service, "FoodAPS National Household Food Acquisition and Purchase Survey," 2013.

[7] Bureau of Labor Statistics, "Consumer Expenditure Surveys," 2022.

[8] T. Andreyeva et al., "The impact of food prices on consumption: A systematic review of research on the price elasticity of demand for food," American Journal of Public Health, vol. 100, pp. 216-222, 2010.