Proceedings of the 4th International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/2026.39400

Calibrating Large Language Model Agents for Cross-Experiment Prediction Through Preference-Weighted Sampling

Zhiqiao Qi

School of Digital Media and Design Arts, Beijing University of Posts and Telecommunications, Beijing, China
qizhiqiao77@gmail.com

Abstract. Using DeepSeek-v4-flash, this study replicates social-preference calibration in the single-stage allocation games examined by Horton and colleagues. The paper estimates sampling weights for three social-preference agent types by constrained least squares and randomly draws agents one by one according to those weights. The paper additionally estimates weights by binomial maximum likelihood using the original human choice counts, and compares the cross-experiment predictions of human choices in two-stage games produced by agents sampled under the two weighting methods. Relative to an unprompted sample, the constrained least-squares calibrated sample reduces prediction mean squared error (MSE) by 51.1% and lowers mean absolute error (MAE) to 0.142833. The binomial-likelihood sample reduces MSE by 51.2%, but its MAE rises to 0.160000, showing no consistent cross-experiment improvement. The error in Berk18 suggests that fixed-preference agents may fail to distinguish game outcomes with equal total payoffs. The results indicate that theory-based social-preference-weighted sampling improves cross-experiment prediction, whereas changing the weight-estimation method alone may be insufficient to resolve errors in agent-behavior prediction.

Keywords: large language models, behavioral simulation, preference-weighted sampling, cross-experiment prediction

1. Introduction

Large language models can generate discrete economic choices when given payoffs, information, and role conditions, and can therefore be viewed as programmable simulators of behavior [1, 2]. Horton, Filippas, and Manning call this implicit model of human behavior Homo silicus and examine the correspondence between AI agents and human behavior in experiments on fairness judgments, social preferences, status quo bias, choices involving complexity, and minimum wages [1]. Compared with conventional human-subject experiments, model calls offer lower costs, faster execution, and greater control over experimental conditions. Yet the ability to generate choices does not imply an ability to represent the target population accurately: unconditioned models often compress behavioral heterogeneity, and their behavior can vary with model version, prompt

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.39400

wording, and role specification [3-5]. Thus, calibrating AI agent samples with limited human data is crucial if these agents are to be used for prediction.

The method of Manning and Horton shows that optimizing the proportions of theory-based agents to match the distribution of human responses in one environment can improve their predictions in related environments [6]. "Theory-based" means that a persona is driven by a theory that it generally understands, can evaluate clearly, and can reasonably expect to predict relevant human behavior—for example, how fairness, self-interest, and efficiency agents reflect the original economic model proposed by Charness and Rabin [7]. Argyle and colleagues find that conditioning GPT-3 on information about human subgroups can reproduce some subgroup differences in survey-response distributions [8]. A specific empirical question, then, is whether a calibrated choice distribution transfers to experiments with a different structure.

The behavioral interpretation in this study is grounded in games involving social preferences. Charness and Rabin use simple allocation games to distinguish motives involving efficiency, inequality, and self-interest [7]. Fehr and Schmidt develop a model of inequity aversion that explains why individual choices may depart from the maximization of personal payoffs [9]. These motives can each be expressed as a prompt to the same underlying model and then combined with weights to simulate observed human choice distributions using relatively few parameters. Leng, Sang, and Agarwal investigate how to calibrate model-generated samples to human targets [10]. Bui and colleagues learn a mixture distribution from language models with different personas for population simulation [11]. These studies support the approach of simulating population distributions.

Agent responses may appear plausible in aggregate while concealing systematic errors, so their performance must be tested on tasks that were not used for calibration. Gao and colleagues identify limitations of treating large language models as human surrogates [3]. Qu and Wang document differences in performance across settings for public-opinion simulation [4]. Larooij and Törnberg place external validation at the center of generative social simulation [12]. The paper estimates sampling weights using constrained least squares and binomial likelihood, respectively, and then compares the cross-experiment predictions of human choices in two-stage games made by the two resulting agent samples.

2. Methods

2.1. Single-stage experiments

Following Horton and colleagues, the paper uses six single-stage games based on the allocation games of Charness and Rabin [1, 7]. The paper provides the LLM with each game description and asks it to choose its preferred allocation without any additional information. In each call, DeepSeek-v4-flash acts as the decision maker, Person B, and chooses between the monetary allocations Left and Right. The numbers in each payoff pair denote the payoffs to Person A (the other person) and Person B (the decision maker), in that order. For example, (400, 600) means that the other person receives 400 and the decision maker receives 600. Figure 1 shows the allocations in all six games.

The game experiment has four prompting conditions. The no-persona condition contains no social-preference instruction. Consistent with the original experiment, the Efficient, Inequity-averse, and Self-interested conditions add, respectively, "you only care about the total payoff of both players", "you only care about fairness between players", and "you only care about your own payoff".

The paper uses DeepSeek-v4-flash with temperature = 1. The model is called 100 times for each combination of six games and four prompting conditions, yielding 2,400 choices. The paper

DOI: 10.54254/2755-2721/2026.39400

calculates the proportion choosing Left from the number of Left choices. For each of the three preference-agent types, these proportions in the six games are used to estimate sampling weights.

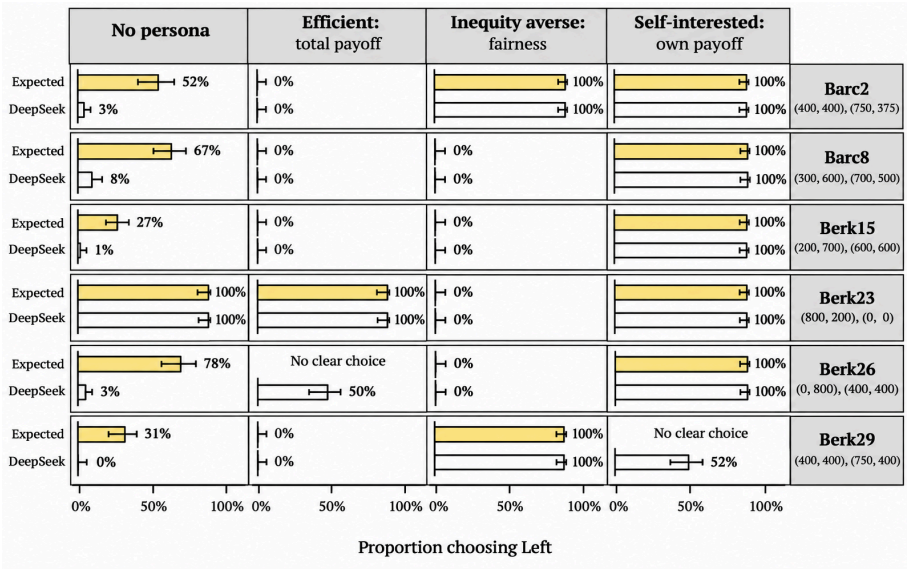


Figure 1. DeepSeek responses in single-stage dictator games (photo/picture credit: original)

Note: The horizontal axis shows the proportion choosing Left. Yellow indicates observed human choices or theoretical predictions from Charness and Rabin [7]; white indicates DeepSeek choices.

Figure 1 summarizes the single-stage results. The horizontal axis shows the proportion choosing Left; white bars show DeepSeek choices, and yellow bars show the observed human choices or theoretical predictions reported by Charness and Rabin. In the no-persona condition, DeepSeek differs markedly from human choice proportions in five of the six games, with Berk23 as the exception. For example, approximately 52%, 67%, and 78% of human participants chose Left in Barc2, Barc8, and Berk26, respectively, whereas the corresponding model proportions were only 3%, 8%, and 3%. Unprompted agent responses therefore cannot directly represent this human choice distribution.

With social-preference prompts, the three agent types exhibit different choice patterns. In games where the theory identifies an unambiguous choice, the observed proportions for the efficient, inequity-averse, and self-interested agents match the expected values. The two options in Berk26 yield equal total payoffs, and 50% of efficient agents choose Left. The two options in Berk29 give the decision maker the same payoff, and 52% of self-interested agents choose Left. In these two cases, the corresponding preference criterion does not identify a uniquely optimal option.

2.2. Replication of constrained least-squares preference-weight estimation

To determine the sampling probabilities of the three preference-agent types, the paper estimates mixture weights from choices in the six single-stage games. Let pgk denote the proportion of type-k agents choosing Left in game g, and let yg denote the corresponding proportion of human participants choosing Left. Here, $k \in \{E, I, S\}$ indexes efficient, inequity-averse, and self-interested agents, respectively. Given weights $w = (w_E, w_I, w_S)$, the predicted proportion choosing Left for the mixture population in game g is

$$q_g(\mathbf{w}) = w_E p_{gE} + w_I p_{gI} + w_S p_{gS} \tag{1}$$

DOI: 10.54254/2755-2721/2026.39400

Following the calibration method of Horton and colleagues, it seeks the weights that minimize the mean squared error between the predicted and human choice proportions across the six games [1]:

$$\mathbf{w}^{*}_{\mathrm{MSE}} = \underset{\mathbf{w}}{\operatorname{argmin}} \ \tfrac{1}{6} \sum_{g=1}^{6} \left[q_g(\mathbf{w}) - y_g\right]^2, \quad w_E + w_I + w_S = 1, \quad w_E, w_I, w_S \geq 0. \tag{2}$$

The resulting weights are $(w_E, w_I, w_S) = (0.487236, 0.019420, 0.493344)$.

2.3. Binomial-likelihood preference-weight estimation

To use the original human choice counts from the six single-stage games, the paper introduces a method that estimates the sampling weights by maximizing binomial likelihood. As before, pgk denotes the proportion of type-k agents choosing Left in game g. If Ng human participants play game g and Lg of them choose Left, the observed count is modeled as

$$L_g \sim \mathrm{Binomial}(N_g, q_g(\mathbf{w})) \tag{3}$$

Under the constraints that the weights are nonnegative and sum to one, it maximizes the joint log-likelihood across the six games:

$$w^{*}_{MLE} = \sum_{g=1}^{6} [L_g log q_g(w) + (N_g - L_g) log(1 - q_g(w))], \tag{4}$$

$$w_E + w_I + w_S = 1, \qquad w_E, w_I, w_S \geq 0$$

The binomial coefficients, which do not depend on the weights, are omitted from this expression. For example, 25 of the 48 human participants in Barc2 chose Left. The contribution of this game to the objective is $25 \log q_{\mathrm{Barc2}} + 23 \log(1 - q_{\mathrm{Barc2}})$. The six games jointly determine the optimal weights, yielding $(w_E, w_I, w_S) = (0.472210, 0, 0.527790)$. It then randomly samples agents one by one according to these weights and evaluates their predictions in the two-stage games.

2.4. Sampling agents by preference weight

After obtaining the least-squares weights, it assigns sampling probabilities of 0.487236, 0.019420, and 0.493344 to efficient, inequity-averse, and self-interested agents, respectively. With a fixed random seed, it draws 100 times from this preference-weight distribution and saves the resulting list of preference types as the agent sample for the two-stage experiment. This draw yielded 52 efficient, 3 inequity-averse, and 45 self-interested agents. These counts are realized outcomes of random sampling, rather than preset quotas. Applying the same sampling procedure to the binomial-likelihood weights yielded 44 efficient, 0 inequity-averse, and 56 self-interested agents.

2.5. Two-stage experiments

In the two-stage games, the LLM simulates Person A deciding whether to complete an allocation or delegate the choice to Person B. Without knowing A's choice, Person B chooses between two allocations. Figure 2 shows the monetary payoffs. Efficiency, fairness, and self-interest continue to influence the two-stage task, so the change in game structure tests whether calibration from single-stage games transfers to prediction in two-stage games.

DOI: 10.54254/2755-2721/2026.39400

The no-persona control answers 100 times in each combination of six games and two roles, A and B. Each of the two sampled groups contains 100 agents, and each agent answers five times in each combination of six games and two roles. Table 1 summarizes the 15,600 model responses.

Table 1. Experimental conditions and numbers of responses

Stage	Design	Responses
Single-stage calibration	6 games × 4 prompting conditions × 100	2,400
Two-stage, no persona	6 games × 2 roles × 100	1,200
Two-stage, constrained least-squares mixture sample	100 agents × 6 games × 2 roles × 5	6,000
Two-stage, binomial-likelihood mixture sample	100 agents × 6 games × 2 roles × 5	6,000
Total	Single- and two-stage model responses	15,600

For each of the 12 game-by-role conditions, it computes the proportions of agents and human participants choosing Left. The paper evaluates prediction using the mean squared error (MSE) and mean absolute error (MAE) between these proportions.

3. Results and analysis

Relative to the no-persona control, the original constrained least-squares calibration reduces the overall MSE across the 12 conditions from 0.082310 to 0.040286, an improvement of 51.1%. MAE falls from 0.242500 to 0.142833, a reduction of 41.1%. Green connecting lines in Figure 2 indicate reductions in error after calibration. These results show that using human data from single-stage games alone to determine the persona composition can substantially improve prediction in structurally different two-stage games. This finding is consistent with the overall trend reported by Horton and colleagues across multiple models and with the idea of transferring theory-driven agents across related tasks [1, 6].

The binomial-likelihood calibrated sample has a cross-experiment MSE of 0.040162, an improvement of 51.2% over the no-persona control. Its MSE is 0.000124 lower than the 0.040286 of the original constrained least-squares sample, a relative improvement of only 0.31%; however, MAE rises from 0.142833 to 0.160000, an increase of approximately 12.0%. Compared with the no-persona control, its prediction improves in seven of the 12 game-by-role conditions, worsens in one, and is unchanged in four. The largest improvement occurs for Person A–Berk21, where absolute error falls from 53 to 9 percentage points; the largest deterioration occurs for Person A–Berk18, where error rises from 1 to 40.6 percentage points. Table 2 summarizes overall cross-experiment prediction error for the three samples.

DOI: 10.54254/2755-2721/2026.39400

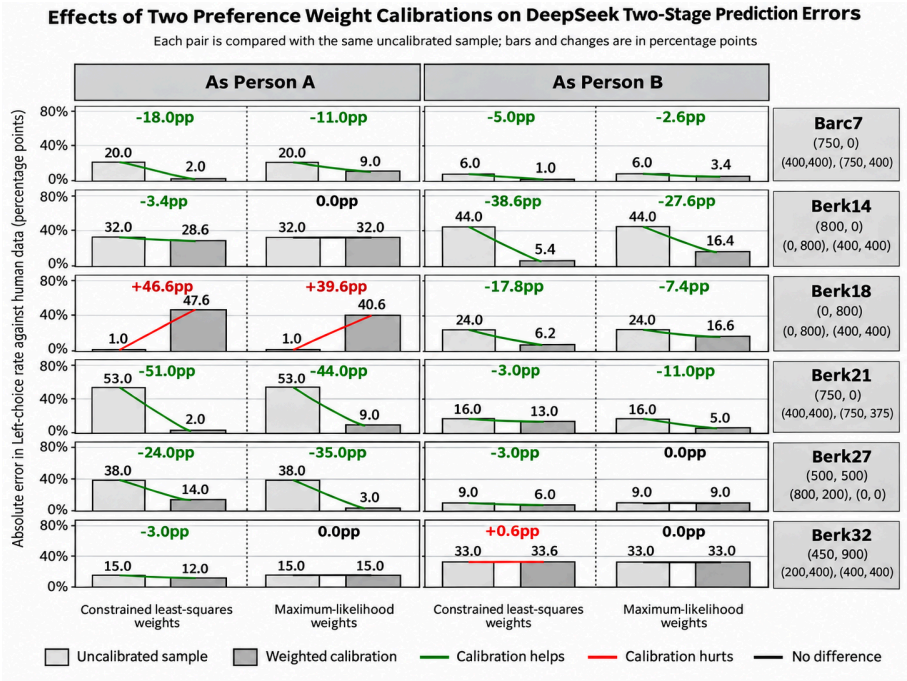


Figure 2. Effects of two preference-weight calibrations on DeepSeek's absolute two-stage prediction errors (photo/picture credit: original)

Table 2. Overall cross-experiment prediction error for the three samples

Sample	Composition	MSE	MAE	MSE change vs. no persona
No persona	100 per condition	0.082310	0.242500	—
Constrained least-squares calibration	52E+3I+45S	0.040286	0.142833	−51.1%
Maximum-likelihood calibration	44E+0I+56S	0.040162	0.160000	−51.2%

The most pronounced structural failure occurs for Person A–Berk18. In this setting, all three possible outcomes have a total payoff of 800: A can choose (0, 800) directly or delegate the choice to B, who then chooses between (0, 800) and (400, 400). For an efficient agent instructed only to "maximize the total payoff of both players", all three outcomes are tied under the efficiency criterion. The prompt specifies neither how to break the tie nor whether to consider control over the decision, certainty of the outcome, or one's own payoff. Consequently, when the efficient-agent weight approaches one half, this unspecified behavior recurs throughout large samples. The absolute error for the original calibrated sample increases to 47.6 percentage points in this condition. Although the maximum-likelihood calibrated sample reduces it to 40.6 percentage points, the error remains high.

DOI: 10.54254/2755-2721/2026.39400

4. Conclusion

This study uses human data from single-stage games to estimate preference weights and evaluates the cross-experiment predictive performance of theory-based, preference-weight-sampled agent groups in two-stage games. Relative to no-persona prompting, the constrained least-squares sample reduces prediction MSE and MAE by 51.1% and 41.1%, respectively. The binomial-likelihood sample has a two-stage MSE 0.31% lower than the least-squares sample, but an MAE 12.0% higher. Under the three fixed preference-agent types and the present sampling setup, binomial-likelihood weight estimation does not show a consistent cross-experiment predictive advantage over constrained least squares. This suggests that changing the mixture-weight estimation method alone may not be sufficient to improve the cross-experiment predictive performance of LLMs. The error in Berk18 also points to the need to examine how agents decide among outcomes tied in total payoff. Future work could allow the strengths of preferences for efficiency, fairness, and self-interest to vary continuously, specify choices in the event of a tie, and compare results across two-stage games with different allocation amounts.

References

[1] Horton, J. J., Filippas, A., & Manning, B. S. (2023). Large language models as simulated economic agents: What can we learn from Homo Silicus? NBER Working Paper 31122, revised 2026. https://doi.org/10.3386/w31122

[2] Aher, G. V., Arriaga, R. I., & Kalai, A. T. (2023). Using large language models to simulate multiple humans and replicate human subject studies. *Proceedings of the 40th International Conference on Machine Learning, PMLR 202*, 337–371. https://proceedings.mlr.press/v202/aher23a.html

[3] Gao, Y., Lee, D., Burtch, G., & Fazelpour, S. (2025). Take caution in using LLMs as human surrogates. *Proceedings of the National Academy of Sciences, 122*(24), e2501660122. https://doi.org/10.1073/pnas.2501660122

[4] Qu, Y., & Wang, J. (2024). Performance and biases of large language models in public opinion simulation. *Humanities and Social Sciences Communications, 11*, 1095. https://doi.org/10.1057/s41599-024-03609-x

[5] Fan, C., Chen, J., Jin, Y., & He, H. (2024). Can large language models serve as rational players in game theory? A systematic analysis. *Proceedings of the AAAI Conference on Artificial Intelligence, 38*(16), 17960–17967. https://doi.org/10.1609/aaai.v38i16.29751

[6] Manning, B. S., & Horton, J. J. (2025). General agents. arXiv:2508.17407v1. https://arxiv.org/abs/2508.17407v1

[7] Charness, G., & Rabin, M. (2002). Understanding social preferences with simple tests. *The Quarterly Journal of Economics, 117*(3), 817–869. https://doi.org/10.1162/003355302760193904

[8] Argyle, L. P., Busby, E. C., Fulda, N., et al. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis, 31*(3), 337–351. https://doi.org/10.1017/pan.2023.2

[9] Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics, 114*(3), 817–868. https://doi.org/10.1162/003355399556151

[10] Leng, Y., Sang, Y., & Agarwal, A. (2024). Reducing disparity between LLMs and humans: Optimal LLM sample calibration. SSRN Working Paper 4802019. https://doi.org/10.2139/ssrn.4802019

[11] Bui, N., Nguyen, H. T., Kumar, S., et al. (2025). Mixture-of-personas language models for population simulation. *Findings of ACL 2025*, 24761–24778. https://doi.org/10.18653/v1/2025.findings-acl.1271

[12] Larooij, M., & Törnberg, P. (2026). Validation is the central challenge for generative social simulation: A critical review of LLMs in agent-based modeling. *Artificial Intelligence Review, 59*, 15. https://doi.org/10.1007/s10462-025-11412-6