SIMULATING RESPONDENTS, NOT SINGLE QUESTIONS: COHERENT SURVEY GENERATION WITH LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models are increasingly used to simulate response distributions in social surveys. Prior work has achieved accurate population-level simulation for individual questions. Real-world questionnaires, however, typically require each respondent to answer a sequence of related questions. A simulated respondent should therefore exhibit coherent preferences across the entire questionnaire, rather than merely produce accurate distributions for isolated items. Existing single-item simulation methods can closely match item-level response distributions, but they do not accurately reproduce how the same respondent answers a complete survey. To address this limitation, we propose FullRespondent-LLM (FR-LLM), a framework for simulating complete virtual survey respondents. FR-LLM first fine-tunes two specialized LLMs: a marginal model that estimates the response distribution of each item and a respondent-level autoregressive model that captures dependencies among answers across the questionnaire. We combine these models through Marginal-Constrained Joint Projection (MCJP), which projects the autoregressive joint distribution onto the set of distributions satisfying the item-level marginals learned by the marginal model. By modeling item distributions and cross-item relationships separately, FR-LLM generates complete questionnaires that reproduce realistic cross-item relationships while retaining the item-level accuracy of strong single-item simulators. Experiments on two real-world social survey datasets show that FR-LLM more accurately reproduces response patterns across multiple questions, while maintaining competitive single-item accuracy and generalizing better to unseen respondent populations and survey questions. We also analyze a small commercial-survey dataset, use the responses generated by each method to make the same pricing and stocking decision, and compare the resulting profits. FR-LLM produces the highest profit, demonstrating its potential for practical survey-based decision making.

1 INTRODUCTION

Large language models (LLMs) are increasingly used to simulate how people answer surveys. Moving beyond prompting, Cao et al. (2025) fine-tune first-token probabilities of valid options against country-level answer distributions; Suh et al. (2025) train on subpopulation–question response distributions. Huang et al. (2026) further align how answer distributions change across respondent backgrounds, improving single-question simulation.

Given a respondent's background and one question, these methods can accurately estimate how a population answers that item. A real survey, however, asks the *same person* several related questions, and its applications often depend on a plausible complete set of answers. Someone reporting high life satisfaction, for example, will often also report happiness. A simulator that predicts the two questions separately may match the proportion of satisfied and happy people while failing to reproduce which answers come from the same person (Figure 1). Such questionnaires can look accurate one item at a time but be misleading when used to group respondents or inform decisions based on multiple answers. They can also have the wrong reliability: Cronbach's α depends on how answers to different questions vary together, which their separate response distributions cannot reveal (Williams et al., 2026; Lukauskas & Šarkauskaitė, 2026).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

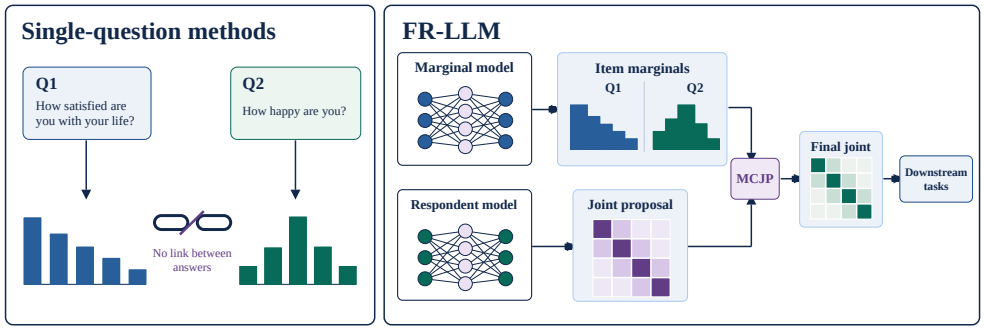


Figure 1: A conceptual satisfaction–happiness example (not an ESS or TALIS item pair). Matching each item's response rates does not ensure that the same simulated respondent gives coherent answers to both; FR-LLM models and calibrates complete response vectors.

A natural alternative is to answer the questions one after another, showing the model its earlier answers before it answers the next question. This lets it learn how a person's answers relate. But learning to produce a complete questionnaire does not separately ensure that the overall proportion choosing each option for each question matches the real survey. In our experiments, this fixed-order sequential approach better captures relationships between answers than single-question fine-tuning, yet usually matches each question's response distribution less accurately.

We propose FR-LLM to learn these two aspects separately and then bring them together: one model learns how often each answer is chosen, while another learns which answers tend to occur in the same person. The first LLM is fine-tuned on individual background–question–answer observations and estimates each question's response distribution. The second is fine-tuned on complete questionnaires from real respondents; it learns to answer using previous responses as context, with question order randomized during training, and generates complete candidate questionnaires. At inference, Marginal-Constrained Joint Projection (MCJP) adjusts how often each candidate is selected so that the final population matches the first model's item-level predictions. Formally, it chooses the closest distribution to the respondent model's proposal in reverse KL. Because it reweights and resamples *whole* questionnaires rather than changing individual answers, the final virtual respondents retain the relationships learned by the second model while matching the single-question distributions learned by the first.

We evaluate FR-LLM on ESS11 and TALIS 2018 with two LLM families under held-out questions, respondent populations, and both together. It improves multi-item accuracy while maintaining competitive item-level accuracy. On a separately collected commercial-survey dataset, we use synthetic responses to choose a selling price and stocking quantity, then assess profit on held-out responses. FR-LLM produces the highest profit. Our contributions are:

- We introduce FR-LLM, which combines separately learned item marginals and complete-response proposals through constrained projection for questionnaire-level tasks.
- We evaluate questionnaire reliability and validity with Cronbach's α and AVE alongside item and joint distributional errors.
- Across two large social surveys and a small commercial decision case, we demonstrate stronger multi-item accuracy, competitive item accuracy, and the practical value of coherent virtual respondents.

2 FullRespondent-LLM

2.1 Background and Notation

A *construct* is an attribute measured by several related items, such as self-efficacy across classroom situations. Survey-defined item groupings are used for evaluation; their names and definitions are withheld from the models.

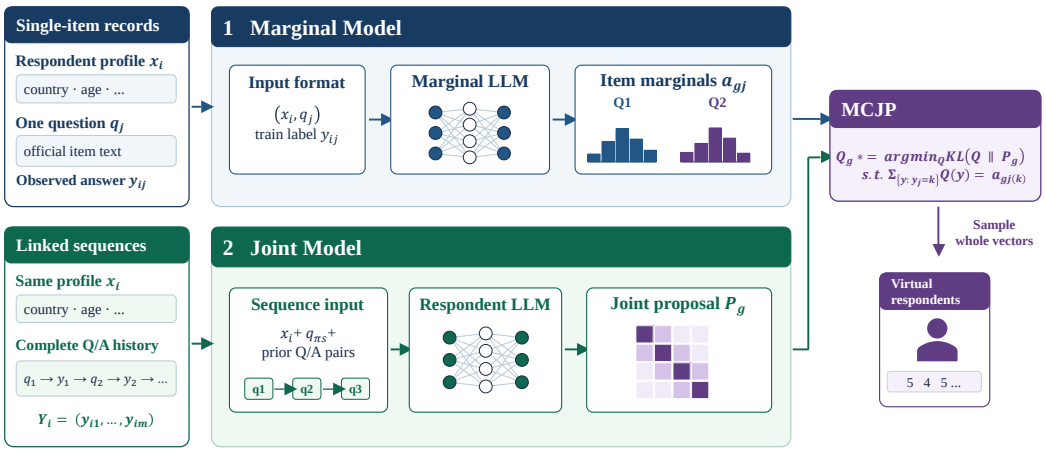


Figure 2: FR-LLM training and generation. Stage 1 scores one observed item at a time to estimate population marginals. Stage 2 trains on linked question–answer sequences from each respondent to propose complete vectors. MCJP reweights those vectors to match the learned marginals within each populated cell; it does not replace individual answers.

For r scored items, Cronbach's $\alpha = \frac{r}{r-1}\left(1 - \frac{\sum_j \mathrm{Var}(Y_j)}{\mathrm{Var}(\sum_j Y_j)}\right)$ measures internal consistency (Cronbach, 1951). Average variance extracted, $\mathrm{AVE} = r^{-1}\sum_j \lambda_j^2$ for standardized one-factor loadings λ_j, summarizes convergent validity (Fornell & Larcker, 1981). Both depend on linked answers. We compare their generated and human values across constructs as diagnostics, not training targets or stand-alone proof of validity.

Let q_j and $Y_j \in \{1, \ldots, K_j\}$ be item j and its answer, and $Y = (Y_1, \ldots, Y_m)$ a complete response. Respondent i has background x_i and linked answers Y_i; $g = G(x)$ is a prespecified population cell. We target the human joint $T_g(y) = p_{\mathrm{human}}(Y = y \mid G(X) = g, q_{1:m})$, not just item marginals. At test time we receive backgrounds but no target answers or construct names. Figure 2 shows how the stages learn margins, propose complete responses, and combine them; only the second LLM sees prior answers.

2.2 Stage 1: Learning Item Marginals

This stage aims to reproduce how often each option is chosen in a population cell. Accurate item rates matter even when the eventual output is a complete questionnaire, but a model trained on isolated questions cannot learn which answers tend to occur together. We therefore train a separate LoRA-adapted LLM (Hu et al., 2022) on one observed item at a time. Given respondent background x_i, item q_j, and its response scale (Appendix Figure 5), it scores only valid response tokens to produce probabilities $a_{ij}(k) = \mathrm{softmax}_k\, f_\phi(x_i, q_j)$. Ordinary cross-entropy minimizes $\mathcal{L}_{\mathrm{marg}}(\phi) = \mathbb{E}[-\log a_{ij}(Y_{ij})]$ over observed respondent–item pairs. In expectation, this is the entropy of the human item distribution plus its forward KL to the model distribution; hard answers thus provide samples for learning population rates.

At inference, the model predicts each item for the frozen target backgrounds. Within cell g, we average these probabilities using normalized protocol weights $\bar{w}_i$, obtaining the anchor $a_{gj}(k) = \sum_{i:G(x_i)=g} \bar{w}_i a_{ij}(k)$. This averaging retains the target cell's background composition without using any held-out answers. The anchors specify desired item rates for Stage 3, not respondent-level answer vectors.

2.3 Stage 2: Learning Complete Respondents

This stage provides what isolated-item training cannot: linked answers that reflect a single respondent's preferences across questions. The respondent model uses the same background and item wording, but also sees that person's preceding question–answer pairs (Appendix Figure 6). In a

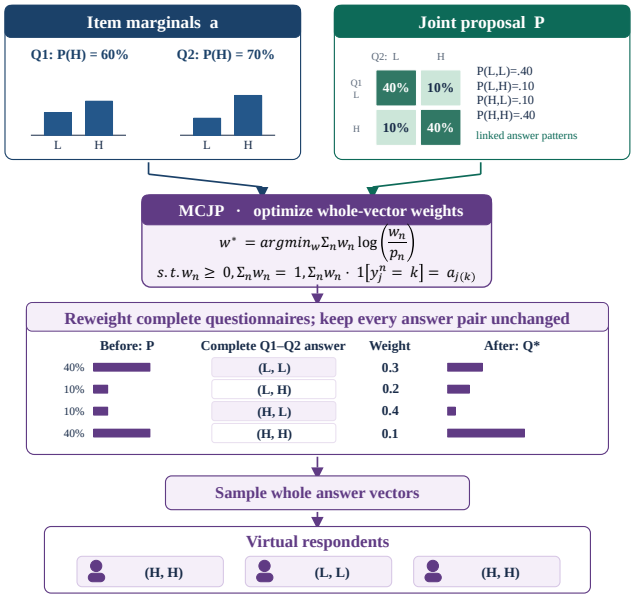


Figure 3: Schematic of MCJP on two binary questions, not an ESS or TALIS evaluation. Stage 1 supplies item targets, the respondent model supplies linked answer pairs, and projection reweights complete pairs before resampling. The displayed weight column is illustrative, not the numerical optimum for the displayed targets; actual weights are determined by Eq. 3 and IPF.

sampled item order π, it predicts each answer conditional on the previous answers, so the product of these conditional probabilities defines a joint distribution P_θ^π over complete questionnaires. Training teacher-forces the observed history, sums the losses over all items for one respondent, and performs one backward pass:

$$\mathcal{L}_{\text{joint}}(\theta) = \mathbb{E}_{(x_i,Y_i),\pi}\left[-\sum_{s=1}^{m} \log p_\theta(Y_{\pi_s} \mid x_i, q_{\pi_{\leq s}}, Y_{\pi_{<s}})\right]. \tag{1}$$

By the probability chain rule, this is a complete-response likelihood objective, averaged over sampled orders rather than tied to the questionnaire display order. At inference, the model uses its own generated histories and a fresh order for each candidate. Averaging across orders and the target backgrounds gives the cell-level proposal

$$P_g(y) = \sum_{i:G(x_i)=g} \bar{w}_i \, \mathbb{E}_{\pi\sim\Pi} P_\theta^\pi(y \mid x_i, q_{1:m}). \tag{2}$$

This proposal retains linked answer patterns. It need not, however, reproduce each item's response rate as accurately as Stage 1, which motivates the final combination.

2.4 Stage 3: Marginal-Constrained Joint Projection

The two trained models offer complementary information: Stage 1 supplies accurate item rates, while Stage 2 supplies plausible combinations of answers. This stage combines them without retraining either model or independently replacing answers. For a populated cell, let $P = P_g$ be the proposal in Eq. 2 and $a = (a_{gj})_j$ its Stage-1 target rates. Write $\varphi_{jk}(y) = \mathbb{1}\{y_j = k\}$ for each item and all but one redundant option, and let $\mathcal{C}(a)$ contain joint distributions whose expected indicators equal a. MCJP finds the distribution closest to the proposal subject to those rates:

$$Q_a = \arg\min_{Q\in\mathcal{C}(a)} D_{\text{KL}}(Q\|P). \tag{3}$$

Here reverse KL means changing the proposal only as needed to satisfy the calibrated rates. With feasible interior margins, the solution has the form $Q_a(y) \propto P(y)\exp\{\lambda_a^\top \varphi(y)\}$, with weights

chosen so its item marginals equal a. These item–option weights reweight whole answer vectors; they do not fit another model of cross-item interactions.

Figure 3 illustrates the reweighting of entire answer patterns. In practice, we sample $M = 16$ complete candidates per respondent with fresh item orders and pool them within populated cells. Iterative proportional fitting (IPF) adjusts their weights to match a, after which randomized systematic resampling produces three rollouts of whole vectors. If the candidate pool lacks the support needed for exact matching, we report an adjustment to the nearest feasible margin. No held-out human response enters MCJP.

Projection guarantee. The projection also separates two sources of error. Let T be the human joint, $t = \mathbb{E}_T \varphi$ its true margins, and Q_t the oracle projection onto $\mathcal{C}(t)$. Under full support and feasible interior margins,

$$\boxed{D_{\mathrm{KL}}(T\|Q_a) = \underbrace{D_{\mathrm{KL}}(T\|Q_t)}_{\text{dependence residual}} + \underbrace{D_{\mathrm{KL}}(Q_t\|Q_a)}_{\text{marginal-anchor error}}.} \tag{4}$$

Thus the final discrepancy consists of a residual dependence error and an error due to imperfect marginal anchors. Exact anchors cannot worsen forward KL relative to the proposal; with imperfect anchors, the outcome depends on their accuracy. The ideal reweighting also preserves conditional odds ratios. Appendix A gives the proof, a quantitative bound for anchor error, and the finite-support qualification relevant to IPF.

3 Experiments

3.1 Settings

Datasets. We evaluate on ESS11's 21-item Human Values Scale (ESS ERIC, 2024; Schwartz et al., 2012) and TALIS 2018's 12-item teacher self-efficacy battery (OECD, 2019). ESS11 trains on four constructs (9 items) and holds out six (12 items); TALIS trains on two constructs (8 items) and holds out one (4 items). **M1** tests unseen constructs in seen populations, **M2** tests seen constructs in unseen countries, and **M3** tests both unseen constructs and countries. Table 1 gives the respondent counts. Development countries are disjoint from test countries and used only for checkpoint selection; methods share the same target respondents within each mode. We additionally use 488 records from our Little-Treat Consumption survey for a separate held-out pricing-and-stocking task, not the ESS11/TALIS transfer benchmark (Appendix F.6).

Models. We use Qwen3.5-9B (Qwen Team, 2026) and Ministral-3-8B (Mistral AI, 2025). Learned methods share construct-free item wording and valid-option scoring. LoRA settings, prompts, and optimization are in Appendix B; checkpoints use development countries, and reported results average three generation rollouts.

Metrics. All four errors are lower-is-better. Item JSD compares human and generated answer distributions per question; Construct JSD compares complete answer tuples within each construct, testing cross-item patterns. α and one-factor AVE MAE average the absolute human–generated difference across constructs, measuring agreement in reliability and convergent-validity summaries. JSD is calculated before averaging; Appendix D specifies weights and aggregation.

Baselines. **Zero-shot** scores valid answers without survey fine-tuning. **Single FT** follows the single-question setting of Cao et al. (2025); Suh et al. (2025): it fine-tunes and samples each item independently, without answer history. **Sequential FT** extends their question-wise fine-tuning setting to complete questionnaires: it processes items in a fixed order, conditions on preceding answers, and makes a separate update for each item. This fixed-order, history-conditioned variant is our baseline implementation, not a method claimed by those papers.

3.2 Main results

Tables 2 and 3 compare both backbones in all three modes on ESS11 and TALIS. Across the resulting 12 settings, FR-LLM has the lowest Cronbach's α MAE, AVE MAE, and Construct JSD in every

Table 1: Distinct respondents by split. M1: unseen constructs; M2: unseen countries; M3: both. M2 and M3 share respondents but use different items.

Dataset	Train n	M1 test n	M2 test n	M3 test n
ESS11	28,268	28,268	7,312	7,312
TALIS 2018	12,000	8,000	8,000	8,000

Table 2: ESS11 results for both backbone models. Lower is better. Bold is best and underline is second best within each backbone and transfer mode; ties receive the same mark.

Mode	Method	Qwen3.5-9B				Ministral-3-8B			
		Cronbach's α MAE	AVE MAE	Item JSD	Construct JSD	Cronbach's α MAE	AVE MAE	Item JSD	Construct JSD
M1	Zero-shot	.507	.348	.075	.124	.442	.316	.106	.168
	Single FT	.363	.272	**.024**	<u>.037</u>	.367	.274	**.031**	<u>.048</u>
	Sequential FT	<u>.084</u>	<u>.075</u>	<u>.044</u>	.047	<u>.114</u>	<u>.103</u>	<u>.089</u>	.091
	FR-LLM	**.078**	**.071**	**.024**	**.021**	**.046**	**.040**	**.031**	**.032**
M2	Zero-shot	.625	.405	.084	.175	.524	.370	.107	.201
	Single FT	.493	.346	**.014**	<u>.039</u>	.496	.346	**.013**	<u>.041</u>
	Sequential FT	<u>.089</u>	<u>.082</u>	<u>.045</u>	.049	<u>.111</u>	<u>.103</u>	<u>.052</u>	.048
	FR-LLM	**.028**	**.030**	**.014**	**.009**	**.044**	**.050**	**.013**	**.011**
M3	Zero-shot	.470	.319	.078	.141	.399	.284	.110	.183
	Single FT	.378	.273	**.038**	<u>.061</u>	.375	.271	**.046**	<u>.072</u>
	Sequential FT	<u>.099</u>	<u>.084</u>	<u>.054</u>	.075	<u>.130</u>	<u>.113</u>	<u>.090</u>	.102
	FR-LLM	**.078**	**.066**	**.038**	**.045**	**.047**	**.037**	**.046**	**.056**

case. Its unrounded Item JSD is best or tied in 7 settings, and best or tied in 11 at the tables' three-decimal precision. Thus the gain in complete-response accuracy usually comes with competitive single-question accuracy. The clearest item-level exception is TALIS M1 with Qwen: Sequential FT attains Item JSD .015 versus .017 for FR-LLM, but FR-LLM improves Construct JSD from .058 to .043 and α/AVE MAE from .094/.142 to .008/.007.

The baselines expose the trade-off. Single FT matches item distributions but has large α and AVE errors, so independent answers poorly represent complete respondents. Sequential FT improves these errors by conditioning on earlier answers, but worsens Item JSD relative to Single FT in 11 of 12 settings. FR-LLM instead preserves competitive item accuracy while improving construct-level joint responses across both surveys.

3.3 Low-data results

Many real surveys collect relatively few complete questionnaires. We therefore repeat ESS11 M1 with Qwen using 1,000 training respondents and evaluate the low-data methods on the same 1,000 target respondents. As Table 4 shows, FR-LLM retains Single FT's item-level accuracy while reproducing the within-construct joint distribution and questionnaire-level reliability and validity summaries more accurately. Thus, even a small linked sample can provide useful information about which answers occur together, making complete-respondent simulation plausible for narrower survey applications. The full-data rows reproduce Table 2 only as context: they use a larger test population and three rollouts rather than two, so the cross-row differences are not a controlled learning curve.

3.4 A commercial pricing and stocking simulation

Beyond matching survey distributions, we test whether synthetic questionnaires support a downstream pricing-and-stocking decision. The Little-Treat survey has 294 training, 97 development, and 97 held-out test respondents. Each model generates six attitude answers, purchase frequency, and the price range of the respondent's most recent purchase. We use frequency to estimate monthly demand

Table 3: TALIS 2018 results for both backbone models. Lower is better. Bold is best and underline is second best within each backbone and transfer mode; ties receive the same mark.

Mode	Method	Qwen3.5-9B				Ministral-3-8B			
		Cronbach's α MAE	AVE MAE	Item JSD	Construct JSD	Cronbach's α MAE	AVE MAE	Item JSD	Construct JSD
M1	Zero-shot	.809	.395	.061	.296	.808	.392	.141	.474
	Single FT	.642	.362	.017	.144	.831	.396	**.020**	.168
	Sequential FT	.094	.142	**.015**	.058	.114	.181	.029	.075
	FR-LLM	**.008**	**.007**	.017	**.043**	**.049**	**.058**	**.020**	**.052**
M2	Zero-shot	.786	.387	.069	.309	.811	.388	.150	.486
	Single FT	.772	.383	**.012**	.150	.817	.387	**.016**	.157
	Sequential FT	.076	.110	.015	.034	.114	.180	.026	.069
	FR-LLM	**.003**	**.004**	**.012**	**.013**	**.037**	**.042**	**.016**	**.015**
M3	Zero-shot	.796	.392	.078	.311	.821	.393	.149	.486
	Single FT	.776	.388	**.018**	.159	.833	.393	**.023**	.174
	Sequential FT	.095	.141	.023	.040	.118	.186	.036	.062
	FR-LLM	**.001**	**.003**	**.018**	**.029**	**.045**	**.055**	**.023**	**.046**

Table 4: ESS11 M1 low-data results (Qwen). Full-data rows reproduce Table 2; bold marks the best low-data value. Lower is better.

Method	Train n	Test n	α MAE	AVE MAE	Item JSD	Construct JSD
Zero-shot	0	1,000	.494	.329	.073	.131
Single FT	1,000	1,000	.406	.290	**.054**	.042
FR-LLM	1,000	1,000	**.066**	**.061**	**.054**	**.022**
Single FT	28,268	28,268	.363	.272	.024	.037
FR-LLM	28,268	28,268	.078	.071	.024	.021

and the historical price range as a *proxy* spending ceiling, not an elicited willingness to pay. At each candidate price p, the model estimates demand $\widehat{D}_m(p)$ from these two linked answers, chooses the price maximizing $(p-c)\widehat{D}_m(p)$ at unit cost c, and stocks $\widehat{D}_m(p)$ units per prospective customer. Applying that decision to held-out human demand gives sales, unsold inventory, and profit after stocking costs. Figure 4 reports two post-hoc cost scenarios with zero salvage, extrapolated to 1,000 customers. Monetary values are USD equivalents at a fixed 7 CNY per USD; the original-currency optimization and quantities are unchanged.

FR-LLM has the highest profit point estimate at both costs and avoids overstocking, whereas Zero-shot overestimates demand and leaves about 880 units unsold per 1,000 prospective customers. The advantage over Sequential FT is small: paired bootstrap intervals over the 97 held-out respondents include zero. This is a conditional decision simulation using historical spending, not evidence of causal price elasticity, realized market profit, or a statistically resolved advantage over that baseline.

3.5 Analysis

Background shifts. We change education, age, residence, or two background fields in otherwise fixed ESS11 profiles and compare the resulting item-mean changes with the corresponding human group differences. Shift MAE averages the absolute errors across held-out questions, capturing both the direction and size of each change. Table 5 shows that FR-LLM has the lowest macro-average error across the five contrasts shared by all three methods (.164, versus .169 for Single FT and .262 for Zero-shot). It leads on four contrasts, but Single FT is better on the combined residence–gender change. These prompt edits test conditional associations, not causal demographic effects; the excluded gender-only contrast is discussed in Appendix F.3.

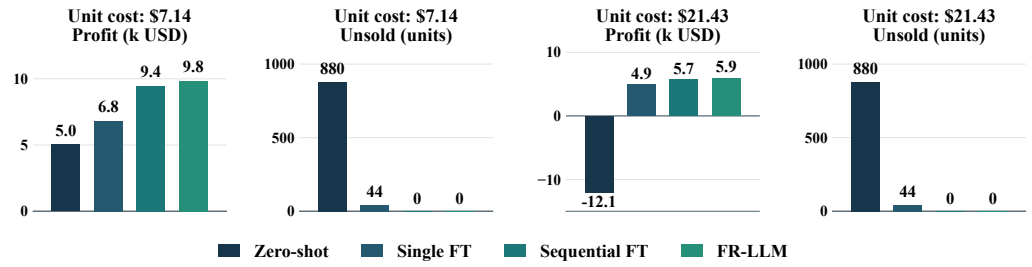


Figure 4: Little-Treat pricing and stocking: profit (thousand USD) and unsold units per 1,000 prospective customers at two unit costs; zero salvage.

Table 5: ESS11 background-shift MAE across five contrasts (response-scale points; lower is better).

Background change	Zero-shot	Single FT	FR-LLM
Low to high education	.261	.153	**.149**
Age 18–29 to 60+	.399	.231	**.216**
Urban to rural	.123	.112	**.111**
Low/young to high/old	.298	.198	**.177**
Urban/male to rural/female	.229	**.150**	.168
Macro average	.262	.169	**.164**

Human sampling reference. We estimate ordinary sampling variation from 1,000 pairs of independent, stratified respondent-level ESS11 M1 resamples. Table 6 reports each method's excess error above this human–human reference. Relative to Zero-shot, FR-LLM closes 86% of the gap in reliability error, 81% in validity error, and 83% in construct-level JSD. It remains measurably above human resampling error, so this empirical reference should not be interpreted as an attainable universal lower bound.

Component and sampling ablation. Table 7 follows the path from Zero-shot and Single FT to a complete-response proposal and its MCJP-corrected outputs. The proposal supplies cross-item dependence; projection brings its item margins closer to those learned by Single FT. At 16 candidates, this lowers construct JSD from .028 to .022. Increasing the projected candidate pool from one to 16 improves α and AVE errors, although not every metric changes monotonically. Because MCJP needs complete-response proposals, these rows compare components and sampling budgets rather than successive fine-tuning checkpoints. Appendix F.1 gives the full candidate-budget and anchor-path audits.

4 Related Work

Simulating survey responses. Prompted virtual samples established the appeal of LLM-based survey simulation (Argyle et al., 2023), while first-token specialization and subpopulation fine-tuning improved single-question response distributions (Cao et al., 2025; Suh et al., 2025). Background-shift alignment further targets how those distributions change across groups (Huang et al., 2026). Work on virtual respondents and restricted closed-ended generation broadens the task and clarifies the importance of response format (Zhao et al., 2026; Ahnert et al., 2026). Our distinction is to train on linked questionnaires while separately preserving calibrated item rates, then evaluate both properties on held-out constructs and countries.

Joint structure and calibration. Williams et al. (2026) show that good item margins can conceal poor cross-item correlations; Krsteski et al. (2026) study the complementary use of limited human data to rectify population estimates. We instead project a learned complete-response distribution toward margins predicted without held-out answers. The underlying I-projection and iterative fitting have a long statistical history (Csiszár, 1975; Deming & Stephan, 1940); our contribution is their use to couple two specialized survey LLMs while retaining whole-response candidates.

Table 6: ESS11 M1 gap to human resampling (Qwen). Excess error is model error minus the mean human–human error over 1,000 stratified resample pairs; lower is better. Gap closed is $(E_{\mathrm{ZS}} - E_{\mathrm{FR}})/(E_{\mathrm{ZS}} - E_{\mathrm{human}})$; higher is better.

	Human	Excess error above human resampling				FR gap
Metric	error	Zero-shot	Single FT	Sequential FT	FR-LLM	closed
α MAE	.007	.500	.356	.077	**.071**	85.9%
AVE MAE	.007	.342	.265	.068	**.065**	81.1%
Item JSD	.007	.067	**.017**	.037	**.017**	74.4%
Construct JSD	$<$.001	.124	.037	.047	**.021**	82.9%

Table 7: ESS11 M1 component and candidate-sampling ablation (Qwen; lower is better). The joint proposal uses 16 samples without projection; MCJP uses learned Stage-1 margins. Rows with different candidate counts are frozen-model re-evaluations.

Configuration	α MAE	AVE MAE	Item JSD	Construct JSD
Zero-shot	.507	.348	.075	.124
Single FT	.363	.272	**.024**	.037
Joint proposal (16 samples)	**.076**	**.070**	.027	.028
MCJP (1 sample)	.087	.080	**.024**	.022
MCJP (16 samples)	.078	.071	**.024**	**.022**

5 Discussion

The two learned models contribute different information: one-dimensional response rates and linked answer patterns. MCJP combines them by changing first-order potentials while preserving the proposal's conditional odds ratios in the ideal projection. It never trains on α, AVE, JSD, or downstream profit. Marginal validation, support adjustment, and effective sample size help diagnose when the projected answer pool cannot realize the intended rates; mixed gains with Ministral underline the importance of anchor quality.

Our evidence is limited to closed-ended batteries, two 8–9B backbones, one fine-tuning seed, and three generation rollouts. Finite candidate support weakens the ideal full-support KL guarantee (Appendix A.8); country holdout does not establish transfer to every demographic intersection. The method also requires two fine-tunes and multiple proposal samples. Finally, α and AVE are descriptive diagnostics, not proof that synthetic respondents can replace human validation or support causal and consequential population claims.

6 Conclusion

Survey simulation should be evaluated at the level at which surveys are observed: linked answers from respondents. FR-LLM uses two ordinary LoRA adapters and one constrained distributional operation: the marginal model estimates how often answers occur, the respondent model estimates which answers occur together, and MCJP finds the closest proposal distribution consistent with the calibrated rates. Across two surveys, two backbones, and three transfer regimes, this separation improves multi-item structure without giving up single-item accuracy. The remaining gap to human resampling shows that virtual respondents are not solved; the contribution is a formulation in which both improvement and failure are measurable.

AI USE STATEMENT

Generative AI assisted with research planning and methodology, experiment code, mathematical claims and proof drafts, literature search, figures, result interpretation, and manuscript drafting and editing. The authors checked source papers, code, equations, and numerical claims against saved artifacts and take responsibility for all content; no AI system is an author.

ETHICS STATEMENT

We use public-use ESS/TALIS files and de-identified responses from our Little-Treat survey under their respective data-use terms. Synthetic responses may distort minority views or be mistaken for real population evidence; they must not replace human surveys in consequential decisions. We will not release non-redistributable respondent records.

REPRODUCIBILITY STATEMENT

Appendix C documents splits, prompts, training, sampling, and metrics; Appendix A gives proofs. We will release code and aggregate outputs, while raw survey files remain with their providers.

REFERENCES

Georg Ahnert, Anna-Carolina Haensch, Barbara Plank, and Markus Strohmaier. Survey response generation: Generating closed-ended survey responses in-silico with large language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*, pp. 41554–41577, 2026. doi: 10.18653/v1/2026.acl-long.1927.

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351, 2023. doi: 10.1017/pan.2023.2.

Yong Cao, Haijiang Liu, Arnav Arora, Isabelle Augenstein, Paul Röttger, and Daniel Hershcovich. Specializing large language models to simulate survey response distributions for global populations. In *Proceedings of NAACL-HLT*, pp. 3141–3154, 2025. doi: 10.18653/v1/2025.naacl-long.162.

Lee J. Cronbach. Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3):297–334, 1951.

Imre Csiszár. I-divergence geometry of probability distributions and minimization problems. *The Annals of Probability*, 3(1):146–158, 1975.

W. Edwards Deming and Frederick F. Stephan. On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics*, 11(4): 427–444, 1940.

ESS ERIC. ESS round 11: European social survey round 11 data. Data file edition 2.0, Sikt – Norwegian Agency for Shared Services in Education and Research, 2024.

Claes Fornell and David F. Larcker. Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research*, 18(1):39–50, 1981.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Ji Huang, Mengfei Li, and Shuai Shao. Distribution shift alignment helps LLMs simulate survey response distributions. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 9395–9409, 2026. doi: 10.18653/v1/2026.findings-acl.457. URL https://aclanthology.org/2026.findings-acl.457/.

Stefan Krsteski, Giuseppe Russo, Serina Chang, Robert West, and Kristina Gligorić. Valid survey simulations with limited human data: The roles of prompting, fine-tuning, and rectification. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*, pp. 10887–10906, 2026. doi: 10.18653/v1/2026.acl-long.498.

Mantas Lukauskas and Viktorija Šarkauskaitė. Plausible but not valid: A psychometric audit of LLMs as synthetic survey respondents. *arXiv preprint arXiv:2608.14606*, 2026.

Mistral AI. Ministral 3 8B model card. Mistral AI model documentation, 2025. URL https://docs.mistral.ai/models/ministral-3-8b-25-12.

OECD. TALIS 2018 technical report. Technical report, Organisation for Economic Co-operation and Development, 2019. URL https://www.oecd.org/content/dam/oecd/en/about/programmes/edu/talis/talis2018supportmaterials/TALIS_2018_Technical_Report.pdf.

Qwen Team. Qwen3.5-9B model card. Hugging Face model documentation, 2026. URL https://huggingface.co/Qwen/Qwen3.5-9B.

Shalom H. Schwartz, Jan Cieciuch, Michele Vecchione, Eldad Davidov, Ronald Fischer, Constanze Beierlein, Alice Ramos, Markku Verkasalo, Jan-Erik Lonnqvist, Kursad Demirutku, Ozlem Dirilen-Gumus, and Mark Konty. Refining the theory of basic individual values. *Journal of Personality and Social Psychology*, 103(4):663–688, 2012.

Joseph Suh, Erfan Jahanparast, Suhong Moon, Minwoo Kang, and Serina Chang. Language model fine-tuning on scaled survey data for predicting distributions of public opinions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pp. 21147–21170, 2025. doi: 10.18653/v1/2025.acl-long.1028.

Tristan Williams, Franziska Weeber, Sebastian Padó, and Alan Akbik. Beyond marginal distributions: A framework to evaluate the representativeness of demographic-aligned LLMs. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 4806–4826, 2026. doi: 10.18653/v1/2026.findings-acl.236.

Jianpeng Zhao, Chenyu Yuan, Weiming Luo, Haoling Xie, Guangwei Zhang, Steven Jige Quan, Zixuan Yuan, Pengyang Wang, and Denghui Zhang. Large language models as virtual survey respondents: Evaluating sociodemographic response generation. *arXiv preprint arXiv:2509.06337*, 2026. URL https://arxiv.org/abs/2509.06337v2.

A Complete Theory

This appendix distinguishes the ideal population version of MCJP from the finite-candidate implementation. The first yields exact KL identities; the second adds support approximation and Monte Carlo error.

A.1 Setup and assumptions

Fix a populated projection cell g and omit it from notation. A complete questionnaire is $Y = (Y_1, \dots, Y_m) \in \Omega = \prod_j \{1, \dots, K_j\}$. Let T be the cell-level human joint distribution, P the cell-level mixture proposal in Eq. 2, a the Stage-1 cell-marginal vector, and t the true marginal vector. Remove the final redundant category for each item and define

$$\varphi(Y) = (\mathbb{1}\{Y_j = k\})_{j=1:m,k=1:K_j-1} \in \mathbb{R}^d, \qquad t = \mathbb{E}_T \varphi(Y). \tag{5}$$

For a feasible mean parameter u, let $\mathcal{C}(u) = \{Q : \mathbb{E}_Q \varphi = u\}$ and

$$Q_u = \arg\min_{Q \in \mathcal{C}(u)} D_{\mathrm{KL}}(Q \| P). \tag{6}$$

We use the following standard conditions.

(A1) $T \ll P$: every human response pattern has positive proposal probability.

(A2) a and t lie in the relative interior of the marginal polytope induced by the support of P.

(A3) For the quantitative error bound, $\nabla^2 A(\lambda) = \mathrm{Cov}_{Q_\lambda}[\varphi(Y)] \succeq \mu I$ along the segment joining the dual parameters of a and t, for some $\mu > 0$.

Dropping one category per item is necessary for A3 because a full one-hot vector is linearly dependent.

A.2 From respondent training to a cell-level proposal

The respondent objective is optimized for individual backgrounds and randomized item orders, whereas MCJP acts on their population mixture. This aggregation does not create an unaccounted risk term. Let w_i be normalized weights inside cell g, let T_i be the human conditional joint for background x_i, and let P_i^π be the order-specific respondent proposal. Define

$$T_g = \sum_i w_i T_i, \qquad P_g = \sum_i w_i \mathbb{E}_{\pi\sim\Pi} P_i^\pi. \tag{7}$$

Joint convexity of KL and Jensen's inequality give

$$D_{\mathrm{KL}}(T_g \| P_g) \le \sum_i w_i D_{\mathrm{KL}}(T_i \| \mathbb{E}_\pi P_i^\pi) \le \sum_i w_i \mathbb{E}_\pi D_{\mathrm{KL}}(T_i \| P_i^\pi). \tag{8}$$

The final quantity is the population analogue of the order-randomized autoregressive loss, up to the entropy of T_i. Thus Stage 2 controls an upper bound on the respondent-model term used below; it does not require any one display order to be canonical.

A.3 Unique exponential-tilt solution

Theorem A.1. Under A1–A2, Eq. 6 has a unique solution of the form

$$Q_u(y) = P(y)\exp\{\lambda_u^\top \varphi(y) - A(\lambda_u)\}, \quad A(\lambda) = \log \mathbb{E}_P \exp\{\lambda^\top \varphi(Y)\}, \tag{9}$$

where $\nabla A(\lambda_u) = u$.

Proof. Introduce Lagrange multipliers for normalization and the mean constraint in Eq. 6. Differentiation with respect to each positive $Q(y)$ gives

$$\log \frac{Q(y)}{P(y)} = \lambda_u^\top \varphi(y) - c. \tag{10}$$

Normalization yields $c = A(\lambda_u)$, and differentiating the log partition function gives $\nabla A(\lambda_u) = \mathbb{E}_{Q_u}\varphi = u$. KL is strictly convex in Q on the support of P, while $\mathcal{C}(u)$ is convex; the feasible stationary point is therefore unique. □

A.4 Exact decomposition and non-degradation

Theorem A.2. Under A1–A2,

$$D_{\mathrm{KL}}(T\|Q_a) = D_{\mathrm{KL}}(T\|Q_t) + D_{\mathrm{KL}}(Q_t\|Q_a), \tag{11}$$

$$D_{\mathrm{KL}}(T\|P) = D_{\mathrm{KL}}(T\|Q_t) + D_{\mathrm{KL}}(Q_t\|P). \tag{12}$$

Proof. Equation 9 implies

$$\log \frac{Q_t(y)}{Q_a(y)} = (\lambda_t - \lambda_a)^\top \varphi(y) - A(\lambda_t) + A(\lambda_a), \tag{13}$$

an affine function of $\varphi(y)$. Because T and Q_t share the same true mean parameter t,

$$\mathbb{E}_T \log \frac{Q_t(Y)}{Q_a(Y)} = \mathbb{E}_{Q_t} \log \frac{Q_t(Y)}{Q_a(Y)} = D_{\mathrm{KL}}(Q_t\|Q_a). \tag{14}$$

Adding $\mathbb{E}_T \log(T/Q_t)$ proves Eq. 11. Likewise, $\log(Q_t/P)$ is affine in φ and $T, Q_t \in \mathcal{C}(t)$, so

$$\mathbb{E}_T \log \frac{Q_t(Y)}{P(Y)} = \mathbb{E}_{Q_t} \log \frac{Q_t(Y)}{P(Y)} = D_{\mathrm{KL}}(Q_t\|P), \tag{15}$$

which proves Eq. 12. □

Combining the identities gives

$$D_{\mathrm{KL}}(T\|Q_a) = D_{\mathrm{KL}}(T\|P) - D_{\mathrm{KL}}(Q_t\|P) + D_{\mathrm{KL}}(Q_t\|Q_a). \tag{16}$$

If $a = t$, then $Q_a = Q_t$ and

$$D_{\mathrm{KL}}(T\|Q_a) = D_{\mathrm{KL}}(T\|P) - D_{\mathrm{KL}}(Q_a\|P) \le D_{\mathrm{KL}}(T\|P). \tag{17}$$

For imperfect margins, projection improves on the respondent model if and only if

$$D_{\mathrm{KL}}(Q_t\|Q_a) < D_{\mathrm{KL}}(Q_t\|P). \tag{18}$$

This condition is important when interpreting the Ministral experiments: a weak Stage-1 anchor can cost more than the correction gains.

A.5 COMPONENT-TO-FUSION ERROR BOUND

Let A^* be the convex conjugate of A. Exponential-family duality gives

$$D_{\mathrm{KL}}(Q_t \| Q_a) = D_{A^*}(t \| a). \tag{19}$$

Under A3, A^* is $1/\mu$-smooth, hence

$$D_{\mathrm{KL}}(Q_t \| Q_a) \le \frac{1}{2\mu} \|t - a\|_2^2. \tag{20}$$

Together with Theorem A.2,

$$D_{\mathrm{KL}}(T \| Q_a) \le R_{\mathrm{dep}} + \frac{1}{2\mu} \|t - a\|_2^2, \qquad R_{\mathrm{dep}} = D_{\mathrm{KL}}(T \| Q_t). \tag{21}$$

For each item, Pinsker's inequality gives $\|t_j - a_j\|_2^2 \le \|t_j - a_j\|_1^2 \le 2 D_{\mathrm{KL}}(t_j \| a_j)$. Therefore

$$\boxed{D_{\mathrm{KL}}(T \| Q_a) \le R_{\mathrm{dep}} + \frac{1}{\mu} \sum_{j=1}^{m} D_{\mathrm{KL}}(t_j \| a_j).} \tag{22}$$

Because $R_{\mathrm{dep}} = D_{\mathrm{KL}}(T \| P) - D_{\mathrm{KL}}(Q_t \| P) \le D_{\mathrm{KL}}(T \| P)$, we also obtain the directly trainable bound

$$\boxed{D_{\mathrm{KL}}(T \| Q_a) \le D_{\mathrm{KL}}(T \| P) + \frac{1}{\mu} \sum_{j=1}^{m} D_{\mathrm{KL}}(t_j \| a_j).} \tag{23}$$

The second term is the population risk estimated by Stage-1 cross-entropy, while Eq. 8 connects the first to Stage 2. If both component risks converge to zero and $\inf_n \mu_n > 0$, then the fused joint distribution is forward-KL consistent.

A.6 EXACT RECOVERY AND INTERACTION PRESERVATION

Theorem A.3 (exact recovery). Suppose the respondent model differs from the human distribution only through univariate potentials:

$$\log \frac{T(y)}{P(y)} = c + \sum_{j=1}^{m} u_j(y_j). \tag{24}$$

If $a = t$, then $Q_a = T$.

Proof. Equation 24 places T in the exponential family generated from base measure P and sufficient statistic φ. It also belongs to $\mathcal{C}(t)$. By uniqueness in Theorem A.1, it is the unique exponential tilt with those margins, so $T = Q_t = Q_a$. □

Theorem A.4 (conditional odds-ratio invariance). For two items r, s, options u, u', v, v', and fixed responses z to all remaining items, define

$$\mathrm{OR}_{rs}^{Q}(u, u'; v, v' \mid z) = \frac{Q(u, v, z) Q(u', v', z)}{Q(u, v', z) Q(u', v, z)}. \tag{25}$$

Whenever all probabilities are positive,

$$\mathrm{OR}_{rs}^{Q_a}(u, u'; v, v' \mid z) = \mathrm{OR}_{rs}^{P}(u, u'; v, v' \mid z). \tag{26}$$

Proof. The ratio $Q_a(y)/P(y)$ in Eq. 9 factorizes into one term for each item. Every factor appears once in the numerator and once in the denominator of the cross-product ratio and cancels. □

Thus the ideal projection changes only first-order log-linear potentials. It need not preserve Pearson correlations, whose values depend on both margins and interactions.

A.7 What is and is not unbiased

Let the Stage-1 anchor $\widehat{a}$ be random over training samples and suppose exact feasible projection ensures $\mathbb{E}_{Q_{\widehat{a}}}\varphi = \widehat{a}$.

Theorem A.5 (marginal and additive-estimand unbiasedness). If $\mathbb{E}_{\text{train}}\widehat{a} = t$, then

$$\mathbb{E}_{\text{train}}\mathbb{E}_{Q_{\widehat{a}}}\varphi = t. \tag{27}$$

For every additive survey statistic $f(Y) = c + \sum_j f_j(Y_j)$,

$$\mathbb{E}_{\text{train}}\mathbb{E}_{Q_{\widehat{a}}}f(Y) = \mathbb{E}_T f(Y). \tag{28}$$

Proof. The first statement follows by iterated expectation. Every function of a finite categorical item is a linear combination of its one-hot indicators; linearity proves the second. □

This covers item means, response proportions, total-score means, and construct-average means. It does *not* imply $\mathbb{E}[Q_{\widehat{a}}] = Q_t$, because projection is nonlinear in $\widehat{a}$. It therefore does not establish unbiasedness of correlations, Alpha, AVE, or complete joint cell probabilities.

Given fixed candidate weights $w_{1:M}$, randomized systematic resampling is conditionally unbiased: if N_i is the number of copies of candidate i in N outputs, then $\mathbb{E}[N_i \mid w] = Nw_i$. Hence sample averages of any $g(Y)$ equal the weighted candidate expectation in expectation, although they retain Monte Carlo variance.

A.8 Finite candidate support

The implementation uses an empirical support $\mathcal{S}_M$ consisting of complete respondent-model candidates. Let $\delta_{\text{miss}} = T(\mathcal{S}_M^c)$, $T_M = T(\cdot \mid \mathcal{S}_M)$, and $t_M = \mathbb{E}_{T_M}\varphi$. Let $\widetilde{a}$ be the feasible support-adjusted anchor, $Q_{\widetilde{a}}^M$ the empirical MCJP solution, and $Q_{t_M}^M$ the corresponding oracle. Under the empirical-support analogue of A3 with constant μ_M,

$$D_{\text{KL}}(T_M \| Q_{\widetilde{a}}^M) \le R_{\text{dep},M} + \frac{\|\widetilde{a} - t_M\|_2^2}{2\mu_M}, \tag{29}$$

where $R_{\text{dep},M} = D_{\text{KL}}(T_M \| Q_{t_M}^M)$. Since $Q_{\widetilde{a}}^M$ has zero probability outside $\mathcal{S}_M$, full-population forward KL may be infinite. Total variation remains bounded:

$$\text{TV}(T, Q_{\widetilde{a}}^M) \le \delta_{\text{miss}} + \sqrt{\frac{1}{2}\left(R_{\text{dep},M} + \frac{\|\widetilde{a} - t_M\|_2^2}{2\mu_M}\right)}. \tag{30}$$

Writing $\delta_{\text{support}} = \|\widetilde{a} - \widehat{a}\|_2$ and $\varepsilon_{\text{stage1}} = \|\widehat{a} - t\|_2$ gives

$$\|\widetilde{a} - t_M\|_2 \le \delta_{\text{support}} + \varepsilon_{\text{stage1}} + \sqrt{d}\,\delta_{\text{miss}}. \tag{31}$$

Suppose the N_g candidate questionnaires are conditionally independent draws (not necessarily identically distributed) and every required item–option event has probability at least $p_{\min} > 0$ in every draw. A union bound yields

$$\Pr(\text{any required event absent}) \le d' \exp(-N_g p_{\min}), \tag{32}$$

where d' is the number of required positive-margin events. This guarantees only marginal feasibility, not coverage of all exponentially many questionnaires.

A.9 Why item marginals cannot identify a respondent

For two binary items, define $T(0,0) = T(1,1) = 1/2$ and $R(0,1) = R(1,0) = 1/2$. Every item marginal agrees, but the supports are disjoint and $\text{JSD}(T, R) = \log 2$. Conversely, the data-processing inequality for f-divergences implies

$$\text{JSD}(g_{\#}T, g_{\#}Q) \le \text{JSD}(T, Q) \tag{33}$$

for every item subset or construct mapping g. Complete-joint closeness implies low-dimensional closeness; low-dimensional closeness does not imply complete-joint closeness.

A.10 PSYCHOMETRIC SUMMARIES

Cronbach's Alpha is a nonlinear covariance functional. Scale r items to $[0, 1]$ and let Σ_R be their covariance under distribution R:

$$\alpha(R) = \frac{r}{r-1}\left(1 - \frac{\mathrm{tr}(\Sigma_R)}{\mathbf{1}^\top \Sigma_R \mathbf{1}}\right). \tag{34}$$

On finite response spaces, total variation controls all bounded first and second moments. Consequently, Alpha is locally continuous whenever total-score variance is bounded away from zero. This is a regularity statement, not a uniform guarantee: near a degenerate total score, the ratio can be unstable. AVE additionally depends on the stability of a fitted factor model; local continuity requires item variances bounded away from zero, a unique CFA solution, and a nonsingular Jacobian of its first-order conditions. We therefore make no unconditional Alpha or AVE claim. They are held-out structural diagnostics that are neither constrained nor optimized by MCJP.

B PROMPTS AND TRAINING DETAILS

B.1 CONSTRUCT-FREE PROMPT PROTOCOL

Both stages use the same item wording and response scale. Construct names and definitions are omitted. The ESS11 template is schematically:

```
Survey respondent profile
Country: [country]
Gender: [gender]
Age band: [age]
Education band: [education]
Domicile: [domicile]
Response scale: 1 = Very much like me;
...; 6 = Not like me at all.
Answer as the same respondent throughout the questionnaire.
[Optional previous question and sampled/true answer history]
Question: [official item text]
Answer:
```

Stage 1 omits history (Figure 5). Stage 2 teacher-forces true history during training and uses generated history at inference (Figure 6). The visuals show two ESS items and illustrative observed labels; they abbreviate the repeated instruction line in the serialized header above. TALIS replaces the background fields and uses its official four-option scale. Direct option-token scoring prevents prose, refusal, or formatting variation from being interpreted as a response.

B.2 OPTIMIZATION

Both stages use LoRA rank 8, scaling 16, dropout 0.05, bfloat16 arithmetic, SDPA attention, gradient accumulation 16, and seed 20260727. ESS uses learning rate 10^{-4}, Stage-1 batch size 2 for one epoch, and Stage-2 batch size 1 for at most four epochs with patience 2. TALIS uses batch size 1 for at most ten epochs with patience 2; the Qwen and Ministral learning rates are 2×10^{-4} and 10^{-4}, respectively. Stage 2 averages validation loss over three sampled item orders. Checkpoint selection uses development-country validation, never test-country labels. Stage 1 selects item negative log-likelihood; Stage 2 selects respondent-level autoregressive negative log-likelihood.

For the final projection, the respondent model generates 16 candidate questionnaires per target respondent and IPF operates within population cells. Three outputs per target are drawn by randomized systematic resampling. All baselines also produce three final rollouts. ESS analysis weights and TALIS teacher weights are retained when aggregating human and generated distributions.

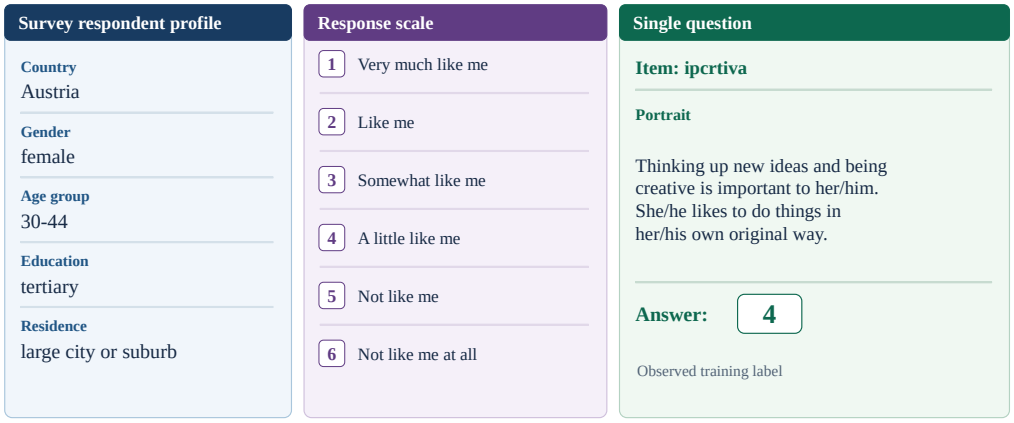


Figure 5: Stage-1 single-item training prompt, illustrated with one ESS11 response. The model sees the respondent profile, option scale, and one portrait; cross-entropy supervises the observed answer token. No previous question or construct name is supplied.

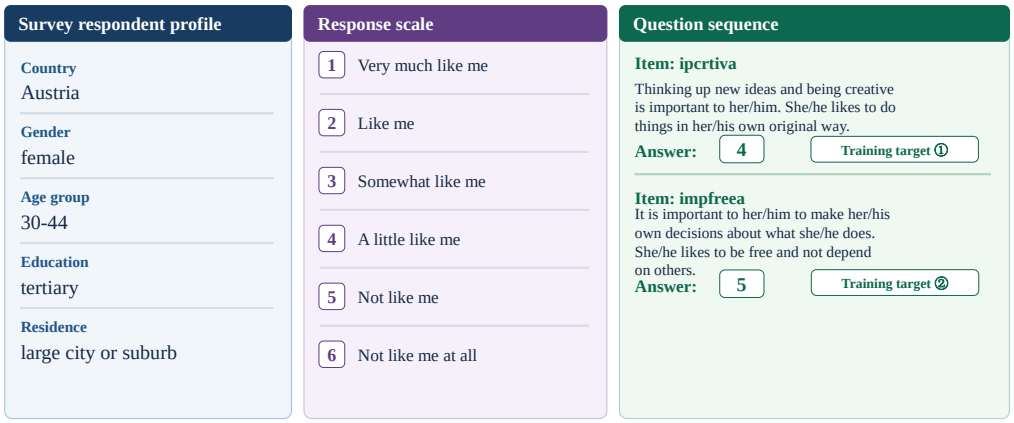


Figure 6: Stage-2 linked-question training prompt for the same respondent. Each preceding observed answer is available when the next answer is predicted, and losses at all answer positions are included in one respondent-level training example. Actual training randomizes item order and uses all eligible items, rather than only the two pictured here.

C Data Protocol Details

C.1 ESS11

Eligibility requires valid 1–6 responses for all 21 PVQ-21 items and known gender and age, leaving 42,551 respondents. Education and domicile nonresponse are explicit "unknown" prompt values. The train countries are disjoint from five development and five test countries. The training constructs are Universalism (3 items), Security (2), Achievement (2), and Hedonism (2). The unseen constructs are Self-Direction, Power, Stimulation, Conformity, Tradition, and Benevolence, each with two items. M1 uses 144 populated country–gender–age cells; M2 and M3 use 40 held-out-country cells.

C.2 TALIS 2018

The international teacher file contains 261,426 records, of which 245,450 have valid answers to all 12 self-efficacy items. Classroom Management and Instruction contain four items each and are observed in training; Student Engagement contains four held-out items. Thirty-three systems form the training pool, seven are development-only, and seven are test-only. The frozen training sample has 12,000

teachers; each mode has 8,000 generated targets. Distribution metrics group by country/system and gender while the prompt additionally contains education and first-career-choice status.

D METRIC DEFINITIONS

Item JSD. For each populated evaluation cell g and item j, human and generated weighted histograms over response categories are normalized to distributions p_{gj} and q_{gj}. We compute

$$\mathrm{JSD}(p,q) = \tfrac{1}{2} D_{\mathrm{KL}}(p \| \tfrac{1}{2}(p+q)) + \tfrac{1}{2} D_{\mathrm{KL}}(q \| \tfrac{1}{2}(p+q)), \tag{35}$$

then average over cells and items. Empty cells are not invented; the frozen protocol uses only populated human cells. A Dirichlet pseudocount of 0.5 is applied to the human reference histogram where specified by the saved protocol.

Construct JSD. For construct c, each respondent contributes one categorical tuple in the Cartesian product of its item response spaces. We compare human and generated tuple histograms and average the resulting JSD after calculation across constructs. Paired-construct JSD applies the same operation to all item tuples in $c \cup c'$. It is not the JSD between an average item histogram and therefore cannot cancel item-specific errors.

Reliability and convergent validity. Cronbach's Alpha is computed from the respondent-level item covariance matrix for every official construct. AVE is computed from standardized one-factor CFA loadings using $\mathrm{AVE} = r^{-1} \sum_{j=1}^{r} \lambda_j^2$. We report the mean absolute difference between generated and human construct values, not the generated value alone. Failed or degenerate fits are recorded rather than silently replaced.

Human sampling reference. Complete respondents, not individual answers, are resampled within country–gender–age strata with survey weights. We run 1,000 replicates at the model evaluation sample size. "Human versus human" compares two independent resamples and is an empirical sampling reference, not a universal information-theoretic lower bound.

E TWO-CONSTRUCT RESULTS

The main tables omit two-construct JSD for readability. Table 8 reports it separately. TALIS evaluates only one construct in M1 and M3, so the metric is undefined in those modes.

F ADDITIONAL ANALYSES

F.1 PROPOSAL BUDGET

F.2 MARGINAL-ANCHOR INTERPOLATION

Holding the same 16-candidate proposal fixed, we interpolate between its own margins ($\rho = 0$) and the learned Stage-1 anchor ($\rho = 1$). The monotone improvement in the two joint metrics shows that the gain is produced by constrained fusion rather than by regenerating candidates. Effective-sample fraction decreases because stronger correction necessarily concentrates weights.

An analysis-only oracle-smoothed anchor derived from held-out labels reaches Item/construct/two-construct JSD of .0024/.0022/.0202. Shuffling learned anchors across background cells instead yields .0450/.0226/.0622. The oracle is not a deployable method; together, these controls isolate anchor quality and correct background assignment as the relevant mechanism.

F.3 CONTROLLABILITY AND RESPONSE BEHAVIOR

The ESS11 country-swap analysis pairs each source country with a target country and compares item-mean changes generated by changing only the prompt field with observed human differences. Across 216 country–item effects, Pearson correlation and cosine similarity are 0.655 and direction

Table 8: Two-construct JSD. Lower is better. Bold is best and underline is second best within each dataset, backbone, and applicable mode.

Data / backbone	Method	M1	M2	M3
ESS11 / Qwen	Zero-shot	.247	.366	.296
	Single FT	<u>.099</u>	<u>.142</u>	<u>.160</u>
	Sequential FT	.110	.165	.184
	FR-LLM	**.059**	**.081**	**.123**
ESS11 / Ministral	Zero-shot	.291	.376	.340
	Single FT	<u>.119</u>	<u>.150</u>	<u>.182</u>
	Sequential FT	.179	.171	.230
	FR-LLM	**.080**	**.086**	**.145**
TALIS / Qwen	Zero-shot	–	.571	–
	Single FT	–	.347	–
	Sequential FT	–	<u>.150</u>	–
	FR-LLM	–	**.115**	–
TALIS / Ministral	Zero-shot	–	.674	–
	Single FT	–	.359	–
	Sequential FT	–	<u>.196</u>	–
	FR-LLM	–	**.123**	–

Table 9: ESS11 M1 candidate-support audit. Item JSD remains anchored while joint support and feasibility improve.

Candidate questionnaires	Item JSD	Two-construct JSD	Maximum marginal error	Adjusted cells	Effective sample fraction
1	.0244	.0693	.1952	469	.330
2	.0244	.0621	.0748	201	.422
4	.0244	.0596	.0190	76	.506
8	.0244	.0596	.0112	18	.555
16	.0244	.0592	.0013	2	.578

agreement is 0.724 for effects larger than 0.02. Removing all backgrounds worsens Alpha MAE from 0.078 to 0.102, AVE MAE from 0.071 to 0.094, Item JSD from 0.024 to 0.042, and pair JSD from 0.059 to 0.096.

We additionally evaluate targeted background changes on the 12 held-out ESS11 items. For each source–target rule r and item j, $\Delta_{rj}^{\text{human}}$ is the weighted difference between the corresponding real target and source groups. $\Delta_{rj}^{\text{model}}$ is the change obtained by editing only the designated prompt fields for the same source respondents. Held-out human target responses are used only to score this audit. We report

$$\text{ShiftMAE}_c = \frac{1}{|\mathcal{R}_c|m} \sum_{r \in \mathcal{R}_c} \sum_{j=1}^{m} \left| \Delta_{rj}^{\text{model}} - \Delta_{rj}^{\text{human}} \right|, \tag{36}$$

which jointly penalizes incorrect direction and magnitude in ESS response-scale points. Table 5 compares the five contrasts shared by all three methods, with equal weight in the macro average.

A gender-only contrast was run for the two learned methods but excluded from the three-method aggregate because Zero-shot was not generated; FR-LLM and Single FT obtain .081 and .079, respectively. The differences in Table 5 are not statistically resolved under one generation seed, and prompt-field edits do not identify causal demographic effects.

A respondent-bootstrap audit also compares extreme-response rate, central-response rate, straightlining, within-person standard deviation, entropy, item-correlation RMSE, and construct-profile correlations. The results are mixed across these secondary diagnostics: FR-LLM improves several profile-level errors over Single FT and Sequential FT, but not every response-style statistic. We therefore do not treat marginally constrained dependence as a complete behavioral model.

Table 10: ESS11 M1 anchor path. The proposal and candidate support are fixed; only the target margin changes.

ρ	Item JSD	Construct JSD	Two-construct JSD	Support adjustment	Effective sample fraction
0.00	.0267	.0280	.0708	.0000	1.000
0.25	.0251	.0254	.0661	.0013	.962
0.50	.0242	.0240	.0636	.0027	.860
0.75	.0240	.0224	.0606	.0040	.722
1.00	.0244	.0217	.0596	.0053	.578

F.4 Option-label robustness

A 256-respondent ESS11 M1 pilot reverses the displayed option labels and measures how much each method's predicted distribution moves. Mean item-level shift JSD is 0.123 for Single FT and 0.112 for FR-LLM. Thus the complete method attenuates, but does not eliminate, sensitivity to option presentation. Because this is a small diagnostic rather than the frozen full-population protocol, we do not mix it into the primary tables.

F.5 Synthetic-to-real downstream validity

We also ask whether a complete synthetic questionnaire carries usable dependence for leave-one-item-out imputation. For each evaluation mode, an L2-regularized multinomial logistic model is trained on a 70% synthetic split to predict each item from the other items and demographics, then tested on the matched 30% held-out human split. In M1, mean NLL is 1.411 when trained on human questionnaires and 1.421 when trained on FR-LLM questionnaires. The small gap shows that the generated records retain useful multivariate signal, although this task-specific diagnostic is not evidence that synthetic data can replace human data.

The apparently high M2/M3 NLL of the human-trained reference is not a claim that synthetic data is superior to human data. The reference classifier is trained on one finite human split under country shift, whereas synthetic generators pool inductive bias from model pretraining and the survey fine-tuning protocol. The comparison is a task-specific validity diagnostic, not a replacement argument.

F.6 Little-Treat decision simulation

We use our Little-Treat Consumption questionnaire. Of 488 complete records, 294 train the survey models, 97 are used for development, and 97 are held out for evaluation. We retain six attitude items, purchase frequency, and the most recent purchase-price range. The latter is treated as a proxy spending ceiling, not an elicited maximum willingness to pay. For a candidate price p, the joint synthetic answers estimate monthly demand per respondent $\widehat{D}_m(p)$ by weighting the frequency of respondents whose proxy ceiling is at least p. The same calculation on held-out human answers gives $D_{\mathrm{real}}(p)$.

The original saved pricing run used a unit cost of 8 CNY and stored each method's predicted profit curve. The pricing-and-stocking analysis in Figure 4 is a subsequent recomputation from those curves, *not* a new model-generation run or a preregistered cost comparison. For each price in {10,20,40,80,150,300,600} CNY, we recover predicted demand as the stored profit divided by $p - 8$. At a hypothetical new unit cost c, each method selects $p_m = \arg\max_p (p - c)\widehat{D}_m(p)$ and stocks $Q_m = \widehat{D}_m(p_m)$ per respondent. For zero salvage, realized proxy profit per respondent is $p_m \min(Q_m, D_{\mathrm{real}}(p_m)) - cQ_m$; unsold stock is $(Q_m - D_{\mathrm{real}}(p_m))_+$. Main-text totals multiply these per-respondent quantities by 1,000 for readability; only 97 respondents were actually held out. Figure 4 converts costs and profits to USD equivalents at a fixed 7 CNY per USD solely for presentation; all decisions were computed in the original currency.

At costs of 50 and 150 CNY, FR-LLM has the highest profit point estimate, and its lead also remains under salvage values of 25% or 50% of unit cost. Zero-shot selects the same 300 CNY price but predicts 1.197 monthly purchases per respondent there, versus 0.317 in the held-out records, causing

substantial overstock. In 10,000 paired bootstrap resamples of the 97 human test respondents with model decisions fixed, the 95% intervals for FR-LLM minus Sequential FT include zero at both costs. These results illustrate sensitivity to the joint frequency–spending distribution; they do not establish a significant advantage over that baseline, causal price elasticity, or revenue realized in a live market.