# ADAPTIVE RESOURCE ALLOCATION FOR EFFECTIVE AND EFFICIENT LLM SOCIAL SURVEY SIMULATION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Large Language Models (LLMs) enable scalable social survey simulation, yet existing pipelines typically assign the same strong general-purpose model and a fixed, often large, respondent history to every respondent–question request. This uniform design overlooks three properties of survey simulation. ***(1) Model capability.*** A stronger general-purpose model may rely more on its own knowledge and deviate from respondent-specific evidence, while also costing more. ***(2) Respondent history.*** More history can help when evidence is insufficient, but irrelevant responses may introduce noise and lengthen the input. ***(3) Model–history interaction.*** Changing the history budget can change the preferred model, while changing the model can change the preferred history budget. These considerations call for joint request-level allocation under an explicit accuracy–cost trade-off. We propose ***E$^2$Sim***, an adaptive resource allocation framework that treats each model–history configuration as a joint request-level decision. Conditioned on the respondent persona, ranked response history, and target question, a lightweight policy predicts the accuracy–cost score of every configuration and selects the most suitable one. To reduce sensitivity to incomplete or variable histories, we apply respondent-history drop-and-swap augmentation; to handle differences in allocation ambiguity, we use a margin-based curriculum that progresses from clearly separated configurations to harder cases. Experiments on four real-world social-survey datasets, multiple model pools, and different history-budget spaces demonstrate improvements over oracle best-fixed configurations, with gains of up to 6.3 percentage points in Accuracy and reductions of up to 55.3% in cost. Our code is available at https://anonymous.4open.science/r/E2Sim-DE66.

## 1 INTRODUCTION

Social surveys are the empirical backbone of computational social science: they record how attitudes, values, and policy preferences are distributed across a heterogeneous population (Squazzoni et al., 2020; Hamill and Gilbert, 2009; Wang et al., 2025b). Such evidence, however, is hard to scale and expensive to obtain, since fielding a questionnaire repeatedly and at scale requires substantial human effort (Wright et al., 2010; Heffetz and Reeves, 2019; Kalton, 2009; Groves et al., 2011; Roopa and Rani, 2012; Tourangeau et al., 2000). Large Language Models (LLMs) (Minaee et al., 2024; Achiam et al., 2023) offer a scalable complement: conditioned on a respondent's demographic profile and past answers, an LLM can simulate how that specific person would answer a new question (Argyle et al., 2023; Aher et al., 2023; Cao et al., 2023; Durmus et al., 2023). Given a target question the respondent has not answered, a simulator retrieves the $k$ past responses of that respondent most relevant to the question, which serve as personal references that ground the prediction in the respondent's own expressed views, and prompts a model $m$ with the respondent's persona, these $k$ responses, and the question to predict the answer. For example, for the question *"Is the government spending too much, too little, or about the right amount on healthcare?"*, the input is a retired respondent's persona together with that respondent's retrieved earlier answers on government responsibility and on medical expenses, and the output is one of the given options.

However, existing LLM-based survey simulation pipelines typically use the same high-capability general-purpose model and a fixed, often large, respondent history for every respondent–question request. This one-size-fits-all assumption overlooks two forms of heterogeneity and their interaction: ***(1) Model capability (m).*** Different survey requests require different levels of model capability. A



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.

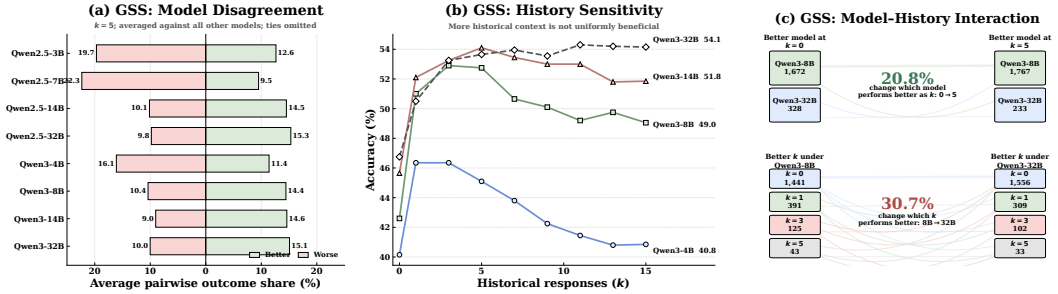



Figure 1: Analysis of 2,000 GSS respondent–question requests. (a) Proportions of pairwise wins and losses for each model. (b) Accuracy as the history budget varies. (c) Model–history interaction: 20.8% of requests change their better-performing model as $k$ increases from 0 to 5, while 30.7% change their better-performing $k$ when switching from Qwen3-8B to Qwen3-32B.

stronger general-purpose model is not always better for survey simulation, because the goal is to predict a respondent's likely choice from limited personal evidence rather than rely solely on general knowledge to derive the most reasonable answer. Requests with sparse or conflicting evidence may benefit from stronger reasoning capabilities. However, when the respondent's pattern is clear, a smaller model may already suffice, while a stronger model may introduce additional priors that are less aligned with the respondent-specific pattern and incur higher inference cost. As shown in Figure 1(a), no single model performs best across all requests, motivating adaptive selection of $m$ at the request level rather than using one fixed model for every request. ***(2) Respondent history (k).*** Different survey requests require different amounts of respondent history. We rank a respondent's historical responses by their relevance to the target question and include the most relevant ones first. If the most relevant responses are insufficient, increasing $k$ may provide useful evidence. However, once the available history is sufficient, further increases mainly introduce lower-ranked responses that may add little value or distract the model, while increasing prompt length and inference cost. As shown in Figure 1(b), larger $k$ can yield diminishing or even negative returns, motivating adaptive selection of the history budget rather than using one fixed value for every request. ***(3) Model–history interaction (m, k).*** Model choice and history budget are interdependent. For the same request, changing the model $m$ may change the most suitable history budget $k$, because models differ in how effectively they extract useful signals from respondent history. Conversely, changing the history budget $k$ may change the most suitable model $m$, because the amount and quality of available evidence affect how much model capability is useful. As shown in Figure 1(c), increasing $k$ from 0 to 5 changes the better-performing model for 20.8% of requests. Switching from Qwen3-8B to Qwen3-32B changes the better-performing history budget for 30.7% of requests. This interaction motivates selecting $(m, k)$ jointly at the request level rather than choosing the model and history budget independently.

Motivated by these observations, we propose **E$^2$Sim**, an adaptive resource allocation framework for **E**ffective and **E**fficient LLM social survey **Sim**ulation. E$^2$Sim treats each respondent–question pair as a distinct allocation unit. We score each candidate configuration by its Accuracy–Cost Trade-off (ACT), $ACT = \text{Accuracy} - \alpha \cdot \text{Cost}$, in which the hyperparameter $\alpha$ sets how heavily inference cost is penalized relative to a correct prediction. Using offline evaluations of candidate configurations as supervision, a lightweight policy learns to predict their expected ACT from the respondent persona, ranked response history, and target question, and jointly selects $(m, k)$ at inference time. Learning this policy raises two challenges. The first is that the retrieved history is not fixed: the same respondent may have fewer responses available in another setting, and a small change in retrieval similarity can change their order. If the policy is trained on a single version of the history, it may tie its decision to that exact version and choose a different configuration when one response is missing or two are reordered. We therefore train on perturbed histories, dropping and swapping retrieved responses, so that the policy gives the same request a consistent allocation and decides from how informative the history is as a whole. The second is that requests are not equally easy to allocate. When the history clearly points to an answer, one configuration stands out; when the evidence is sparse or self-contradictory, several configurations perform almost identically and the supervision is close to a tie. We quantify this difficulty by the ACT margin between the best two configurations, and train with a curriculum that starts from large-margin requests and gradually adds ambiguous ones.

**Contributions.** ❶ We identify complementary, request-dependent roles of model capability and respondent history, and show that their benefits are interdependent. ❷ We propose E$^2$Sim, a cost-aware

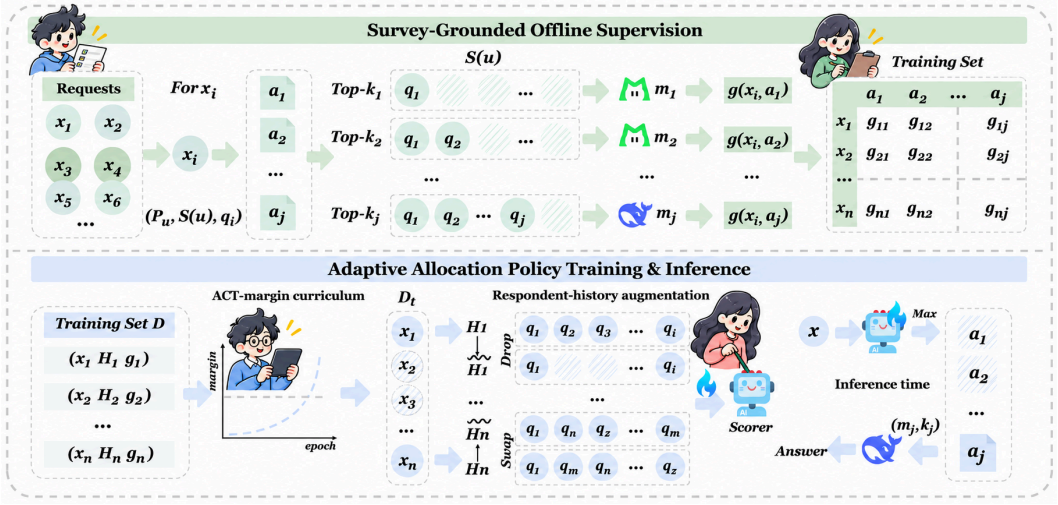



Figure 2: Overview of E$^2$Sim. The upper panel shows the construction of survey-grounded supervision: it retrieves target-relevant respondent history, evaluates every model–history action, and records the complete ACT vector. The lower panel trains a joint configuration scorer with respondent-history Drop/Swap and an ACT-margin curriculum; at inference, the policy invokes only its selected $(m, k)$.

joint allocation framework with respondent-history augmentation and a margin-based curriculum for robust policy learning. ❸ Across four surveys, we evaluate E$^2$Sim under varied model pools and history budgets, demonstrating its effectiveness and robustness.

## 2 Problem Formulation and Definition

**Formulation: LLM-based Social Survey Simulation.** Let $u$ denote a respondent with a textual persona $P_u$, and let $S(u)$ denote the respondent's observed responses to other survey questions. For a target question $q$, including its response options, a simulation request is defined as

$$x = (P_u, S(u), q). \tag{1}$$

The amount of respondent history provided as personalized context is controlled by a history budget $k$. Given $k$, the simulator retrieves an ordered context $H_k$ containing the $k$ historical responses most relevant to the target question, and model $m$ predicts the respondent's answer:

$$H_k = \mathrm{Retrieve}_k(q, S(u)), \qquad \hat{y}_{m,k} \sim P_m(\cdot \mid P_u \oplus H_k \oplus q), \tag{2}$$

where $\oplus$ denotes prompt concatenation.

**Definition: Adaptive Resource Allocation.** Let $\mathcal{M}$ be the candidate model pool and $\mathcal{K}$ the available history budgets. Their Cartesian product defines the joint action space

$$\mathcal{A} = \mathcal{M} \times \mathcal{K}, \qquad a = (m, k) \in \mathcal{A}. \tag{3}$$

An allocation policy maps each request to one joint action,

$$\pi : \mathcal{X} \to \mathcal{A}, \qquad \pi(x) = (m, k). \tag{4}$$

This joint formulation captures the interdependence between model choice and history budget: the relative performance of candidate models can change with $k$ for a given respondent–question request, while the predictive value of additional history can likewise change with $m$.

## 3 Method

In this section, we present E$^2$Sim, the adaptive resource allocation framework illustrated in Figure 2. We first construct survey-grounded offline supervision by evaluating every candidate model–history configuration against observed survey responses (Section 3.1). We then learn a joint model–history allocation policy with respondent-history augmentation and an ACT-margin curriculum (Section 3.2). At deployment, request-level adaptive inference selects one model and history budget for each request and invokes the chosen configuration (Section 3.3).

### 3.1 SURVEY-GROUNDED OFFLINE SUPERVISION

Each training request $x_i$ is paired with the response $y_i$ recorded in the survey. We run every action $a = (m, k)$ on the same request, using model $m$ and the top-$k$ historical responses relevant to its target question. The generated answer, input-token usage and model price determine the Accuracy, Cost, and ACT of that action. Collecting these scores across $\mathcal{A}$ gives one complete supervision vector $\mathbf{g}_i$ for learning which configuration suits the request.

Respondent histories span many social topics, so the context is formed relative to the target question. Let $S(u_i)$ be respondent $u_i$'s observed response history. A fixed embedding model ranks its questions by cosine similarity and returns the top-$k$ question–answer pairs:

$$\mathbf{e}(q) = \mathrm{Embed}(q), \quad s(q_i, q) = \frac{\mathbf{e}(q_i)^\top \mathbf{e}(q)}{\|\mathbf{e}(q_i)\|_2 \, \|\mathbf{e}(q)\|_2}, \quad H_{i,k} = \mathrm{TopK}_{\substack{(q',y') \in S(u_i) \\ q' \neq q_i}} s(q_i, q'). \tag{5}$$

For each action $a = (m, k)$, model $m$ receives the persona, retrieved history, and target question. Its output and exact-match accuracy against the observed response $y_i$ are

$$\hat{y}_i(a) \sim P_m(\cdot \mid P_{u_i} \oplus H_{i,k} \oplus q_i), \qquad \mathrm{Acc}(x_i, a) = \mathbb{1}[\hat{y}_i(a) = y_i]. \tag{6}$$

Because different actions trade predictive accuracy against inference cost, we define a unified score that captures both. The cost of an action depends on two factors: model choice determines the token price, while the history budget determines the prompt length. Since the structured output contains only an option index, we measure inference cost using only price-weighted input tokens. To make cost comparable with accuracy, we normalize it to the $[0, 1]$ scale and define

$$C(x_i, a) = \frac{c_m^{\mathrm{in}} T(x_i, a)}{\max\limits_{1 \le j \le N,\ a' = (m', k') \in \mathcal{A}} c_{m'}^{\mathrm{in}} T(x_j, a')}, \qquad g_i(a) = \mathrm{Acc}(x_i, a) - \alpha C(x_i, a), \tag{7}$$

where $T(x_i, a)$ is the number of input tokens, $c_m^{\mathrm{in}}$ is the model-specific input-token price, and $g_i(a)$ is the Accuracy–Cost Trade-off score (ACT). The coefficient $\alpha$ controls the trade-off between accuracy and cost. Evaluating every action gives the complete score vector $\mathbf{g}_i = (g_i(a))_{a \in \mathcal{A}}$. With $H_i = H_{i,k_{\max}}$ and $k_{\max} = \max \mathcal{K}$, the offline supervision is

$$\mathcal{D} = \{(x_i, H_i, \mathbf{g}_i)\}_{i=1}^{N}. \tag{8}$$

Thus, each request retains the relative accuracy–cost outcomes of every model–history alternative rather than only a best-action label.

### 3.2 JOINT MODEL–HISTORY ALLOCATION POLICY

**Joint configuration scoring.** Each survey request provides three complementary signals for allocation: the persona describes relatively stable sociodemographic characteristics, the respondent history records issue-specific choices, and the target question specifies the attitude to be inferred. We combine them into the request representation

$$z_i = P_{u_i} \oplus H_i \oplus q_i, \tag{9}$$

where $\oplus$ denotes concatenation and $H_i = H_{i,k_{\max}}$ is the context under the predefined maximum history budget $k_{\max} = \max \mathcal{K}$. Providing the policy with the maximum available history gives it sufficient information to compare configurations with different values of $k$. A lightweight scorer $f_\theta$ then predicts the ACT of every joint action in one pass:

$$f_\theta(z_i) = \widehat{\mathbf{g}}_\theta(z_i) = \big(\widehat{g}_\theta(z_i, a)\big)_{a \in \mathcal{A}} \in \mathbb{R}^{|\mathcal{A}|}. \tag{10}$$

Each coordinate corresponds to one model–history configuration and is supervised by its matching entry in $\mathbf{g}_i$. Joint scoring lets the same survey evidence determine the predicted values of both $m$ and $k$, preserving their interaction within a single allocation decision.

**Robustness to incomplete and variable histories.** In practice, response histories may be incomplete due to item nonresponse, while small differences in retrieval similarity can alter the ordering of otherwise comparable responses. If the policy relies on memorizing whether a response appears or

where it appears in the history, its allocations may become unstable. To improve robustness, we construct alternative views of the same request by dropping and swapping responses:

$$\widetilde{H}_i = \mathrm{Swap}_{p_{\mathrm{swap}}}\Big(\mathrm{Drop}_{p_{\mathrm{drop}}}(H_i)\Big), \qquad \widetilde{z}_i = P_{u_i} \oplus \widetilde{H}_i \oplus q_i. \tag{11}$$

where $p_{\mathrm{drop}}$ is the probability of independently removing each response, and $p_{\mathrm{swap}}$ is the probability of exchanging the positions of two selected responses. These limited perturbations preserve the respondent persona and target question while varying only the retrieved history, yielding alternative views of the same request. Training on these views acts as a local invariance regularizer, encouraging stable allocation under small changes in history content and order.

**ACT-margin curriculum for allocation ambiguity.** Survey requests differ in how clearly their available evidence identifies an appropriate resource configuration. Target-relevant history may sharply separate one action from the alternatives, whereas sparse or mixed evidence can leave several actions nearly tied. The fully observed score vector makes this allocation ambiguity directly measurable. Let $g_i^{(1)}$ and $g_i^{(2)}$ be the largest and second-largest action scores; their margin is

$$\Delta_i = g_i^{(1)} - g_i^{(2)}. \tag{12}$$

A large $\Delta_i$ indicates a clear allocation preference, while a small margin indicates ambiguity. Rather than exposing the policy to both from the outset, we begin with large-margin requests and expand the admitted fraction. At epoch $t$, the fraction $\rho_t$ and corresponding training subset $\mathcal{D}_t$ are

$$\rho_t = \begin{cases} \rho_0 + (1-\rho_0)\dfrac{t-1}{T-1}, & 1 \le t \le T, \\ 1, & t > T, \end{cases} \qquad \mathcal{D}_t = \mathrm{Top}_{\lceil \rho_t N \rceil}\left(\{x_i\}_{i=1}^N;\, \Delta_i\right). \tag{13}$$

where $\rho_0$ is the initial fraction, $T$ is the number of curriculum epochs, and $\mathcal{D}_t$ contains the requests with the largest margins. Thus, the policy first learns from requests with clear resource requirements and progressively incorporates more ambiguous allocation decisions as $\rho_t$ increases.

**Joint score regression.** A best-action label would discard how strongly one configuration is preferred: a narrow and a decisive ACT gap would receive identical supervision. This distinction matters for survey requests whose allocation ambiguity varies. We therefore regress the complete score vector. At curriculum epoch $t$, the Smooth-$L_1$ objective is

$$\mathcal{L}(\theta) = \frac{1}{|\mathcal{D}_t|} \sum_{x_i \in \mathcal{D}_t} \ell_{\text{smooth-}L_1}(\widehat{\mathbf{g}}_\theta(\widetilde{z}_i), \mathbf{g}_i). \tag{14}$$

Each coordinate is supervised by the observed score of the corresponding model–history action, preserving both the ranking and the utility differences among all candidate configurations.

### 3.3 Request-Level Adaptive Inference

At deployment, each respondent–question request is allocated independently. We retrieve the predefined maximum history once and form $z(x) = P_u \oplus H_{k_{\max}}(u, q) \oplus q$. The policy directly selects

$$(m^*, k^*) = \arg\max_{(m,k)\in\mathcal{A}} \widehat{g}_\theta(z(x), (m, k)). \tag{15}$$

The policy examines up to $k_{\max}$ responses and selects a survey model $m^*$ and history budget $k^*$ for each request. The selected model $m^*$ then predicts the answer using only the top $k^*$ responses, so the additional history examined by the policy does not increase the survey model's prompt length. At deployment, each request requires one policy evaluation and one call to the selected survey model.

## 4 Experiments

We investigate four research questions. **(RQ1)** How effectively does E$^2$Sim improve the accuracy–cost trade-off, and how quickly can its offline supervision cost be amortized? **(RQ2)** How does $\alpha$ control the cost–accuracy trade-off? **(RQ3)** How robust is E$^2$Sim across different history-budget spaces and model-pool sizes? **(RQ4)** How much does each component contribute?

Table 1: Overall performance under two model configurations. Improved reports the difference between E$^2$Sim and the strongest baseline by ACT.

| Dataset | Metric | Qwen3-8B + Qwen3-32B | | | | | | Qwen3-8B + DS-V4-Flash | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | WCR | Random | LLM-S | BF | Ours | Improved | WCR | Random | LLM-S | BF | Ours | Improved |
| ANES | Accuracy | 0.2047 | 0.5737 | 0.6360 | 0.6552 | 0.6880 | **0.0328** | 0.2097 | 0.5880 | 0.6585 | 0.6850 | 0.7063 | **0.0213** |
| | Cost | 0.5559 | 0.3917 | 1.5882 | 0.5736 | 0.3614 | **0.2122** | 0.5399 | 0.3910 | 1.6052 | 0.5690 | 0.3577 | **0.2113** |
| | ACT | 0.1770 | 0.5542 | 0.5566 | 0.6266 | 0.6699 | **0.0433** | 0.1828 | 0.5684 | 0.5782 | 0.6565 | 0.6884 | **0.0319** |
| BSA | Accuracy | 0.1092 | 0.4275 | 0.4675 | 0.4825 | 0.5102 | **0.0277** | 0.1095 | 0.4472 | 0.4843 | 0.5155 | 0.5222 | **0.0067** |
| | Cost | 0.5871 | 0.3848 | 1.5262 | 0.5680 | 0.3870 | **0.1810** | 0.5692 | 0.3831 | 1.5458 | 0.5620 | 0.3706 | **0.1914** |
| | ACT | 0.0799 | 0.4083 | 0.3912 | 0.4541 | 0.4909 | **0.0368** | 0.0810 | 0.4281 | 0.4070 | 0.4874 | 0.5037 | **0.0163** |
| GSS | Accuracy | 0.1787 | 0.5085 | 0.5242 | 0.5278 | 0.5543 | **0.0265** | 0.1872 | 0.5210 | 0.5385 | 0.5785 | 0.5790 | **0.0005** |
| | Cost | 0.5445 | 0.3582 | 1.3959 | 0.3205 | 0.2873 | **0.0332** | 0.5272 | 0.3556 | 1.3874 | 0.6370 | 0.4738 | **0.1632** |
| | ACT | 0.1515 | 0.4906 | 0.4545 | 0.5117 | 0.5399 | **0.0282** | 0.1609 | 0.5032 | 0.4691 | 0.5466 | 0.5553 | **0.0087** |
| WVS | Accuracy | 0.1888 | 0.4973 | 0.5407 | 0.5533 | 0.5687 | **0.0154** | 0.1935 | 0.5258 | 0.5447 | 0.5935 | 0.5998 | **0.0063** |
| | Cost | 0.4935 | 0.3138 | 1.2424 | 0.6050 | 0.2704 | **0.3346** | 0.4853 | 0.3188 | 1.2817 | 0.4710 | 0.3457 | **0.1253** |
| | ACT | 0.1641 | 0.4816 | 0.4786 | 0.5230 | 0.5552 | **0.0322** | 0.1692 | 0.5098 | 0.4807 | 0.5700 | 0.5825 | **0.0125** |

### 4.1 Experimental Setup

**Datasets.** Experiments are conducted on four real-world social-survey datasets: **ANES** (American National Election Studies, 2024 wave), **BSA** (British Social Attitudes, 2024 wave), **GSS** (General Social Survey, 2024 wave), and **WVS** (World Values Survey, Wave 7). Dataset statistics and split construction are detailed in Appendix A.

**Model and History Configuration.** The candidate model pool consists of three LLMs: Qwen3-8B, Qwen3-32B, and DeepSeek-V4-Flash. These models span a range of inference costs and reasoning capabilities. For our controlled cost model, we assign relative input-token weights of 2, 4, and 4 to Qwen3-8B, Qwen3-32B, and DeepSeek-V4-Flash, respectively. These weights are used consistently in the normalized Cost and ACT calculations. The history-budget space is $\mathcal{K} = \{0, 1, 3, 5\}$, where $k$ specifies the number of historical responses and $k = 0$ denotes the no-history setting.

**Baselines.** We compare E$^2$Sim with four baselines: (1) Worst-Case Reference (WCR), which has access to the observed action scores and selects the configuration with the lowest ACT for each request, providing a worst-case reference; (2) Random, which samples a model–history configuration uniformly from $\mathcal{A}$; (3) LLM-based Selector (LLM-S), which uses Qwen3-32B with the prompt in Appendix B.2 to predict a suitable configuration, with the associated Qwen3-32B inference cost included in the reported Cost; and (4) Oracle Best Fixed (BF), which uses the observed action scores on the evaluation split to select the single configuration with the highest average ACT and applies it to every request. BF is therefore an oracle upper bound for policies restricted to one fixed configuration.

**Evaluation Metrics.** We report three metrics: (1) Accuracy, the proportion of requests where the predicted answer matches the ground truth; (2) Cost, the average price-weighted input-token cost normalized by the maximum weighted cost among all candidate configurations for all requests; and (3) Accuracy–Cost Trade-off score (ACT), the combined metric $\text{Acc} - \alpha \cdot \text{Cost}$ that balances effectiveness and efficiency, where higher Accuracy and ACT and lower Cost indicate better performance.

**Implementation Details.** Question embeddings use OpenAI's `text-embedding-3-small`; LLM responses use temperature 0 with one evaluation per request–configuration. The allocation policy is implemented with a ModernBERT-base encoder followed by a linear scoring head. Each input is truncated to 1,250 tokens. The model is optimized with AdamW using a learning rate of $1 \times 10^{-4}$, a 0.1 linear warmup ratio, and a batch size of 12. The complete training schedule uses 10 curriculum epochs followed by 10 additional full-data epochs. During the curriculum phase, the initial training fraction is 0.3 and increases linearly to include all requests. For respondent-history augmentation, Drop is applied with probability 0.25, removing at most two responses and at most 40% of the context, while Swap is applied with probability 0.15. The lightweight ModernBERT-base policy model (149M parameters) is deployed locally; we price it at $2 \times 149\text{M}/8\text{B} \approx 0.0373$ on the same relative scale and include this cost in the reported Cost for E$^2$Sim, alongside the LLM-based Selector's Qwen3-32B overhead (Appendix C). All reported E$^2$Sim results are averaged over three training seeds (41, 42, and 43). Further implementation details are available in our released code at `https://anonymous.4open.science/r/E2Sim-DE66`.

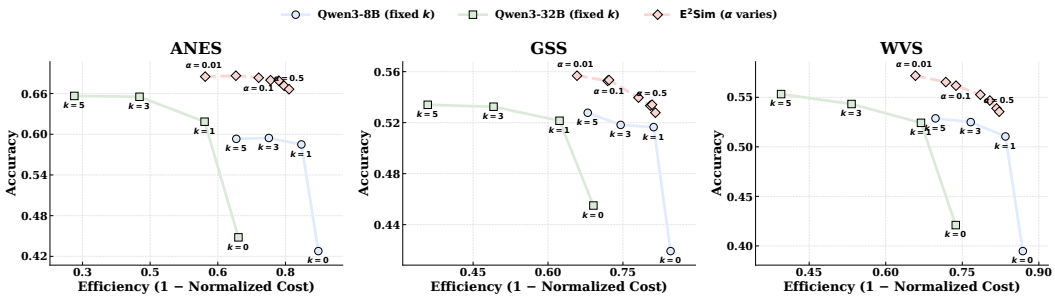



Figure 3: Cost-accuracy tradeoff under different alpha values. The plot shows fixed baselines (Qwen3-8B and Qwen3-32B with different history budgets $k$) and ours at various alpha values, demonstrating the trade-off between cost efficiency and accuracy.

### 4.2 Main Results

**Overall Performance.** Table 1 presents the performance of E$^2$Sim and four baselines across four social-survey datasets and two model configurations, using $\mathcal{K} = \{0, 1, 3, 5\}$ and $\alpha = 0.05$. E$^2$Sim outperforms both Random and LLM-based Selector (LLM-S) in terms of ACT across all datasets. More importantly, E$^2$Sim also improves upon Oracle Best Fixed (BF), the evaluation-set oracle upper bound over fixed configurations. Under Qwen3-8B + Qwen3-32B, E$^2$Sim improves Accuracy by 1.5–3.3 percentage points, reduces normalized Cost by 10.4–55.3%, and raises ACT by 2.8–4.3 percentage points across the four datasets. Under Qwen3-8B + DeepSeek-V4-Flash, E$^2$Sim improves Accuracy by 0.1–2.1 percentage points, reduces Cost by 25.6–37.1%, and raises ACT by 0.9–3.2 percentage points relative to BF on all four datasets. These results show that adaptive model–history allocation consistently improves the effectiveness–efficiency objective across all four datasets and both model configurations, with especially substantial savings in inference cost.

**Offline-Supervision Amortization.** We calculate the cost of collecting offline supervision by summing the price-weighted token costs of evaluating eight actions for each of 16,000 training requests. Because E$^2$Sim improves Accuracy while reducing Cost, we amortize this one-time expense using its per-request ACT gain over BF. Table 5 shows break-even points of 57K–78K requests across the four datasets, only 3.6–4.9 times the training volume. WVS reaches break-even after 62K requests, or 485 respondent-equivalents at the observed density of 127.4 requests per respondent; its original population of 97,220 respondents is approximately 200 times this scale. The other three datasets also exceed their break-even points at their original survey scales, by factors ranging from 2.9 to 21.9. Appendix D provides the exact accounting and conversion.

**Cost-Accuracy Tradeoff.** Figure 3 compares three performance curves: Qwen3-8B with different history budgets, Qwen3-32B with different history budgets, and E$^2$Sim under different values of $\alpha$. The first two curves represent fixed model–history configurations, whereas the third traces the configurations produced by E$^2$Sim as the cost preference changes. Across the evaluated values of $\alpha$, E$^2$Sim occupies the upper-right region relative to the two fixed-model curves, indicating a better Cost–Accuracy trade-off than either fixed approach. Moreover, varying $\alpha$ provides a mechanism for controlling the effectiveness–efficiency trade-off: larger values place greater emphasis on reducing inference Cost, while smaller values prioritize higher Accuracy. Thus, E$^2$Sim can operate at different points along the trade-off curve instead of being restricted to a single $\alpha$ configuration.

### 4.3 Robustness Analysis

**History-Budget-Space Size.** We first test whether E$^2$Sim is robust to the set of available history budgets. Table 2 compares four nested action spaces, $\mathcal{K} = \{0\}$, $\{0, 1\}$, $\{0, 1, 3\}$, and $\{0, 1, 3, 5\}$, using Qwen3-8B and Qwen3-32B as candidate models with $\alpha = 0.05$. Across both datasets and all four spaces, E$^2$Sim consistently achieves higher Accuracy and ACT and lower normalized Cost than BF. In particular, the ACT improvement ranges from 1.7 to 4.3 percentage points. E$^2$Sim outperforms BF even in the no-history setting $\mathcal{K} = \{0\}$, where the policy adapts only the model choice; this advantage persists as additional history budgets are introduced. This consistency indicates that the effectiveness of adaptive allocation is not tied to a particular history-budget-space definition.

Table 2: Performance across four history-budget-space sizes. Improved reports the difference between E$^2$Sim and the strongest baseline by ACT.

| Dataset | Method | $\mathcal{K}=\{0\}$ | | | $\mathcal{K}=\{0,1\}$ | | | $\mathcal{K}=\{0,1,3\}$ | | | $\mathcal{K}=\{0,1,3,5\}$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **Accuracy** | **Cost** | **ACT** | **Accuracy** | **Cost** | **ACT** | **Accuracy** | **Cost** | **ACT** | **Accuracy** | **Cost** | **ACT** |
| ANES | WCR | 0.3068 | 0.7106 | 0.2712 | 0.2425 | 0.6471 | 0.2101 | 0.2150 | 0.5555 | 0.1872 | 0.2047 | 0.5559 | 0.1770 |
| | Random | 0.4355 | 0.5762 | 0.4067 | 0.5245 | 0.5016 | 0.4994 | 0.5567 | 0.4083 | 0.5363 | 0.5737 | 0.3917 | 0.5542 |
| | LLM-S | 0.4485 | 2.3777 | 0.3296 | 0.6040 | 2.1173 | 0.4981 | 0.6150 | 1.6820 | 0.5309 | 0.6360 | 1.5882 | 0.5566 |
| | BF | 0.4480 | 0.7485 | 0.4106 | 0.6185 | 0.7263 | 0.5822 | 0.6552 | 0.6756 | 0.6215 | 0.6552 | 0.5736 | 0.6266 |
| | Ours | 0.4652 | 0.4886 | 0.4408 | 0.6412 | 0.4503 | 0.6187 | 0.6753 | 0.3775 | 0.6564 | 0.6880 | 0.3614 | 0.6699 |
| | Improved | **0.0172** | **0.2599** | **0.0302** | **0.0227** | **0.2760** | **0.0365** | **0.0201** | **0.2981** | **0.0349** | **0.0328** | **0.2122** | **0.0433** |
| BSA | WCR | 0.2165 | 0.6420 | 0.1844 | 0.1375 | 0.6094 | 0.1070 | 0.1175 | 0.5897 | 0.0880 | 0.1092 | 0.5871 | 0.0799 |
| | Random | 0.3683 | 0.5288 | 0.3418 | 0.4010 | 0.4644 | 0.3778 | 0.4145 | 0.4083 | 0.3941 | 0.4275 | 0.3848 | 0.4083 |
| | LLM-S | 0.3593 | 2.2453 | 0.2470 | 0.4377 | 1.9700 | 0.3393 | 0.4490 | 1.6337 | 0.3673 | 0.4675 | 1.5262 | 0.3912 |
| | BF | 0.3867 | 0.6660 | 0.3535 | 0.4597 | 0.6767 | 0.4259 | 0.4825 | 0.7046 | 0.4473 | 0.4825 | 0.5680 | 0.4541 |
| | Ours | 0.3960 | 0.5075 | 0.3706 | 0.4855 | 0.4693 | 0.4620 | 0.5100 | 0.3929 | 0.4904 | 0.5102 | 0.3870 | 0.4909 |
| | Improved | **0.0093** | **0.1585** | **0.0171** | **0.0258** | **0.2074** | **0.0361** | **0.0275** | **0.3117** | **0.0431** | **0.0277** | **0.1810** | **0.0368** |

Table 3: Performance across two model-pool sizes. Improved reports the difference between E$^2$Sim and the strongest baseline by ACT.

| Dataset | Metric | One Model | | | | | | Three Models | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **WCR** | **Random** | **LLM-S** | **BF** | **Ours** | **Improved** | **WCR** | **Random** | **LLM-S** | **BF** | **Ours** | **Improved** |
| ANES | Accuracy | 0.2752 | 0.5510 | 0.5897 | 0.5945 | 0.6575 | **0.0630** | 0.1810 | 0.5917 | 0.6522 | 0.6850 | 0.7053 | **0.0203** |
| | Cost | 0.6213 | 0.5206 | 2.5468 | 0.5736 | 0.5071 | **0.0665** | 0.5727 | 0.4339 | 1.8424 | 0.5622 | 0.3434 | **0.2188** |
| | ACT | 0.2442 | 0.5250 | 0.4624 | 0.5658 | 0.6321 | **0.0663** | 0.1524 | 0.5701 | 0.5601 | 0.6569 | 0.6881 | **0.0312** |
| BSA | Accuracy | 0.1792 | 0.4095 | 0.4507 | 0.4435 | 0.4730 | **0.0295** | 0.0902 | 0.4585 | 0.4760 | 0.5155 | 0.5262 | **0.0107** |
| | Cost | 0.6394 | 0.5106 | 2.4984 | 0.5680 | 0.4633 | **0.1047** | 0.6095 | 0.4223 | 1.7778 | 0.5486 | 0.3768 | **0.1718** |
| | ACT | 0.1473 | 0.3840 | 0.3258 | 0.4151 | 0.4498 | **0.0347** | 0.0598 | 0.4374 | 0.3871 | 0.4881 | 0.5074 | **0.0193** |

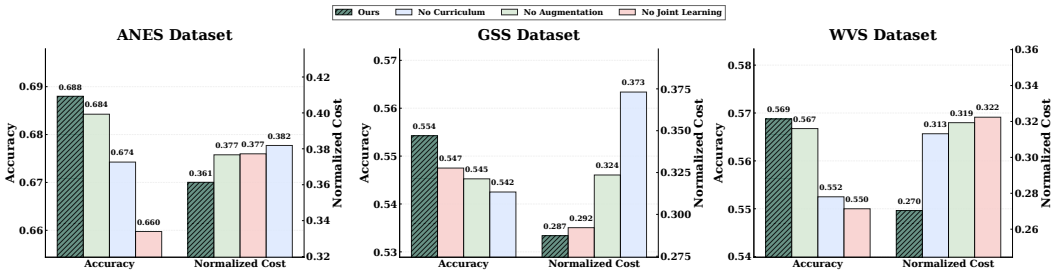



Figure 4: Ablation study of E$^2$Sim on ANES, GSS, and WVS. We compare the full framework with variants removing the curriculum, respondent-history augmentation, or joint model–history selection. The left and right axes show Accuracy and Normalized Cost, respectively.

**Model-Pool Size.** We next vary the number of candidate models while fixing $\mathcal{K} = \{0, 1, 3, 5\}$ and $\alpha = 0.05$. Table 3 considers a single-model pool containing Qwen3-8B and a heterogeneous three-model pool containing Qwen3-8B, Qwen3-32B, and DeepSeek-V4-Flash. With one model, E$^2$Sim can adapt only the history budget. Relative to BF, it improves Accuracy by 6.3 and 3.0 percentage points, reduces Cost by 11.6% and 18.4%, and raises ACT by 6.6 and 3.5 percentage points on ANES and BSA, respectively. With three models, where both the model and history budget can be selected, E$^2$Sim improves Accuracy by 2.0 and 1.1 percentage points, reduces Cost by 38.9% and 31.3%, and raises ACT by 3.1 and 1.9 percentage points on ANES and BSA, respectively. Thus, E$^2$Sim retains its effectiveness–efficiency advantage under both restricted and expanded model pools.

### 4.4 Ablation Study

We conduct ablation studies to examine the contribution of the ACT-margin curriculum, respondent-history augmentation, and joint model–history selection in E$^2$Sim. All experiments use Qwen3-8B and Qwen3-32B, the history-budget space $\mathcal{K} = \{0, 1, 3, 5\}$, and $\alpha = 0.05$. Figure 4 reports the results on ANES, GSS, and WVS. We evaluate the following variants:

- No Curriculum: removes the ACT-margin curriculum and trains on all requests immediately.

- No Augmentation: disables historical-response deletion and permutation during training.
- No Joint Learning: replaces joint model–history selection with a two-stage procedure that first selects the model and then selects the history budget (Appendix E).

As shown in Figure 4, the full E$^2$Sim consistently achieves higher Accuracy and lower Cost than all three ablated variants across the three datasets. Removing the curriculum causes the largest degradation on GSS, while replacing joint selection with the two-stage procedure has the greatest impact on ANES and WVS. Removing respondent-history augmentation also leads to lower Accuracy and higher Cost on every dataset. These results demonstrate that all three components contribute to the effectiveness–efficiency trade-off of E$^2$Sim.

## 5 Related Work

We review two lines of work most relevant to our study: LLM-based survey simulation and LLM Routing and Adaptive Inference.

**LLM-based Survey Simulation.** Large language models have increasingly been used as synthetic respondents in survey research. Early studies showed that LLMs conditioned on sociodemographic profiles or persona descriptions can reproduce aspects of human attitudes and response distributions, motivating their use for scalable survey simulation (Argyle et al., 2023; Aher et al., 2023; Durmus et al., 2023; Cao et al., 2023). Subsequent work further examined whose opinions LLMs reflect and the limitations of using synthetic responses as substitutes for human survey data (Santurkar et al., 2023; Bisbee et al., 2024). More recently, research has moved toward specialized models, richer respondent conditioning, and more systematic evaluation (Chen et al., 2026). These include fine-tuning models to reproduce population-level response distributions (Cao et al., 2025; Suh et al., 2025; Huang et al., 2026), as well as respondent-level simulation using sociodemographic personas, personality information, and historical survey responses (Wang et al., 2025a; Lutz et al., 2025; Liu et al., 2025; Ku et al., 2026). Recent studies also show that survey simulation is sensitive to persona construction, attribute selection, response-generation strategies, and prompt design, and that providing more respondent information does not necessarily improve fidelity (Lutz et al., 2025; Morocho et al., 2026; Ahnert et al., 2026; Fröhling et al., 2026). However, these methods generally treat the model and respondent context as study-level configuration choices. In contrast, E$^2$Sim treats them as request-dependent resources and jointly allocates model capability and historical-response context for each respondent–question pair under an explicit accuracy–cost objective.

**LLM Routing and Adaptive Inference.** LLM routing improves inference efficiency by dynamically assigning each query to a model with an appropriate capability–cost trade-off. Existing approaches use signals such as preference feedback, query representations, contrastive supervision, or model–query relations to learn routing policies, while cascading methods progressively invoke stronger models when additional capability is needed (Ong et al., 2025; Mohammadshahi et al., 2024; Chen et al., 2024; Feng et al., 2024; Dekoninck et al., 2024). Recent work further studies budget-aware online routing and large-scale evaluation of routing policies (Panda et al., 2025; Huang et al., 2025). More broadly, our work considers adaptive inference as a resource-allocation problem rather than model selection alone. In survey simulation, different respondent–question requests may require different levels of model capability and different amounts of respondent history. E$^2$Sim therefore jointly allocates these two resources at the request level, selecting both the model and the history budget under an explicit accuracy–cost objective.

## 6 Conclusion

In this paper, we identified the limitations of fixed configurations in LLM-based social survey simulation by examining the request-dependent effects of model capability, respondent history, and their interaction. To address this problem, we proposed E$^2$Sim, an adaptive resource allocation framework that jointly selects the model and history budget for each request. The framework combines ACT-margin curriculum learning with respondent-history drop-and-swap augmentation. Across the evaluated settings, E$^2$Sim generally achieves higher observed ACT than the compared baselines while improving Accuracy and reducing inference Cost in most cases. Future work will extend the action space and adapt model-cost weights to deployment-specific pricing (Appendix F).

## AI USE STATEMENT

Language models were used as experimental components in this work: the candidate LLMs generated simulated survey responses, and Qwen3-32B served as the LLM-based configuration-selection baseline. Separately, generative AI tools provided limited assistance with manuscript language editing, translation, formatting, and software troubleshooting. They were not used to originate the research idea, formulate the method, design the experiments, alter the datasets or reported results, or interpret the findings. All AI-assisted edits and suggestions were reviewed by the authors, who take responsibility for the final manuscript and all of its claims, code, and other artifacts.

## ETHICS STATEMENT

This work uses publicly released social-survey datasets and publicly available language models, subject to their respective terms of use and licenses. We do not collect new data from human participants. The simulated responses are research artifacts and should not be treated as substitutes for informed responses from real individuals or as authoritative statements about any demographic group. Because language-model outputs can reproduce biases or errors present in their training data and prompts, the reported simulations should be interpreted with appropriate caution.

## REPRODUCIBILITY STATEMENT

We provide the formal problem definition and training procedure in Sections 2 and 3, the experimental settings and evaluation protocol in Section 4, and the evaluation prompt templates in Appendix B. The source code, experiment launchers, figures, and table-generation scripts are released at `https://anonymous.4open.science/r/E2Sim-DE66`. Exact reproduction additionally requires access to the survey data specified in the code, model endpoints or checkpoints, the embedding model, and the documented model-price settings.

## REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Gati V Aher, Rosa I Arriaga, and Adam Tauman Kalai. Using large language models to simulate multiple humans and replicate human subject studies. In *International conference on machine learning*, pages 337–371. PMLR, 2023.

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 (Volume 1: Long Papers)*, pages 41554–41577, 2026.

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.

James Bisbee, Joshua D Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer M Larson. Synthetic replacements for human survey data? the perils of large language models. *Political Analysis*, 32 (4):401–416, 2024.

Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello Piqueras, Min Chen, and Daniel Hershcovich. Assessing cross-cultural alignment between chatgpt and human societies: An empirical study. In *Proceedings of the first workshop on cross-cultural considerations in NLP (C3NLP)*, pages 53–67, 2023.

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 the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3141–3154, 2025.

Shuhao Chen, Weisen Jiang, Baijiong Lin, James Kwok, and Yu Zhang. Routerdc: Query-based router by dual contrastive learning for assembling large language models. *Advances in Neural Information Processing Systems*, 37:66305–66328, 2024.

Xu Chen, Yuanzi Li, Lei Wang, Nan Lu, Yang Wang, Anding Wang, Lei Shi, Xiaoxing Fu, and Ji-Rong Wen. Benchmarking llms for community governance simulation with life-history narratives. *arXiv preprint arXiv:2605.23783*, 2026.

Jasper Dekoninck, Maximilian Baader, and Martin Vechev. A unified approach to routing and cascading for llms. *arXiv preprint arXiv:2410.10347*, 2024.

Esin Durmus, Karina Nguyen, Thomas I Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, et al. Towards measuring the representation of subjective global opinions in language models. *arXiv preprint arXiv:2306.16388*, 2023.

Tao Feng, Yanzhen Shen, and Jiaxuan You. Graphrouter: A graph-based router for llm selections. *arXiv preprint arXiv:2410.03834*, 2024.

Leon Fröhling, Jens Rupprecht, Markus Strohmaier, and Claudia Wagner. When persona attributes improve population alignment in large language models. *arXiv preprint arXiv:2609.02526*, 2026.

Robert M Groves, Floyd J Fowler Jr, Mick P Couper, James M Lepkowski, Eleanor Singer, and Roger Tourangeau. *Survey methodology*. John Wiley & Sons, 2011.

Lynne Hamill and Geoffrey Gilbert. Social circles: A simple structure for agent-based social network models. *Journal of Artificial Societies and Social Simulation*, 12(2), 2009.

Ori Heffetz and Daniel B Reeves. Difficulty of reaching respondents and nonresponse bias: Evidence from large government surveys. *Review of Economics and Statistics*, 101(1):176–191, 2019.

Ji Huang, LI Mengfei, and Shuai Shao. Distribution shift alignment helps llms simulate survey response distributions. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 9395–9409, 2026.

Zhongzhan Huang, Guoming Ling, Yupei Lin, Yandong Chen, Shanshan Zhong, Hefeng Wu, and Liang Lin. Routereval: A comprehensive benchmark for routing llms to explore model-level scaling up in llms. *arXiv preprint arXiv:2503.10657*, 2025.

Graham Kalton. Methods for oversampling rare subpopulations in social surveys. *Survey methodology*, 35(2):125–141, 2009.

Chan-Tung Ku, Chan Hsu, Pei-Cing Huang, Frank Cheng-shan Liu, I Cheng, Yihuang Kang, et al. Silicon sampling via cross-survey transfer. *arXiv preprint arXiv:2607.03091*, 2026.

Haijiang Liu, Qiyuan Li, Chao Gao, Yong Cao, Xiangyu Xu, Xun Wu, Daniel Hershcovich, and Jinguang Gu. Beyond demographics: Enhancing cultural value survey simulation with multi-stage personality-driven cognitive reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 18417–18439, 2025.

Marlene Lutz, Indira Sen, Georg Ahnert, Elisa Rogers, and Markus Strohmaier. The prompt makes the person (a): A systematic evaluation of sociodemographic persona prompting for large language models. In *EMNLP (Findings)*, pages 23212–23237, 2025.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. Large language models: A survey. *arXiv preprint arXiv:2402.06196*, 2024.

Alireza Mohammadshahi, Arshad Rafiq Shaikh, and Majid Yazdani. Routoo: Learning to route to large language models effectively. *arXiv preprint arXiv:2401.13979*, 2024.

Erika Elizabeth Taday Morocho, Lorenzo Cima, Tiziano Fagni, Marco Avvenuti, and Stefano Cresci. Assessing the reliability of persona-conditioned llms as synthetic survey respondents. *arXiv preprint arXiv:2602.18462*, 2026.

Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E. Gonzalez, M Waleed Kadous, and Ion Stoica. Routellm: Learning to route llms with preference data, 2025. https://arxiv.org/abs/2406.18665.

Pranoy Panda, Raghav Magazine, Chaitanya Devaguptapu, Sho Takemori, and Vishal Sharma. Adaptive llm routing under budget constraints. In *EMNLP (Findings)*, pages 23934–23949, 2025.

Siddegowda Roopa and Menta Satya Rani. Questionnaire designing for a survey. *Journal of Indian Orthodontic Society*, 46(4_suppl1):273–277, 2012.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? In *International conference on machine learning*, pages 29971–30004. PMLR, 2023.

Flaminio Squazzoni, J Gareth Polhill, Bruce Edmonds, Petra Ahrweiler, Patrycja Antosz, Geeske Scholz, Émile Chappin, Melania Borit, Harko Verhagen, Francesca Giardini, et al. Computational models that matter during a global pandemic outbreak: A call to action. *Jasss*, 23(2), 2020.

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 (Volume 1: Long Papers)*, pages 21147–21170, 2025.

R Tourangeau, LJ Rips, and K Rasinski. The psychology of survey response. cambridge university press; cambridge core, 2000.

Jia Wang, Ziyu Zhao, Tingjuntao Ni, and Zhongyu Wei. Sociobench: Modeling human behavior in sociological surveys with large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 26268–26300, 2025a.

Lei Wang, Jingsen Zhang, Hao Yang, Zhi-Yuan Chen, Jiakai Tang, Zeyu Zhang, Xu Chen, Yankai Lin, Hao Sun, Ruihua Song, et al. User behavior simulation with large language model-based agents. *ACM Transactions on Information Systems*, 43(2):1–37, 2025b.

James D Wright, Peter V Marsden, et al. Survey research and social science: History, current practice, and future prospects. *Handbook of survey research*, pages 3–26, 2010.

# A Dataset and Evaluation Protocols

## A.1 Dataset Statistics

For each survey, we organize the observed responses as a respondent–question matrix and use 20,000 observed entries as the experimental pool. Each entry is a respondent–question pair together with its recorded response. Table 4 reports the numbers of respondents and questions in each pool and the training and test sizes. We randomly assign 20% of observed pairs to test using split seed 0. For both training and test requests, history is retrieved only from training responses, excluding the target pair; test answers are used solely for evaluation.

Table 4: Dataset statistics and training/test split sizes.

| Dataset | Pairs | Respondents | Questions | Train | Test |
|---|---|---|---|---|---|
| ANES | 20,000 | 88 | 332 | 16,000 | 4,000 |
| BSA | 20,000 | 430 | 119 | 16,000 | 4,000 |
| GSS | 20,000 | 265 | 208 | 16,000 | 4,000 |
| WVS | 20,000 | 157 | 175 | 16,000 | 4,000 |

# B Prompt Templates

This section records the prompt templates used for survey-response generation and LLM-based configuration selection.

## B.1 Survey-Response Generation Prompt

The following template is used to generate $\hat{y}_i(a)$ for every model–history configuration (Eq. 6). The placeholders in angle brackets are filled with the respondent persona, retrieved historical responses, and target question for each request.

**Complete system and user prompt template**

```
SYSTEM PROMPT:
You are simulating the survey respondent described by the persona.
Answer the target survey question as that respondent would answer
it. Use past answers only as behavioral evidence. Select exactly
one listed option and return only a JSON object in the form {"
answer": <integer>}.

USER PROMPT:
PERSONA:
<respondent persona>

PAST ANSWERS (most similar first):
Example 1 (semantic similarity=<score>):
<historical question text>
Options:
[0] <option 0>
[1] <option 1>
...
Respondent's answer: [<answer index>] <answer text>

Example 2 (semantic similarity=<score>):
<historical question text>
Options:
[0] <option 0>
[1] <option 1>
...
Respondent's answer: [<answer index>] <answer text>
```

```
<additional examples, if any>

TARGET QUESTION:
<target question text>
Options:
[0] <option 0>
[1] <option 1>
...
Return only: {"answer": <integer>}
```

### B.2 LLM-BASED SELECTOR PROMPT

This section records the exact prompt used by the LLM-based Selector (LLM-S) baseline described in Section 4. Given the respondent persona, target question, and the candidate model–history configurations with their normalized costs, the selector returns the index of one action $a \in \mathcal{A}$.

**Complete LLM-based Selector system and user prompt**

```
SYSTEM PROMPT:
You are an expert system that selects the optimal language model
configuration for survey response prediction.

Given:
1. A respondent persona
2. A survey question to answer
3. Available model configurations with their normalized costs

Your task is to maximize the REWARD function:
  Reward = Accuracy - 0.05 x Cost

Where:
- Accuracy: prediction correctness (0 or 1)
- Cost: normalized computational cost of the selected configuration
- Alpha (0.05): cost penalty weight

This means:
- A one-percentage-point accuracy improvement offsets an increase
of 0.2 in cost (0.01 / 0.05 = 0.2)
- Accuracy is the dominant factor, but cost matters
- Avoid unnecessarily expensive configurations when simpler ones
suffice
- Don't sacrifice accuracy just to save cost

Return ONLY a JSON object in the format: {"selected_action": <
action_index>}

USER PROMPT:
RESPONDENT PERSONA:
<respondent persona>

TARGET SURVEY QUESTION:
<target question with its options>

AVAILABLE RESPONDENT HISTORY (retrieved similar questions):
  Example 1:
  <example question with its options, if available>
  Respondent's answer: [<answer index>] <answer text>

  Example 2:
  <example question with its options, if available>
  Respondent's answer: [<answer index>] <answer text>
```

```
  <additional questions and respondent answers, if available>

AVAILABLE CONFIGURATIONS:
Action 0: Model=<model>, history budget=0 (no history), Normalized
Cost=<normalized_cost>
Action 1: Model=<model>, history budget=<k> (use first <k>
responses above), Normalized Cost=<normalized_cost>
<one line for every candidate action>

OPTIMIZATION OBJECTIVE:
Maximize: Reward = Accuracy - 0.05 x Cost

DECISION STRATEGY:
1. Estimate the expected accuracy for each configuration based on:
   - Question complexity (factual vs. opinion-based, simple vs.
nuanced)
   - Persona characteristics (education, political engagement,
demographic factors)
   - Model capability (larger models handle complex reasoning
better)
   - history relevance (do the responses help with this specific
question?)

2. Evaluate the cost-benefit tradeoff:
   - Use the normalized Cost values listed above
   - Prefer a higher-Cost configuration only when its expected
Accuracy gain exceeds 0.05 times the Cost increase

3. Apply heuristics:
   - Simple factual questions -> cheaper configs often sufficient
   - Complex political/value questions -> larger models often worth
 the cost
   - Highly educated/engaged personas -> may need more nuanced
predictions
   - Generic questions with clear answers -> respondent history may
 not help much

Select the action index that maximizes the expected reward.
```

## C Cost Computation

For a selected survey action $a = (m, k)$ on request $i$, let $T_{i,a}$ denote the API-reported prompt-token count and let $c_m^{\text{in}}$ denote the relative input-token price of model $m$. The execution cost is

$$C_i^{\text{exec}} = c_m^{\text{in}} T_{i,a}. \tag{16}$$

Both LLM-based Selector and E$^2$Sim incur an additional configuration-selection cost before this execution step, since each request must first be routed to an action $a$. For LLM-based Selector, let $T_i^{\text{selector}}$ be the selector prompt-token count and $c_{\text{selector}}^{\text{in}}$ its price multiplier (the Qwen3-32B input-token price). The selector cost is

$$C_i^{\text{selector}} = c_{\text{selector}}^{\text{in}} T_i^{\text{selector}}. \tag{17}$$

For E$^2$Sim, let $T_i^{\text{policy}}$ be the input-token count of the policy input $z(x)$ defined in Section 3.3 and fed to the ModernBERT-base policy model, priced at $c_{\text{policy}}^{\text{in}} = 2 \times 149\text{M}/8\text{B} \approx 0.0373$ on the same relative scale as Qwen3-8B, by parameter count. The policy-model cost is

$$C_i^{\text{policy}} = c_{\text{policy}}^{\text{in}} T_i^{\text{policy}}. \tag{18}$$

Let $C_i^{\text{select}}$ denote the applicable selection cost for a given method: $C_i^{\text{select}} = C_i^{\text{selector}}$ for LLM-based Selector, $C_i^{\text{select}} = C_i^{\text{policy}}$ for E$^2$Sim, and $C_i^{\text{select}} = 0$ for methods without a selection step

Table 5: ACT-based amortization of offline supervision. BE denotes break-even, Scale is BE requests relative to training, and Full/BE is the original survey population relative to BE respondents.

| Dataset | Train Resp. | Full Resp. | Req./Resp. | BE Req. | BE Resp. | Scale | Full/BE |
|---|---|---|---|---|---|---|---|
| ANES | 88 | 5,521 | 227.3 | 57K | 252 | 3.6× | 21.9× |
| BSA | 430 | 4,120 | 46.5 | 66K | 1,417 | 4.1× | 2.9× |
| GSS | 265 | 3,986 | 75.5 | 78K | 1,036 | 4.9× | 3.8× |
| WVS | 157 | 97,220 | 127.4 | 62K | 485 | 3.9× | 200.4× |

(WCR, Random, Oracle Best Fixed). The average total cost of a method is

$$\overline{C}^{\text{total}} = \frac{1}{N}\sum_{i=1}^{N}\left(C_i^{\text{exec}} + C_i^{\text{select}}\right). \tag{19}$$

To ensure comparability, the normalizer is defined once for each dataset, model pool, history-budget action space, and evaluated request set, and the same value is used for every method compared under that setting. It is the maximum candidate execution cost over all evaluated requests and candidate actions:

$$C_{\max} = \max_{i,\, a=(m,k)\in\mathcal{A}} c_m^{\text{in}} T_{i,a}, \qquad \text{Cost} = \frac{\overline{C}^{\text{total}}}{C_{\max}}. \tag{20}$$

The reported LLM-based Selector results include the Qwen3-32B selection cost, and the reported E$^2$Sim results include the ModernBERT-base policy-model cost, in the numerator, so the normalized Cost can exceed one for either method. The reported trade-off score is $\text{Acc} - \alpha\,\text{Cost}$.

## D Offline-Supervision Amortization Details

We measure how many deployment requests are needed to recover the one-time cost of constructing E$^2$Sim's offline supervision. For the primary model pool, each of the $N_{\text{train}} = 16{,}000$ training requests is evaluated under eight actions. Using the recorded prompt-token counts and the same model-price weights as the main experiments, the exact acquisition cost is

$$C_{\text{offline}} = \sum_{i\in\mathcal{D}_{\text{train}}}\sum_{a=(m,k)\in\mathcal{A}} c_m^{\text{in}} T(x_i, a). \tag{21}$$

Let $C_{\max}$ be the shared normalizer from Section C. Under the paper's objective, the offline acquisition contributes the one-time penalty $\alpha C_{\text{offline}}/C_{\max}$. The per-request deployment gain relative to BF is

$$\Delta\overline{g} = \left(\overline{\text{Acc}}_{\text{E2Sim}} - \alpha\overline{C}_{\text{E2Sim}}\right) - \left(\overline{\text{Acc}}_{\text{BF}} - \alpha\overline{C}_{\text{BF}}\right). \tag{22}$$

The ACT-based break-even request volume is therefore

$$N_{\text{BE}} = \frac{\alpha C_{\text{offline}}/C_{\max}}{\Delta\overline{g}}. \tag{23}$$

This formulation credits both the Accuracy improvement and the Cost reduction under the same objective used to train and evaluate E$^2$Sim. The comparison remains conservative toward the fixed baseline because BF is assigned no offline selection cost. The Two-Stage Router is not used as the amortization reference because it consumes the same exhaustive action matrix as E$^2$Sim, so their acquisition costs cancel in a direct comparison. The released analysis file additionally reports a cost-only break-even value as a stricter sensitivity measure.

To express request volume as respondent-equivalents, we use the observed density of the experimental pool, $\bar{Q}_d = 20{,}000/R_d$, where $R_d$ is the number of respondents reported in Table 4. Thus, $R_{\text{BE}} = N_{\text{BE}}/\bar{Q}_d$. The original source surveys contain 5,521, 4,120, 3,986, and 97,220 respondents for ANES, BSA, GSS, and WVS, respectively. "Full/BE" in Table 5 divides these population sizes by $R_{\text{BE}}$; it is a workload projection that holds the observed requests-per-respondent density fixed rather than assuming that every respondent is simulated on every questionnaire item. Exact intermediate values are produced from the released token logs by the amortization-analysis script.

## E TWO-STAGE ABLATION DETAILS

The No Joint Learning variant, illustrated in Figure 5, uses the same ModernBERT-base architecture and Smooth-$L_1$ ACT regression objective as the joint policy, but decomposes allocation into two stages. Stage 1 receives the persona and target question without respondent history and regresses the ACT values of the $k = 0$ action for each candidate model; the model with the highest predicted value is selected. Stage 2 trains one independent history-budget selector per model, where each selector receives the full top-$k_{\max}$ history and regresses the ACT values of that model's actions over $k \in \mathcal{K}$. Both stages use the same ACT-margin curriculum, while respondent-history drop-and-swap augmentation is applied only to the Stage 2 selectors. At inference, Stage 1 selects $m^*$ and the corresponding model-specific Stage 2 selector chooses $k^*$, producing the final action $(m^*, k^*)$.

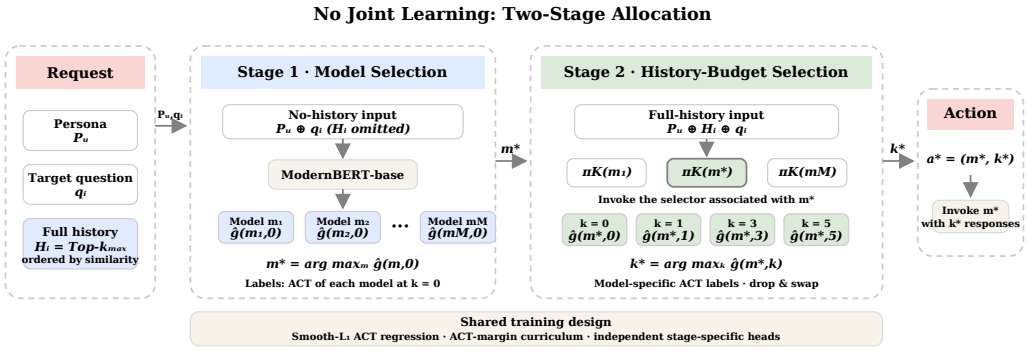



Figure 5: Two-stage allocation in the No Joint Learning ablation. Stage 1 selects a model using the predicted ACT of each model's no-history action. The selector associated with the chosen model then predicts the ACT of each available history budget and selects $k^*$, yielding the final action $(m^*, k^*)$.

## F LIMITATIONS

1. **Predefined action space.** E$^2$Sim operates over a fixed set of candidate models and history budgets. Introducing a new model or history budget requires evaluating the added configurations and updating the allocation policy.
2. **Empirical model costs.** Our experiments use empirical relative model-cost weights to provide controlled comparisons. These weights are not universal estimates of wall-clock latency or monetary cost; in deployment, they should be configured using current API prices or costs measured in the target hardware environment.