Digital Knockoffs

Joint Calibration of LLM Agents

Niuniu Zhang

UCLA Anderson School of Management

PROBLEM

The calibration problem

Social simulations need agents with varied, measurable traits. Describing a person in a prompt need not reproduce their measured profile.

How can we calibrate an LLM agent to multiple traits jointly?

Personality as a concrete test case

The Big Five (OCEAN) summarizes personality along five dimensions. Each target is a vector of scores on a 1–5 scale.

O Openness
Curiosity and receptiveness to new ideas

C Conscientiousness
Organization and dependability

E Extraversion
Sociability and assertiveness

A Agreeableness
Cooperation and consideration

N Neuroticism
Sensitivity to stress and negative emotions

Human target

$$y_p = \begin{bmatrix} 4.5 \\ 4.9 \\ 3.5 \\ 3.8 \\ 4.6 \end{bmatrix}$$

OCEAN order

Example IPIP-50 item
"I am always prepared."
Disagree 1 2 3 4 5 Agree
Score 50 questionnaire answers to obtain five traits.
Goldberg et al., *J. Research in Personality* (2006).

Five controls in a high-dimensional model

Five directions provide five steering controls.

$$\beta = (\beta_O, \beta_C, \beta_E, \beta_A, \beta_N)^\top \in \mathbb{R}^5,$$

$$V = \begin{bmatrix} v_O\, v_C\, v_E\, v_A\, v_N \end{bmatrix} \in \mathbb{R}^{k\times 5}, \qquad \Delta h = V\beta.$$

Columns of V are concept directions; β sets their signed strengths.

$$X \in \mathbb{R}^{T\times d} \xrightarrow[\text{IPIP item}]{\text{LLM}} h + \Delta h \xrightarrow[\text{token representation}]{\text{LLM}} \text{answer 1–5}$$

Add the shift at selected layers; model weights stay fixed.
Steering method: Beaglehole et al., *Science* (2026).

Prompting comparisons

Blank: answer as a generic respondent, without any interventions; the resulting five trait scores define $y_0 = (3.3,\ 3.1,\ 3.2,\ 3.8,\ 3.0)^\top$.
Profile prompting: describe the target in words, then answer the same questionnaire.

"They are structured and dependable in their planning and follow-through…"
Profile excerpt for the example target above.

Steering uses the blank prompt; the target sets β.

METHODS

Learn the response, choose the intervention

Human targets. We randomly sample and fix 500 people from 19,718 complete records in the public OpenPsychometrics IPIP-50 dataset. Real profiles preserve trait correlations and realistic combinations of scores.
Open-Source Psychometrics Project, BIG5 dataset (2014).
Model: Llama 3.1 8B Instruct; the same 50 items for every evaluation.

The central task is to **learn the shared response map**:

$$\underbrace{\beta \in \mathcal{B}}_{\text{five steering strengths}} \longmapsto \underbrace{y(\beta) \in [1,5]^5}_{\text{five measured traits}}, \qquad \mathcal{B} = [-0.3, 0.3]^5.$$

Given a target y_p, choose a dose to minimize trait MAE:

$$\ell_p(z) = \tfrac{1}{5}\|z - y_p\|_1, \qquad \beta_p^\star \in \arg\min_{\beta\in\mathcal{B}} \ell_p(y(\beta)).$$

A common starting library
1,000 uniform interventions, each paired with its observed response: $\mathcal{D} = \{(\beta_i, y_i)\}$.
800 allocated to learning (799 valid); 200 held out for prediction evaluation.

1 Linear predict–then–optimize (PtO)

Estimate J by OLS, then minimize predicted MAE:

$$\hat{y}_{\text{lin}}(\beta) = \underbrace{y_0}_{\text{blank}} + \hat{J}\beta, \qquad \hat{\beta}_p \in \arg\min_{\beta\in\mathcal{B}} \ell_p(\hat{y}_{\text{lin}}(\beta)).$$

2 Bayesian optimization (BO)

Model the unknown map y with five Gaussian processes, each using all five intervention coordinates. At pass $r = 1, \dots, 5$:

$$y_j \mid \mathcal{D}_r \sim \text{GP}(m_{j,r}, k_{j,r}), \qquad j \in \{O, C, E, A, N\}.$$

Recommend a dose by **expected improvement** (EI):

$$\beta_{p,r} \approx \arg\max_{\beta\in\mathcal{B}} \underbrace{\mathbb{E}\left[\left(b_{p,r} - \ell_p(y(\beta))\right)_+ \mid \mathcal{D}_r\right]}_{\text{expected gain over the best observed match}}.$$

$b_{p,r}$ is person p's best observed MAE; $(u)_+ = \max(u, 0)$. EI uses both predicted response and posterior uncertainty.
Frazier, *A Tutorial on Bayesian Optimization* (2018).

$$\mathcal{D}_r \xrightarrow{\text{fit GPs}} y \mid \mathcal{D}_r \xrightarrow{\text{EI}} \{\beta_{p,r}\}_{p=1}^{500} \xrightarrow{\text{execute}} \mathcal{D}_{r+1}$$

Freeze the posterior within each pass. Add all 500 new responses before the next update.

3 Uniform random search

Replace EI with an independent draw $\beta \sim \text{Uniform}(\mathcal{B})$. Both BO and random search use **five passes of 500 doses**: 2,500 interventions per method.
Bergstra and Bengio, *JMLR* (2012).

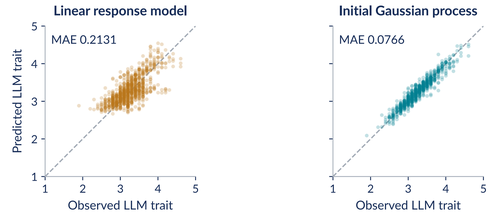
We evaluate trait calibration; downstream behavioral transfer remains untested.

RESULTS

1 Can we predict responses to unseen doses?

Predict the LLM's five traits at 200 held-out doses:

$$\hat{y}_{\text{lin}}(\beta_i) = y_0 + \hat{J}\beta_i, \qquad \hat{y}_{\text{GP}}(\beta_i) = \mathbb{E}[y(\beta_i) \mid \mathcal{D}].$$



Response predictor	Prediction MAE
Blank / no change	0.239
Linear	0.213
Initial GP	0.077
Final GP	0.071

MAE averages absolute errors across 200 doses and five traits. Training: 799 responses for linear / initial GP; 3299 for final GP.

2 Do recommended doses match human targets?

Trait MAE averaged across people, scoring **only the newly assigned dose** for each person in each pass.

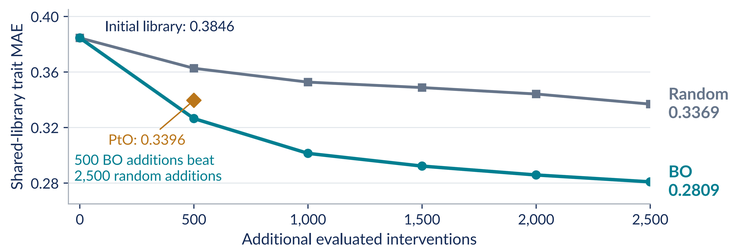
Method	Single run / pass 1	Pass 2	Pass 3	Pass 4	Pass 5
Blank	0.685	–	–	–	–
Profile prompting	0.493	–	–	–	–
Linear PtO†	0.690	–	–	–	–
BO	**0.420**	**0.413**	**0.426**	**0.426**	**0.425**
Uniform random	0.730	0.728	0.731	0.727	0.732

†PtO: 459/500 valid; all other entries: 500. BO also beats PtO on matched people.
BO beats all assigned-response baselines in every pass.

3 Does BO beat collecting more random responses?

A simple baseline collects random doses and selects each person's closest observed response. Initial-library lookup gives **0.3846 MAE**.

$$\text{MAE}_{\text{lookup}}(\mathcal{D}) = \frac{1}{500}\sum_{p=1}^{500} \min_{(\beta,y)\in\mathcal{D}} \ell_p(y).$$



Same initial library and lookup rule; each method adds its own responses. PtO has one 500-dose batch. Held-out doses are excluded.
Prompting references: blank 0.6854; profile 0.4931 (assigned responses, without search). One main run on 500 fixed targets.

Contact: niuniu.zhang.phd@anderson.ucla.edu