SHARED ACTIVATION DIRECTIONS UNDERLYING LLMS' TENDENCY TOWARD SOCIAL RESPONSIBILITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language model (LLM) responses express value orientations across diverse social contexts. However, endorsing socially responsible choices does not necessarily reflect how frequently people make those choices in practice. We investigate this distinction for socially responsible behavior (SRB), encompassing social contribution and harm reduction, across 39 behavioral items and four LLMs. In persona-conditioned sentence completions without explicit response options, all four models exhibit SRB rates above human references on average, even after behaviorally relevant information is added. We estimate model- and layer-specific shared directions from pre-response activations associated with naturally generated SRB and non-SRB responses. Individual-layer interventions modulate mean SRB rates across items in both directions, with model-dependent response ranges and saturation. Multi-layer coefficients selected using known human reference values reduce the mean absolute gap on held-out respondents for the same items under an output-validity criterion. Excluding the target item from direction estimation while retaining selected coefficients preserves the sign of behavioral changes in eligible nonzero interventions and most improvements over baseline. These findings suggest that SRB over-selection across distinct domains includes a shared, adjustable response tendency, connecting value-related responses to their behavioral prevalence in human data.

1 INTRODUCTION

Research has increasingly examined the value systems of large language models (LLMs) and the value orientations reflected in their responses (Ren et al., 2024; Yao et al., 2024; Ye et al., 2025). Among these values, we focus on social responsibility. Social responsibility extends beyond individual interests to encompass contributions to others and the community (Wray-Lake & Syvertsen, 2011; Wray-Lake et al., 2016) and can be expressed through activities such as voting, charitable giving, reporting, and environmental protection. To examine how this value is reflected in LLM responses, we operationalize its behavioral expression, Socially Responsible Behavior (SRB), as actions or behavioral choices that contribute to the community or society, or reduce avoidable harm beyond the actor. Valuing social responsibility, however, does not necessarily translate into practicing it: human action depends on circumstances such as others' cooperation and time constraints (Kessler et al., 2021; Darley & Batson, 1973). Thus, when LLMs consistently generate responses indicating participation in SRB, they may reflect socially positive values while inadequately capturing variation in actual participation across individual circumstances. We therefore examine whether generated SRB response rates systematically exceed human reference rates across everyday domains.

Prior LLM research has documented several response tendencies related to social responsibility, but an integrated analysis connecting them through the shared properties of SRB remains limited. Research on value systems has treated social responsibility as one of several factors (Ye et al., 2025), while responses to specific situations have exhibited preferences for prosocial choices, overestimation of altruistic allocations, and overreporting of voting, charitable giving, and volunteering (Zhou et al., 2025; Capraro et al., 2025; Kazinnik et al., 2026). Connecting these findings through the shared properties of social contribution and harm reduction allows response tendencies observed in different tasks to be analyzed as systematic over-selection of SRB. From this perspective, we inves-

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.

tigate SRB over-selection across behavioral domains and ask whether SRB response tendencies in different behavioral domains can be modulated through a shared activation direction. To this end, we estimate and intervene along a shared activation direction based on the SRB/non-SRB choices in model-generated responses. This extends a unified analysis of SRB response tendencies to an investigation of whether shared activation directions can modulate these responses.

We evaluate four LLMs on 39 behavioral items using common and specialized personas constructed from actual survey respondents' background information. Models complete English behavioral sentences without explicit response options, and we compare SRB rates among codable responses with human reference values derived from survey responses and published aggregate data. We estimate model- and layer-specific directions by pooling pre-response activations across items according to naturally generated SRB/non-SRB labels, and assess their effects through individual-layer interventions. We select multi-layer coefficients on respondents in the coefficient-selection set and apply them unchanged to separate test respondents for the same items. We also exclude each target item from direction estimation, retaining the original coefficients to assess dependence on its contribution.

All four models exhibit SRB rates above human references on average, even with additional behaviorally relevant persona information. Interventions along shared directions modulate mean SRB rates across items in both directions, and selected settings reduce the mean absolute gap on held-out respondents. Target-item exclusion preserves the sign of changes in eligible nonzero interventions and most improvements. Together, these findings suggest that over-selection across distinct behavioral domains includes a shared, adjustable response tendency, rather than comprising only unrelated item-specific discrepancies. Understanding value-related responses therefore requires examining not only which choices models endorse, but also how frequently they generate those choices relative to human data. Our contributions comprise three complementary analyses:

- **Identifying SRB over-selection across behavioral domains.** Using a unified perspective on SRB, we show that persona-conditioned sentence completions exhibit SRB response rates exceeding human reference values across multiple behavioral domains. This tendency persists even after behaviorally relevant persona information is added.
- **Estimating shared directions for response modulation.** We estimate a shared activation direction from SRB/non-SRB response labels across behavioral domains and pre-response activations. Interventions along this direction modulate the mean SRB response rate across items in both directions.
- **Reducing SRB distributional error.** Multi-layer intervention settings selected using human reference values and then held fixed reduce the mean absolute gap relative to baseline on held-out respondents for the same items.

2 Related Work

2.1 Social Responsibility in Human Research

Research on humans conceptualizes social responsibility as a value that extends beyond individual interests to encompass the well-being of others and the community, as well as contributions to the common good (Wray-Lake & Syvertsen, 2011); its scope can extend to society at large and future generations (Schuyt et al., 2010). Research on community and civic engagement treats responsibility and obligation as foundations for participation in shared goals (Nowell & Boyd, 2010; Adler & Goggin, 2005; Zaff et al., 2010) and links a sense of community responsibility to contributions of time and resources and to leadership (Nowell & Boyd, 2014). However, social responsibility does not always translate into action. For example, initially cooperative individuals may cease contributing to a public good after observing others' non-cooperation (Kessler et al., 2021), while people in a hurry may pass someone in distress without helping (Darley & Batson, 1973), illustrating how action depends on others' cooperation and immediate circumstances. This distinction separates the normative importance of social responsibility from the prevalence of corresponding behaviors. We therefore examine whether LLM-generated responses over-select SRB relative to human behavioral distributions across domains.

2.2 LLM Value Orientations

LLM value orientations have been examined using value systems such as Schwartz's basic values and GPLA's five-factor framework (Ren et al., 2024; Yao et al., 2024; Ye et al., 2025). Focusing on the broader structure of LLM values, Ye et al. (2025) treated social responsibility as one of five factors rather than the central object of analysis. In social interaction tasks, models favored prosocial choices even at the expense of goal attainment (Zhou et al., 2025). When predicting human choices in economic games, models underestimated self-interested allocations and overestimated altruistic ones (Capraro et al., 2025). Persona-conditioned simulations did not consistently reproduce human allocation patterns (Ma, 2026), while some models exhibited higher cooperation rates than humans in the repeated prisoner's dilemma (Fontana et al., 2025). Research on social desirability found more favorable self-descriptions during personality assessment (Salecha et al., 2024) and participation rates exceeding external human behavioral benchmarks in responses about voting, charitable giving, and volunteering (Kazinnik et al., 2026). Given that social responsibility is a broader value connecting diverse behaviors (Wray-Lake & Syvertsen, 2011), distributional discrepancies in individual behaviors alone cannot establish whether LLMs exhibit a systematic problem across domains in how they reflect this value in their responses. We adopt a unified perspective on SRB to assess this possibility by examining over-selection relative to human distributions across behavioral domains and test whether shared directions estimated from response contrasts can modulate it.

3 Experimental Methods and Evaluation

3.1 Constructing the Behavioral Evaluation

Behavioral coverage. We evaluate persona-conditioned SRB responses using 39 items spanning six domains: civic and political participation, community contribution and charitable giving, preventive health and blood donation, fraud reporting, transportation, and environmental and consumer behavior. We use individual-level data from ANES 2024 (American National Election Studies, 2026a;b), CPS (U.S. Census Bureau, 2024), BRFSS 2024 (Centers for Disease Control and Prevention, 2024), NHIS 2024 (National Center for Health Statistics, 2025), HPS/HTOPS (U.S. Census Bureau, 2026), NHTS 2022 (Federal Highway Administration, 2022), and U.S. ISSP Environment and Social Networks samples (ISSP Research Group, 2023; 2019), supplemented with aggregate data from OECD EPIC (OECD, 2023) and WHO (World Health Organization, 2026). SRB denotes the response direction corresponding to social contribution or reduced harm to others. Human references primarily use U.S. proportions from binarized valid responses or published aggregate indicators. OECD EPIC references for recycled paper, refillable containers, and reusable shopping bags pool nine countries because U.S.-specific values are unavailable. Both persona conditions share each item's reference. Items without reference values are included in direction estimation but excluded from reference-based evaluation and coefficient selection. Items, domain assignments, and reference calculations are detailed in Appendices A.1–A.2.

Persona construction. We convert actual respondent information into first-person English statements. Common personas describe sociodemographics and living circumstances; specialized personas add domain-relevant information such as political interest, healthcare access, or vehicle ownership. Both exclude target-item responses and closely overlapping past behaviors. A cell is an item–persona-condition pair. Persona variables by source and examples appear in Appendix A.3.

3.2 Measuring SRB Responses

Completion task. We provide the model with a persona and a behavioral sentence containing a blank, and instruct it to generate only a short phrase to fill the blank. For example, the voting item asks the model to complete "I [BLANK] my vote in presidential elections." Both inputs and outputs are in English; neither behavioral response options nor the SRB category label are provided. Behavioral sentences take the form of self-descriptions that preserve the behavioral target and direction of the original survey item while simplifying survey-specific conditions such as recall periods, frequencies, and amounts. Complete prompts and item-specific sentences are provided in Appendix A.3. We evaluate Qwen 3.5 9B and 27B (Qwen Team, 2026), Llama 3.1 8B (Grattafiori et al., 2024), and OLMo 3 7B (Team Olmo et al., 2025) without additional training. We use greedy decoding for coefficient selection, test evaluation, and individual-layer sweeps; verifiable settings for activation

collection used in direction estimation are documented separately in Appendix C.1. The baseline condition applies no activation intervention.

We classify generated responses as SRB, non-SRB, or uncodable using a rule-based parser. SRB and non-SRB correspond, respectively, to performing the target behavior and to not performing it or taking the opposite behavioral direction. Uncodable responses are outputs that are malformed or whose behavioral direction cannot be reliably determined. Coding is based on the meaning of the generated phrase in the context of the original sentence, and uncodable responses are not counted as non-SRB. Let S_c, N_c, and U_c denote the numbers of SRB, non-SRB, and uncodable responses in cell c. We define the conditional SRB rate p_c and the uncodable response rate q_c as

$$p_c = \frac{S_c}{S_c + N_c}, \qquad q_c = \frac{U_c}{S_c + N_c + U_c}. \tag{1}$$

If $S_c + N_c = 0$, p_c is undefined and the cell is excluded from rate aggregation. We define the signed gap g_c and absolute gap a_c relative to the human reference value r_c as

$$g_c = p_c - r_c, \qquad a_c = |p_c - r_c|, \tag{2}$$

which measure the direction and magnitude of the discrepancy, respectively. For each model, we report the mean signed gap, the proportion of cells exceeding the human reference value, and the mean absolute gap, weighting all cells equally. The main analysis includes conditions with $q_c \leq 0.25$ and reports both the number of valid cells and the uncodable response rate. Item-specific coding rules and sensitivity analyses are provided in Appendices B and G.4.

Data splits. For each item, we use 500 respondents as the coefficient-selection set and sample 400 non-overlapping respondents as the test set. In both sets, we construct common and specialized personas for the same respondents. We use the coefficient-selection set to collect baseline activations, estimate directions, select intervention layers, and choose intervention strengths, and apply these settings unchanged to the test set. Model-specific execution settings are provided in Appendix C.

3.3 Steering along Shared Activation Directions

Direction estimation. Motivated by prior work linking model behaviors shared across inputs to low-dimensional activation directions (Arditi et al., 2024; Chen et al., 2025; Rimsky et al., 2024), we estimate a shared direction for each model and layer from naturally generated responses. For each prompt in the coefficient-selection set, we collect the residual-stream activation at the final prompt token immediately before generating the first response token, and associate it with the SRB/non-SRB label of the generated response. Direction estimation uses codable responses from cells with a baseline uncodable response rate of at most 25%, pooled across items and both persona conditions. Model-specific pool composition and class counts are provided in Appendix E. Let $H_{S,\ell}$ and $H_{N,\ell}$ denote the sets of SRB and non-SRB activations at layer ℓ. We normalize the difference between the two group means:

$$\mathbf{d}_\ell = \frac{1}{|H_{S,\ell}|} \sum_{\mathbf{h} \in H_{S,\ell}} \mathbf{h} - \frac{1}{|H_{N,\ell}|} \sum_{\mathbf{h} \in H_{N,\ell}} \mathbf{h}, \qquad \mathbf{v}_\ell = \frac{\mathbf{d}_\ell}{\|\mathbf{d}_\ell\|_2}. \tag{3}$$

Activation intervention. We intervene through additive activation steering, adding the direction to the residual stream during inference without modifying model weights. At token position t in a selected layer ℓ, we modify the activation as

$$\widetilde{\mathbf{h}}_{\ell,t} = \mathbf{h}_{\ell,t} + \alpha_{c,\ell} \mathbf{v}_\ell, \tag{4}$$

applying the same direction during prompt processing and subsequent autoregressive decoding. For each model, we select a layer bundle by comparing candidate layers that distinguish SRB from non-SRB activations on the coefficient-selection set. The final bundles are $\{21, 22, 23\}$ for Qwen 3.5 9B, $\{49, 50, 51, 52, 53\}$ for Qwen 3.5 27B, $\{15, 17, 18\}$ for Llama 3.1 8B, and $\{27, 28, 29, 30, 31\}$ for OLMo 3 7B. Candidate construction and comparison results are provided in Appendix F.1.

Coefficient selection. We treat the intervention coefficients $\alpha_{c,\ell}$ as cell-specific hyperparameters. With the directions and layer bundle fixed, we search over candidate coefficients on the coefficient-selection set. We apply the selected coefficients to the test set without further adjustment and compare the absolute gap and output validity across the baseline, single-layer intervention, and multi-layer bundle intervention. We evaluate changes in the SRB rate as a function of intervention strength

through dose–response analyses that vary the selected layers individually. Details of coefficient search and individual-layer sweeps are provided in Appendices F.2 and F.3, respectively. Sweep aggregation criteria are described in Appendix F.5, and changes in steering results across validity thresholds are reported in Appendix G.4.

Target-item exclusion. To assess whether the shared direction depends on activations from a particular item, we re-estimate the direction after excluding all persona conditions of each target item from the pool. We retain the layer bundle, test respondents, and coefficients selected using the full direction. For cells that are valid under both the full and item-excluded directions, we compare SRB rates, absolute gaps, and uncodable response rates, and assess whether the direction of change and error reduction relative to baseline are preserved. This analysis evaluates robustness to item composition in direction estimation.

4 SRB Over-Selection Is Modulated by Shared Directions

We first examine baseline SRB rates relative to human reference values, then assess whether interventions along a shared direction induce consistent directional changes in responses across items.

4.1 SRB Over-Selection across Behavioral Items

Baseline over-selection. To assess whether persona-conditioned behavioral completions favor SRB responses at rates exceeding human reference values, we compare baseline conditional SRB rates without activation interventions. Following the criteria in Section 3.2 and Equations (1)–(2), we include behavioral cells with an available human reference value and a valid baseline. Evaluation coverage is detailed in Appendix A.1. Table 1 reports positive overall mean signed gaps for all four models, reaching 39.5–51.7 percentage points for Qwen 3.5 27B, Llama 3.1 8B, and OLMo 3 7B. Persona comparisons and item-level exceptions are provided in Appendix D; cell-level baseline rates and human references are reported in Tables 13–16 in Appendix G.1.

Persona effects. Across domains, the overall baseline exhibits positive mean signed gaps in all six domains for Qwen 3.5 27B, Llama, and OLMo, whereas Qwen 3.5 9B exhibits negative mean gaps in community contribution and charitable giving and in transportation. On matched items, adding behaviorally relevant persona information reduces the overall mean gap from +15.5 to +13.9 percentage points for Qwen 3.5 9B, from +48.0 to +32.7 for Qwen 3.5 27B, and from +51.5 to +41.1 for Llama, but increases it from +47.5 to +53.7 for OLMo. These changes also vary across domains. For example, the specialized condition yields larger gaps in preventive health and blood donation for Qwen 3.5 9B and in transportation and environmental and consumer behavior for OLMo. Behaviorally relevant persona information therefore changes response distributions without consistently improving alignment with human reference values, and all four models retain positive mean signed gaps under the specialized condition. The large fraud-reporting gap represents one item and should be interpreted separately from multi-item domains.

4.2 Behavioral Responses to Shared-Direction Steering

Individual-layer interventions. Because steering effects can vary across inputs or even reverse direction (Braun et al., 2025), we assess the applicability of the estimated direction through observed behavioral changes. We evaluate whether the shared direction estimated in Section 3.3 induces changes with a consistent sign across behavioral items. Following the coordinate sweep protocol in Appendix F.3, we intervene on each layer in the bundle individually and compare changes in SRB rates on the test set. Figure 1 reports the mean conditional SRB rate over a fixed set of cells that remain valid across the full coefficient range for each layer. We include cells both with and without human reference values, yielding 60, 72, 60, and 40 candidate cells for Qwen 3.5 9B, Qwen 3.5 27B, Llama 3.1 8B, and OLMo 3 7B, respectively. Cell counts valid across all coefficients at each layer are 59, 59, and 59 for Qwen 3.5 9B; 72, 72, 72, 72, and 72 for Qwen 3.5 27B; 42, 45, and 43 for Llama; and 35, 33, 33, 33, and 38 for OLMo.

Bidirectional response changes. Across all four models, the mean conditional SRB rate generally increases with positive coefficients and decreases with negative coefficients. For Qwen 3.5 9B, the rate is 58.6% at $\alpha = 0$, ranges from 36.7–40.9% across layers at $\alpha = -12$, and reaches 73.5–75.5%

Table 1: Baseline mean signed gaps (pp; bold), with above-reference/valid cell counts below. Shading marks all valid cells per model; indented persona rows use matched items valid under both conditions. Fraud reporting comprises one item.

Model	Overall	Civic/ political	Community/ giving	Health/ blood donation	Fraud reporting	Transport	Environment/ consumption
Qwen 3.5 9B	**+15.3** (32/48)	**+36.5** (10/11)	**-12.0** (3/7)	**+10.0** (5/8)	**+85.2** (2/2)	**-17.1** (0/4)	**+14.7** (12/16)
Common	+15.5 (14/20)	+30.9 (4/5)	+7.0 (2/3)	-3.6 (2/4)	+92.2 (1/1)	-16.5 (0/2)	+18.1 (5/5)
Specialized	+13.9 (13/20)	+33.5 (5/5)	-29.0 (1/3)	+23.6 (3/4)	+78.2 (1/1)	-17.7 (0/2)	+12.1 (3/5)
Qwen 3.5 27B	**+39.5** (52/61)	**+69.3** (14/14)	**+37.3** (9/10)	**+15.2** (5/8)	**+92.4** (2/2)	**+15.6** (2/4)	**+30.4** (20/23)
Common	+48.0 (28/30)	+73.3 (7/7)	+53.5 (5/5)	+21.8 (2/4)	+92.4 (1/1)	+36.3 (2/2)	+37.1 (11/11)
Specialized	+32.7 (24/30)	+65.4 (7/7)	+21.1 (4/5)	+8.7 (3/4)	+92.4 (1/1)	-5.0 (0/2)	+27.4 (9/11)
Llama 3.1 8B	**+44.8** (48/53)	**+73.8** (14/14)	**+29.8** (6/8)	**+34.0** (6/6)	**+92.4** (2/2)	**+50.3** (4/4)	**+27.1** (16/19)
Common	+51.5 (24/24)	+74.8 (7/7)	+37.8 (3/3)	+34.9 (3/3)	+92.4 (1/1)	+61.7 (2/2)	+34.8 (8/8)
Specialized	+41.1 (20/24)	+72.8 (7/7)	+1.6 (1/3)	+33.2 (3/3)	+92.4 (1/1)	+38.8 (2/2)	+25.4 (6/8)
OLMo 3 7B	**+51.7** (35/36)	**+61.9** (6/6)	**+54.1** (8/8)	**+45.3** (5/5)	**+92.4** (2/2)	**+72.6** (4/4)	**+32.4** (10/11)
Common	+47.5 (13/14)	+53.6 (2/2)	+52.1 (3/3)	+31.9 (2/2)	+92.4 (1/1)	+68.7 (2/2)	+26.9 (3/4)
Specialized	+53.7 (14/14)	+48.4 (2/2)	+52.1 (3/3)	+32.8 (2/2)	+92.4 (1/1)	+76.6 (2/2)	+46.8 (4/4)

at $\alpha = 12$. Qwen 3.5 27B starts at 84.3% at $\alpha = 0$ and reaches 71.9–78.4% and 89.1–91.8% at the negative and positive endpoints, respectively, exhibiting a higher baseline rate and a smaller response range than the 9B model. Llama starts at 88.0–90.4% at $\alpha = 0$, decreases to 64.8–73.7% at $\alpha = -4$, and reaches 98.6–100.0% at $\alpha = 12$. OLMo has a high baseline rate of 94.4–95.1% and remains within 87.5–96.3% throughout the sweep. Layer 27 exhibits the largest change, but increases only gradually over the positive coefficient range, remaining close to its high baseline SRB rate. Cell-level response classifications, response ranges, and monotonicity analyses are provided in Appendix F.4; validity-based aggregation criteria and the coverage of the observed data are detailed in Appendix F.5. These results indicate that interventions along a shared direction induce behavioral changes with the expected sign when averaged across items, while response ranges and saturation patterns vary across models.

5 Adjusting SRB Response Distributions

We select intervention coefficients using human reference values to adjust the model's conditional SRB rate along a shared direction, then evaluate the fixed settings on new respondents. We further examine whether the observed effects depend directly on activations from the target item by excluding that item when estimating the direction. This analysis uses activation interventions to adjust response rates toward human behavioral references (Kazinnik et al., 2026) using a shared direction estimated from behavioral contrasts in naturally generated responses, and evaluates both generalization to new respondents under fixed intervention settings and robustness to target-item exclusion.

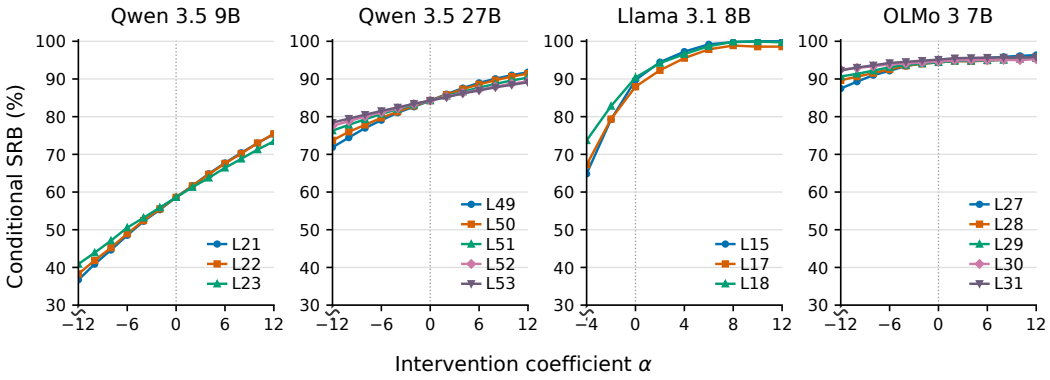


Figure 1: Individual-layer test sweeps along the shared direction. Each curve varies one coefficient with the other bundle coefficients fixed at zero. Conditional SRB rates are equally weighted means over a fixed set of cells with at most 25% uncodable responses at every displayed coefficient. Panels share a 30–100% y-axis; the break omits 0–30%, below the observed minimum of 36.71%.

5.1 Held-Out Steering Results

Evaluation on held-out respondents. We apply the direction, layer bundle, and coefficients selected in Section 3.3 to new test respondents without further adjustment, evaluating steering performance on new respondents for the same behavioral items. We report test results under the primary validity criterion: at most 25% uncodable responses.

Reduced reference gaps. Table 2 shows that the fixed multi-layer intervention reduces the mean absolute gap from the human reference across all four models. On the same cells for which both the baseline and steering conditions are valid, the mean absolute gap decreases from 34.42 to 4.27 percentage points (pp) for Qwen 3.5 9B and from 41.44 to 6.24 pp for Qwen 3.5 27B. It decreases from 46.41 to 8.70 pp for Llama 3.1 8B and from 52.59 to 27.61 pp for OLMo 3 7B. The corresponding mean reductions are 30.15, 35.21, 37.71, and 24.98 pp. However, the substantial residual error for OLMo indicates that adjusting coefficients along a shared direction is insufficient to bring all behavioral response distributions into close agreement with human references.

Output validity. The mean uncodable response rates after steering are 0.65%, 1.48%, 4.65%, and 6.38%, respectively, corresponding to increases of 0.42, 1.33, 3.22, and 3.78 pp over the baseline. All means weight cells equally. Of all selected cells, 47/48, 60/61, 51/53, and 35/36 remain valid under steering on the test set, respectively. Mean reference gaps therefore decrease while most selected cells retain codable outputs within the validity threshold.

Residual errors. Across behavioral domains, OLMo's mean absolute gap decreases from 72.63 to 8.01 pp for transportation but remains at 92.40 pp for fraud reporting. Llama also retains a larger residual gap for community contribution and charitable giving than for other domains, at 24.08 pp after steering. Cell-level response rates and human references are reported in Tables 13–16 in Appendix G.1. These results reveal substantial residual errors despite the average improvements. For Qwen 3.5 27B with the common persona, the test SRB rate for purchasing or boycotting based on a company's political values or business practices is 98.8%, well above the human reference of 35.0%. For Llama 3.1 8B with the specialized persona, volunteering for an organization or association yields an SRB rate of 99.4%, compared with the human reference of 31.4%. For OLMo 3 7B, reporting fraudulent contacts or offers remains at 100.0% under both persona conditions, far from the human reference of 7.6%. These cells satisfy the output validity criterion, underscoring the distinction between producing codable responses and adequately controlling their distribution.

These results show that fixed intervention settings reduce the mean absolute gap relative to the baseline for new respondents answering the same items. The conclusion is restricted to cell-level aggregate distributions adjusted toward known human reference values; it does not imply accurate reproduction of individual behavior or preservation of relationships among behaviors. Optimized single-layer comparisons appear in Appendix G.2. On matched valid cells for Qwen 3.5 9B and

Table 2: Held-out steering performance. (a) Mean absolute gaps (pp) on n cells valid under baseline (B) and steering (S, bold). (b) Mean uncodable rates (%) on these cells and validity counts among all selected cells.

(a) Mean absolute gap by domain

Behavioral domain	Qwen 3.5 9B			Qwen 3.5 27B			Llama 3.1 8B			OLMo 3 7B		
	n	B	**S**	n	B	**S**	n	B	**S**	n	B	**S**
Overall	47	34.42	**4.27**	60	41.44	**6.24**	51	46.41	**8.70**	35	52.59	**27.61**
Civic/political	11	40.26	**4.30**	14	69.33	**11.50**	12	72.84	**9.23**	6	61.86	**11.63**
Community/ giving	7	37.76	**8.39**	10	38.28	**3.91**	8	45.63	**24.08**	8	54.12	**43.48**
Health/ blood donation	8	50.11	**2.33**	8	24.63	**4.09**	6	34.02	**4.31**	4	45.36	**36.37**
Fraud reporting	2	85.15	**0.25**	2	92.40	**1.72**	2	92.40	**3.50**	2	92.40	**92.40**
Transport	4	17.11	**6.30**	4	20.69	**3.92**	4	50.26	**2.39**	4	72.63	**8.01**
Environment/ consumption	15	18.06	**3.34**	22	30.39	**5.56**	19	28.32	**5.15**	11	34.53	**16.95**

(b) Output validity across selected cells

Metric	Qwen 3.5 9B	Qwen 3.5 27B	Llama 3.1 8B	OLMo 3 7B
Baseline uncodable (%)	0.23	0.15	1.43	2.60
Steering uncodable (%)	0.65	1.48	4.65	6.38
Baseline valid / total	48/48	61/61	53/53	36/36
Steering valid / total	47/48	60/61	51/53	35/36

Llama 3.1 8B, multi-layer steering yields smaller mean absolute gaps than a QLoRA-SFT baseline (Appendix G.5). An evaluation in Germany, Spain, and France also finds reduced mean gaps after country-specific direction estimation and coefficient selection (Appendix H).

5.2 Robustness to Target-Item Exclusion

Fixed-coefficient comparison. To assess dependence on activations from the target item, we apply the target-item-excluded directions defined in Section 3.3 using the existing layer bundle and selected coefficients. This tests direction-estimation robustness with fixed coefficients.

Effect retention. Table 3 compares the full and item-excluded directions on jointly valid cells. The mean absolute gap increases from 4.27 to 4.45 pp for Qwen 3.5 9B, from 6.24 to 6.45 pp for Qwen 3.5 27B, from 8.87 to 10.76 pp for Llama 3.1 8B, and from 28.29 to 34.28 pp for OLMo 3 7B. The corresponding increases are 0.19, 0.22, 1.89, and 5.99 pp, respectively. Across all models, every cell in which a nonzero intervention changes the SRB rate under the full direction retains the sign of that change after target-item exclusion. Improvements in the gap relative to baseline are also retained in all eligible cells for both Qwen models and Llama, and in 17 of 18 cells for OLMo. The behavioral control afforded by the excluded directions is therefore strongest for Qwen, partial for Llama, and limited for OLMo.

Relative to the baseline, the mean gaps after target-item exclusion decrease from 34.42 to 4.45 pp for Qwen 3.5 9B and from 41.44 to 6.45 pp for Qwen 3.5 27B. Thus, while exclusion increases the gap by 0.19 and 0.22 pp relative to the full direction, the reductions from baseline remain 29.97

Table 3: Target-item-excluded steering on held-out respondents. Mean absolute gaps (pp) use n jointly valid cells. Baseline: no intervention; Full and Excluded (bold): directions estimated with and without the target item. Retention fractions cover nonzero interventions: Effect preserves Full's rate-change sign; Improvement preserves its gap reduction from baseline. Denominators count cells showing the corresponding effect under Full.

Model	n	Mean gap (pp)			Retained / eligible	
		Baseline	Full	**Excluded**	Effect	Improvement
Qwen 3.5 9B	47	34.42	4.27	**4.45**	41/41	41/41
Qwen 3.5 27B	60	41.44	6.24	**6.45**	49/49	48/48
Llama 3.1 8B	50	47.12	8.87	**10.76**	41/41	41/41
OLMo 3 7B	33	50.99	28.29	**34.28**	18/18	17/18

and 34.99 pp, respectively. For both Qwen models, the performance degradation due to target-item exclusion is therefore small relative to the original steering gains.

Retaining the sign of a change does not guarantee the same accuracy relative to the target value. Definitions of effect and improvement retention are provided in Appendix G.3, and sensitivity to the validity criterion and failure cases in Appendix G.4. Cell-level rates under the baseline, full, and item-excluded directions, with human references and validity flags, appear in Tables 13–16.

6 Discussion

Connecting behavioral contrasts and higher-level values. We estimated shared directions from naturally generated behavioral responses classified using our SRB definition, without explicitly prompting models to endorse social responsibility. These directions modulate mean SRB rates across items in both directions (Section 4.2) and, with cell-specific coefficients, reduce human-reference gaps across behavioral domains (Section 5.1). The connection between behavioral contrasts and intervention effects offers an empirical route to studying higher-level values through their expression in concrete choices. Researchers could use such contrasts to investigate representations shared across behaviors associated with a broader value, then test their relevance through interventions. This also motivates organizing behavioral control around that broader concept, using shared directions to coordinate response changes across otherwise distinct tasks that express it.

Limitations and future work. Our evaluation focuses on behaviors associated with social responsibility. In broader settings, a choice may reflect several values that reinforce or conflict with one another. An important next step is to examine how directions associated with different values interact: whether their effects can be combined, when adjusting one alters the expression of another, and how interventions should account for context-dependent trade-offs. Such work would extend concept-guided behavioral control to choices shaped by multiple values.

7 Conclusion

Across four LLMs, persona-conditioned completions exhibit SRB rates above human references on average, even with additional behaviorally relevant information. Shared directions estimated from naturally generated responses modulate mean SRB rates across items in both directions. Multi-layer coefficients selected using human references reduce gaps on held-out respondents for the same items. Excluding the target item from direction estimation preserves the sign of changes in eligible nonzero interventions and most improvements, although error and uncodable responses increase for some models. These findings support shared directions for modulating SRB responses across behavioral domains.

AI USE STATEMENT

We used LLM-based tools to assist with translation, refine grammar and academic phrasing, and improve terminology consistency, clarity, and concision. The authors reviewed and revised the AI-assisted text to ensure consistency with the intended meaning and research results. The authors take full responsibility for the paper's final content and claims.

ETHICS STATEMENT

This study uses existing survey data and publicly available aggregate statistics to analyze LLM response tendencies. Personas constructed from respondents' background information serve as experimental inputs; model-generated responses should not be treated as those respondents' actual statements or behaviors. Human reference values are empirical benchmarks, not normative targets for ethically desirable conduct. The activation interventions evaluated here can increase or decrease socially responsible responses and could be misused to deliberately distort social simulation results. Applications of these interventions should disclose their use and settings and distinguish generated responses from data collected from human participants.

REPRODUCIBILITY STATEMENT

Section 3 describes the task, response coding, evaluation metrics, direction estimation, and activation interventions. Appendices A and B document the data sources, human reference construction, persona and completion templates, and rule-based response coding. Appendix C specifies model execution settings, coefficient-selection and test splits, and activation extraction and intervention details. Direction-pool composition and weighting are described in Appendix E, while layer selection, coefficient selection, and individual-layer sweeps are detailed in Appendix F. Tables 13–16 provide cell-level test SRB rates, human references, and validity flags for the baseline, single-layer steering, multi-layer steering, and target-item-excluded directions. Additional comparisons and validity-threshold analyses are reported in Appendix G, and the European evaluation protocol and results are presented in Appendix H. To facilitate reproducibility, we provide the codebase required to reproduce our experiments as supplementary material.

AUTHOR CONTRIBUTIONS

If you'd like to, you may include a section for author contributions as is done in many journals. This is optional and at the discretion of the authors.

ACKNOWLEDGMENTS

Use unnumbered third level headings for the acknowledgments. All acknowledgments, including those to funding agencies, go at the end of the paper.

REFERENCES

Richard P Adler and Judy Goggin. What do we mean by "civic engagement"? *Journal of transformative education*, 3(3):236–253, 2005.

American National Election Studies. ANES 2024 time series study full release, 2026a. URL https://electionstudies.org/data-center/2024-time-series-study/. Dataset and documentation, May 19, 2026 version.

American National Election Studies. *User's Guide and Codebook for the ANES 2024 Time Series Voter Validation Supplemental Data*. University of Michigan and Stanford University, Ann Arbor, MI and Palo Alto, CA, 2026b. URL https://electionstudies.org/data-center/2024-time-series-study/. June 23, 2026 version.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. *Advances in Neural Information Processing Systems*, 37:136037–136083, 2024.

Joschka Braun, Carsten Eickhoff, David Krueger, Seyed Ali Bahrainian, and Dmitrii Krasheninnikov. Understanding (un)reliability of steering vectors in language models. In *ICLR 2025 Workshop on Building Trust in Language Models and Applications*, 2025. URL https://openreview.net/forum?id=JZiKuvIK1t.

Valerio Capraro, Roberto Di Paolo, and Veronica Pizziol. A publicly available benchmark for assessing large language models' ability to predict how humans balance self-interest and the interest of others. *Scientific Reports*, 15(1):21428, 2025.

Centers for Disease Control and Prevention. Behavioral risk factor surveillance system survey data, 2024. URL https://www.cdc.gov/brfss/annual_data/annual_2024.html.

Runjin Chen, Andy Arditi, Henry Sleight, Owain Evans, and Jack Lindsey. Persona vectors: Monitoring and controlling character traits in language models. *arXiv preprint arXiv:2507.21509*, 2025.

John M Darley and C Daniel Batson. " from jerusalem to jericho": A study of situational and dispositional variables in helping behavior. *Journal of personality and social psychology*, 27(1): 100, 1973.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 10088–10115. Curran Associates, Inc., 2023. doi: 10.52202/075280-0441. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/1feb87871436031bdc0f2beaa62a049b-Paper-Conference.pdf.

European Social Survey European Research Infrastructure (ESS ERIC). Ess11 - integrated file, edition 4.2, 2026. URL https://ess.sikt.no/en/datafile/242aaa39-3bbb-40f5-98bf-bfb1ce53d8ef.

Federal Highway Administration. 2022 nextgen national household travel survey core data, 2022. URL https://nhts.ornl.gov/. Data version 2.1.

Nicoló Fontana, Francesco Pierri, and Luca Maria Aiello. Nicer than humans: How do large language models behave in the prisoner's dilemma? In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, pp. 522–535, 2025.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The Llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

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. URL https://openreview.net/forum?id=nZeVKeeFYf9.

ISSP Research Group. International social survey programme: Social networks and social resources - issp 2017, 2019. URL https://search.gesis.org/research_data/ZA6980?doi=10.4232/1.13322.

ISSP Research Group. International social survey programme: Environment iv - issp 2020, 2023. URL https://search.gesis.org/research_data/ZA7650?doi=10.4232/1.14153.

Sophia Kazinnik, José Ramón Enríquez, Jacy Anthis, David Nguyen, and Jiaxin Pei. Sex, drugs, and llms: Social desirability bias in language models. *Available at SSRN 6752998*, 2026.

Judd B Kessler, Corinne Low, and Monica Singhal. Social policy instruments and the compliance environment. *Journal of Economic Behavior & Organization*, 192:248–267, 2021.

Ji Ma. Can machines think like humans: A behavioral evaluation of llm agents in dictator games. *Voluntas: International Journal of Voluntary and Nonprofit Organizations*, 37(1):104–118, 2026.

National Center for Health Statistics. National health interview survey, 2024, 2025. URL https://www.cdc.gov/nchs/nhis/documentation/2024-nhis.html. Public-use data file and documentation.

Branda Nowell and Neil Boyd. Viewing community as responsibility as well as resource: Deconstructing the theoretical roots of psychological sense of community. *Journal of community psychology*, 38(7):828–841, 2010.

Branda Nowell and Neil M Boyd. Sense of community responsibility in community collaboratives: Advancing a theory of community as resource and responsibility. *American Journal of Community Psychology*, 54(3-4):229–242, 2014.

OECD. *How Green is Household Behaviour? Sustainable Choices in a Time of Interlocking Crises*. OECD Studies on Environmental Policy and Household Behaviour. OECD Publishing, Paris, 2023. doi: 10.1787/2bbbb663-en. URL https://doi.org/10.1787/2bbbb663-en.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Yuanyi Ren, Haoran Ye, Hanjun Fang, Xin Zhang, and Guojie Song. ValueBench: Towards comprehensively evaluating value orientations and understanding of large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2015–2040, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.111. URL https://aclanthology.org/2024.acl-long.111/.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, 2024.

Aadesh Salecha, Molly E Ireland, Shashanka Subrahmanya, João Sedoc, Lyle H Ungar, and Johannes C Eichstaedt. Large language models display human-like social desirability biases in big five personality surveys. *PNAS nexus*, 3(12):pgae533, 2024.

Theo Schuyt, Rene Bekkers, and JH Smit. The pilantropy scale: a sociological perspective in measuring new forms of pro social behaviour. *Social Work & Society*, 8(1):121–135, 2010.

Team Olmo, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee Chen, Michael Noukhovitch, Nathan Lambert, Pete Walsh, Pradeep Dasigi, Robert Berry, Saumya Malik, Saurabh Shah, Scott Geng, Shane Arora, Shashank Gupta, Taira Anderson, Teng Xiao, Tyler Murray, Tyler Romero, Victoria Graf, Akari Asai, Akshita Bhagia, Alexander Wettig, Alisa Liu, Aman Rangapur, Chloe Anastasiades, Costa Huang, Dustin Schwenk, Harsh Trivedi, Ian Magnusson, Jaron Lochner, Lester James V. Miranda, Maarten Sap, Malia Morgan, Michael Schmitz, Michal Guerquin, Michael Wilson, Regan Huff, Ronan Le Bras, Rui Xin, Rulin Shao, Sam Skjonsberg, Shannon Zejiang Shen, Shuyue Stella Li, Tucker Wilde, Valentina Pyatkin, Will Merrill, Yapei Chang, Yuling Gu, Zhiyuan Zeng, Ashish Sabharwal, Luke Zettlemoyer, Pang Wei Koh, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. Olmo 3, 2025. URL https://arxiv.org/abs/2512.13961.

U.S. Census Bureau. Current population survey, september 2023: Civic engagement and volunteering supplement, 2024. URL https://www.census.gov/data/datasets/2023/demo/cps/cps-volunteer.html. Microdata file; conducted for the Corporation for National and Community Service.

U.S. Census Bureau. HTOPS household pulse april 2025 public use file, 2026. URL https://www.census.gov/programs-surveys/household-pulse-survey/data/datasets.2025.html. Household Trends and Outlook Pulse Survey; data collected April 15–29, 2025; released March 26, 2026.

World Health Organization. *Global status report on blood safety and availability 2025*. World Health Organization, 2026.

Laura Wray-Lake and Amy K Syvertsen. The developmental roots of social responsibility in childhood and adolescence. *New directions for child and adolescent development*, 2011(134):11–25, 2011.

Laura Wray-Lake, Amy K Syvertsen, and Constance A Flanagan. Developmental change in social responsibility during adolescence: An ecological perspective. *Developmental psychology*, 52(1): 130, 2016.

Jing Yao, Xiaoyuan Yi, Yifan Gong, Xiting Wang, and Xing Xie. Value FULCRA: Mapping large language models to the multidimensional spectrum of basic human value. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 8762–8785, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.486. URL https://aclanthology.org/2024.naacl-long.486/.

Haoran Ye, Tianze Zhang, Yuhang Xie, Liyuan Zhang, Yuanyi Ren, Xin Zhang, and Guojie Song. Generative psycho-lexical approach for constructing value systems in large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11968–11991, 2025.

Jonathan Zaff, Michelle Boyd, Yibing Li, Jacqueline V Lerner, and Richard M Lerner. Active and engaged citizenship: Multi-group and longitudinal factorial analysis of an integrated construct of civic engagement. *Journal of Youth and Adolescence*, 39(7):736–750, 2010.

Jinfeng Zhou, Yuxuan Chen, Yihan Shi, Xuanming Zhang, Leqi Lei, Yi Feng, Zexuan Xiong, Miao Yan, Xunzhi Wang, Yaru Cao, Jianing Yin, Shuai Wang, Quanyu Dai, Zhenhua Dong, Hongning Wang, and Minlie Huang. SocialEval: Evaluating social intelligence of large language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 30958–31012, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1496. URL https://aclanthology.org/2025.acl-long.1496/.

A Data, Human References, and Prompts

This appendix documents the correspondence among the analyzed items, human reference values, personas, and input sentences presented to the models. Human reference values serve as benchmarks for comparing generated distributions, rather than normative standards for behavior.

A.1 Item Selection and Coverage

The final analysis comprises 39 items across six domains (Table 4). Domains were assigned according to the target behavior. Political donations were assigned to civic and political participation; donations to environmental organizations, environmental petitions, and environmental demonstrations to environmental and consumer behavior; and whole-blood donation to preventive health and blood donation.

We construct common and specialized personas for each item, yielding 78 possible cells per model. Direction estimation uses codable responses from cells with a baseline uncodable response rate of at most 25%. Under this criterion, 60 cells each are included for Qwen 3.5 9B and Llama 3.1 8B, 72 for Qwen 3.5 27B, and 40 for OLMo 3 7B. Individual-layer sweeps retain these same cells. Human-reference comparisons use 32 behavioral items, excluding four items without comparable references and three perception or intention items (V241215, v26, and Q59); all 39 items remain eligible for direction estimation and individual-layer sweeps. Among cells with existing coefficient selections, the reference-comparison sets contain 48, 61, 53, and 36 cells for Qwen 3.5 9B, Qwen 3.5 27B, Llama 3.1 8B, and OLMo 3 7B, respectively. Test validity is assessed separately for each condition.

Table 4: Behavioral domains, data sources, and item identifiers. Domain counts sum to 39; identifiers are grouped by source. Completion sentences for each item are listed in Table 6.

Behavioral domain	Data source	Item identifiers
Civic and political participation (8)	ANES 2024	`val_turnout24`, `V241215`, `V242029`, `V242030`, `V242031`, `V242034`, `V242036`, `V242037`.
Community contribution and charitable giving (5)	ANES 2024; CPS 2023; ISSP Social Networks	ANES: `V242035`. CPS: `PES6`, `PES7`, `PES18`. ISSP: `v19`.
Preventive health and blood donation (4)	BRFSS 2024; NHIS 2024; WHO	BRFSS: `FLUSHOT7`. NHIS: `SHTFLU12M_A`, `SHTCVD191_A`. WHO: `WHO_whole_blood_per_1000`.
Fraud reporting (1)	HPS/HTOPS 2025	`FRAUD2`.
Transportation (3)	NHTS 2022; OECD EPIC 2022	NHTS: `LAST30_PT`. OECD: `Q50`, `Q59`.
Environmental and consumer behavior (18)	ISSP Environment; OECD EPIC 2022	ISSP: `v26`, `v52`, `v53`, `v54`, `v55`, `v56`, `v57`. OECD: `Q40a`, `Q40b`, `Q40c`, `Q40d`, `Q40e`, `Q74_recycled_paper`, `Q74_refill`, `Q74_reusable_bag`, `Q84g`, `Q85a`, `Q86h`.

A.2 Human Reference Construction

Individual-level reference values are obtained from ANES 2024 and its voter validation data (American National Election Studies, 2026a;b), the CPS 2023 Civic Engagement and Volunteering Supplement (U.S. Census Bureau, 2024), BRFSS 2024 (Centers for Disease Control and Prevention, 2024), NHIS 2024 (National Center for Health Statistics, 2025), the April 2025 HPS/HTOPS data (U.S. Census Bureau, 2026), NHTS 2022 (Federal Highway Administration, 2022), and the U.S. samples of ISSP Environment and Social Networks (ISSP Research Group, 2023; 2019). We use unweighted proportions whose denominators comprise responses valid for the relevant binary contrast among the 500 respondents sampled for each item. The same item-specific value is applied to both persona conditions and all four models.

For items asking whether a behavior was performed, performance is coded as SRB and nonperformance as non-SRB. ANES `val_turnout24` is based on voter-file linkage rather than self-report; only linked records indicating voting or nonvoting are used. For CPS `PES6` and ISSP Social Networks `v19`, performing the behavior at least once is coded as affirmative. Items `v52` and `v53` contrast performing the behavior at least sometimes with never doing so. Refusals, unknown, erroneous, and inapplicable responses are not counted as non-SRB. For HPS `FRAUD2`, the denominator includes only respondents who received fraudulent contacts or offers. The completion sentence asks about reporting scams without explicitly specifying this exposure condition, so the reference describes reporting conditional on exposure rather than an unconditional reporting rate. Thus, original response categories and denominators differ across items; item-specific values are reported in the Human columns of Tables 13–16.

For OECD EPIC 2022 items (OECD, 2023), we use published aggregate values from the report. For recycled paper, refillable containers, and reusable shopping bags, for which U.S. values are not published, we use pooled values across nine countries. No comparable reference values are assigned to `Q40c`, `Q84g`, `Q85a`, or `Q86h`. Item `Q50` concerns public transportation use by the respondent or a household member. We do not recalculate these proportions because the published report does not provide item-level individual responses and denominators. The WHO whole-blood donation indicator (World Health Organization, 2026) records donations per 1,000 population in 2023; we use the U.S. value of 28.2/1,000 for numerical comparison. Because this count includes repeated donations, it cannot be interpreted as the proportion of individuals who have donated blood.

For aggregate data, OECD environmental and consumer items use personas from ISSP Environment, transportation items from NHTS, food-sharing items from CPS, and the WHO blood donation item from NHIS. These personas are not linked to the original respondents underlying the aggregate indicators. Consequently, these comparisons use external aggregate references rather than observed behavior from the individuals supplying the persona profiles. The pooled nine-country OECD references additionally differ in geographic coverage from the U.S. persona samples.

The behavioral completion sentences omit some recall-period, frequency, and amount conditions specified in the original surveys. For example, ANES participation during the past 12 months, the CPS donation threshold of more than 25 dollars, and NHTS use during the past 30 days are reformulated as general first-person behavioral statements. Consequently, differences between model-generated rates and human reference values may reflect measurement differences as well as behavioral direction.

A.3 Persona Fields and Completion Templates

Common personas express eight types of information—age, gender, race/ethnicity or social group, education, income/socioeconomic position, marital/household status, employment/occupation, and residence—as first-person statements in English, adapted to each data source. Specialized personas append the behavior-related context in Table 5 to the same respondent's common profile. Responses to the target item and past behaviors closely overlapping with it are excluded from both conditions. We use only respondents with valid information for both common and specialized personas, applying only the relevant branch of survey items such as ANES party identification strength or leaning. For aggregate items, persona sources and their relationship to the human reference populations are described in Appendix A.2.

Table 5: Variables and information units added to specialized personas. All information in the common persona is retained.

Source	Source variables	Additional context
ANES	`V241005`, `V241177`, `V241221`, `V241222`, `V241223`	Campaign interest, ideology, party identification, and identification strength or leaning (4 units).
CPS	`PES1`, `PES2`, `PES10`	Contact with family/friends, discussion of issues, and news consumption (3 units).
BRFSS	`GENHLTH`, `PRIMINS2`, `PERSDOC3`, `MEDCOST1`	General health, source of health coverage, usual healthcare provider, and cost-related barriers to care (4 units).
NHIS	`PHSTAT_A`, `NOTCOV_A`, `USUALPL_A`, `PAYBLL12M_A`	Health, health coverage status, usual place of care, and difficulty paying medical bills (4 units).
HPS/HTOPS	`TRUST2_POLICE`, `TRUST2_CRIMJUST`	Trust in the police and criminal justice system (2 units).
NHTS	`HHVEHCNT`, `DRIVER`, `RAIL`, `MEDCOND`	Vehicle ownership, driving status, heavy-rail access, and conditions/disabilities limiting mobility (4 units).
ISSP Environment	`v15`, `v17`, `v18`, `v19`, `v22`, `v25`	Environmental concern, causes and global/national impacts of climate change, environmental harm from modern life, and economic growth and the environment (6 units; 4 under structural branching).
ISSP Social Networks	`v1--v10`, `v44`, `v46`, `v47`, `v31--v33`, `v34`, `v35`	Network diversity, opportunities for social contact and mobilization, isolation, and generalized trust (18 source variables summarized into 4 information units).

Figure 2 shows common and specialized personas for the same ANES respondent (source row 0), together with the shared instructions. The specialized condition appends political information to the common profile. Both conditions replace only {`persona`} and {`sentence`} with the respondent profile and item sentence.

Table 6 lists the English sentences substituted for {`sentence`} for the final 39 items. Neither human reference values nor SRB labels are presented to the models.

Common persona

I am a 50-year-old Hispanic man. I have a bachelor's degree, my household income is between $100,000 and $249,999, I am married, I am currently working, and I live in the South.

Specialized persona: common profile +

I am very interested in political campaigns. I identify as conservative. I identify with the Republican Party. I consider that party identification strong.

System message

You will complete a sentence from the first-person profile supplied by the user. Return only the short phrase that fills [BLANK]. Do not repeat the sentence or explain your answer.

User message

```
Respondent profile:
{persona}

Sentence:
"{sentence}"

Fill [BLANK] with a short phrase only:
```

Figure 2: Persona examples and the shared completion prompt. The specialized profile combines the common profile with the additional political information. System and user messages are reproduced as supplied to the models.

Table 6: English completion sentences for all 39 items; models generate only the phrase replacing [BLANK].

Item	Sentence
Civic and political participation	
val_turnout24	I [BLANK] my vote in presidential elections.
V241215	I see voting as a [BLANK].
V242029	I [BLANK] in protests about political or social issues.
V242030	I [BLANK] petitions about political or social issues.
V242031	I [BLANK] to organizations working on political or social issues.
V242034	I [BLANK] public meetings about local issues.
V242036	I [BLANK] federal officials or their staff about issues that matter to me.
V242037	I [BLANK] a company's products because of its political values or business practices.
Community contribution and charitable giving	
V242035	I [BLANK] my time to an organization or association.
PES6	I [BLANK] my neighbors with everyday favors.
PES7	I [BLANK] with my neighbors to improve our community.
PES18	I [BLANK] money or goods to nonprofit organizations.
v19	I [BLANK] in volunteer activities with charitable or religious organizations.
Preventive health and blood donation	
FLUSHOT7	I [BLANK] a flu shot.
SHTFLU12M_A	I [BLANK] a flu shot.
SHTCVD191_A	I [BLANK] vaccinated against COVID-19.
WHO_whole_blood_per_1000	I [BLANK] whole blood.

Continued on next page

Table 6: English completion sentences (continued).

Item	Sentence
Fraud reporting	
FRAUD2	I [BLANK] scams to the authorities.
Transportation	
LAST30_PT	I [BLANK] public transportation.
Q50	I or someone in my household [BLANK] public transportation.
Q59	I [BLANK] my car less when better public transportation is available.
Environmental and consumer behavior	
v26	I [BLANK] extra for environmentally friendly options.
v52	I [BLANK] recyclables from other waste.
v53	I [BLANK] products that I consider harmful to the environment.
v54	I [BLANK] a member of an environmental group.
v55	I [BLANK] petitions about environmental issues.
v56	I [BLANK] to environmental organizations.
v57	I [BLANK] in demonstrations about environmental issues.
Q40a	I [BLANK] off the lights when I leave a room.
Q40b	I [BLANK] my use of heating and air conditioning.
Q40c	I [BLANK] for a full load before running a washing machine or dishwasher.
Q40d	I [BLANK] my hot-water use when I wash dishes or shower.
Q40e	I [BLANK] dry my laundry instead of using a dryer.
Q74_recycled_paper	I [BLANK] paper products made with recycled materials when I shop.
Q74_refill	I [BLANK] refillable containers.
Q74_reusable_bag	I [BLANK] reusable shopping bags.
Q84g	I [BLANK] Fairtrade-certified food when I shop.
Q85a	I [BLANK] food with minimal packaging when I shop.
Q86h	I [BLANK] goods to a food-sharing scheme.

B Response Coding and Validity

B.1 Coding Rules

Using a rule-based parser, we classify responses as SRB, non-SRB, or uncodable by combining the generated phrase with the original behavioral sentence. Direct negation of the target behavior is checked before affirmative expressions, and responses whose behavioral direction cannot be determined are not counted as non-SRB. Table 7 summarizes the main coding categories.

B.2 Ambiguous and Invalid Responses

Outputs expressing uncertainty, ability, willingness, or preference without determining the target behavior, copied personas, multiple answers, repetition, and malformed outputs are coded as uncodable. Full-sentence responses are coded when their behavioral direction is clear, whereas isolated adverbs or incomplete phrases are accepted only under item-specific rules. Thus, the same generated phrase may receive different labels depending on the behavior expressed after the blank.

Table 7: Coding principles and examples by behavioral context. Examples are interpreted within the original sentence rather than applied as standalone keyword rules.

Behavior type	SRB examples	Non-SRB and coding conditions
Voting/participation	vote, attend, participate	Direct negation of the target action verb
Donation/sharing	donate, give, contribute	Direct negation of the target donation behavior
Vaccination	get, receive, am vaccinated	Direct negation of vaccination; interpreted with the tense and complement of the fixed sentence
Fraud reporting	report, notify	Direct negation of reporting; verify that the phrase denotes notifying the relevant authority
Avoiding environmentally harmful products	avoid, refrain from; do not buy	buy, purchase; direct negation of avoidance verbs
Energy/water conservation	reduce, limit, conserve	increase, use more; direct negation of reduction behaviors
Purchases/container use	buy, choose, use	Direct negation of purchasing or using the target product
Incomplete short phrases	Coded only when context uniquely determines the behavioral direction	Other behaviors, repetition, empty outputs, and unclear attitudes are coded as uncodable

C Reproducibility Details

C.1 Models and Generation Settings

We run Qwen 3.5 9B and 27B, Llama 3.1 8B Instruct, and OLMo 3 7B Instruct using the tokenizer and chat template associated with each checkpoint. Models are loaded in BF16 without additional training or quantization, with input and generation lengths limited to 1,024 tokens and 18 new tokens, respectively. Inputs are left-padded, and special tokens are removed from the generated suffix before response coding. Thinking is disabled when supported by the template.

Activations are collected with a batch size of 4. Responses used to assign SRB/non-SRB labels to the collected pre-response activations are sampled with temperature 0.7 and top-p 0.9. Coefficient selection, test evaluation, and individual-layer sweeps use greedy decoding.

C.2 Sampling and Data Splits

For each item, we sample 500 eligible respondents for coefficient selection and then 400 test respondents excluding those selected for coefficient selection on that item. The same item-specific coefficient-selection and test respondent sets are used across all four models, both persona conditions, and intervention conditions. Respondents may be reused across different items. Items without human reference values also receive a test sample of 400 respondents who do not overlap with respondents in the coefficient-selection set.

Directions, layer bundles, and intervention coefficients are selected on the coefficient-selection set and applied unchanged to the test set. Coefficient-selection subsets used for candidate screening are also excluded from test evaluation. Model-specific cell inclusion is reported in Appendix A.1; validity is reassessed for each test condition.

C.3 Activation Extraction and Intervention

We extract activations from each decoder layer at the final prompt token immediately before generation of the first response token. Layers are indexed from zero; activations are collected from 32 layers for Qwen 9B, Llama, and OLMo, and 64 layers for Qwen 27B. Model computation uses BF16, stored activations use FP16, and direction computation uses FP32 or higher precision. Class-

specific sums for Qwen 27B are accumulated in FP64. No additional token-norm-based rescaling is applied to the L2-normalized directions in Equation (3).

The intervention in Equation (4) is implemented through a forward hook on the output of each selected decoder block. During prompt processing, the direction is added only at the final token; during subsequent KV-cache-based decoding, the same direction and coefficient are added at the current generation position. Earlier prompt tokens are not modified, and layers with zero coefficients are skipped. Pool composition is provided in Appendix E.1, and coefficient selection in Appendix F.2.

D Baseline Heterogeneity and Persona Effects

D.1 Evaluation Coverage

Baseline results are evaluated using the difference between the conditional SRB rate and the human reference value; model-specific cell counts are provided in Appendix A.1. Gap calculations use behavioral items with available human references, following Appendix A.1. For the reference-based evaluation cells, Tables 13–16 in Appendix G.1 report conditional SRB rates, human references, and condition-specific validity flags. This appendix presents persona effects and exceptions that are not apparent from the main-text averages alone.

D.2 Paired Persona Comparisons

Persona comparisons use matched reference-based evaluation cells for which both common and specialized conditions are valid. We define the change in conditional SRB rate as $\Delta p = p_{\text{specialized}} - p_{\text{common}}$ and the change in absolute gap as $\Delta a = |p_{\text{specialized}} - r| - |p_{\text{common}} - r|$. Values of $\Delta a < 0$ indicate improvement and $\Delta a > 0$ indicate deterioration; absolute changes of at most 10^{-9} percentage points (pp) are treated as numerical ties. Because conditional SRB rates use the codable responses in each condition as their denominators, Δp is distinct from an individual-level label-transition rate.

Table 8 reports improvements, deteriorations, and ties for paired items. For Qwen 3.5 27B, the absolute gap decreases for 28 of 30 items, yet the specialized condition still exceeds the human reference value for 24 items. OLMo shows three improvements and five deteriorations, with all 14 paired items exceeding the reference value under the specialized condition. Changes induced by additional information do not guarantee agreement with human reference distributions.

Table 8: Effects of additional persona information on the absolute gap. Percentages in parentheses use the number of paired items for the corresponding model as the denominator.

Model	Items	Improved	Worsened	Tied
Qwen 3.5 9B	20	11 (55.0%)	7 (35.0%)	2 (10.0%)
Qwen 3.5 27B	30	28 (93.3%)	1 (3.3%)	1 (3.3%)
Llama 3.1 8B	24	16 (66.7%)	2 (8.3%)	6 (25.0%)
OLMo 3 7B	14	3 (21.4%)	5 (35.7%)	6 (42.9%)

D.3 Item-Level Patterns and Examples

Item-level results show that a decrease in SRB need not reduce error. For nonprofit donations (PES18), Qwen 3.5 9B decreases from 87.3% under the common condition to 12.3% under the specialized condition, overshooting the human reference value of 53.4% and increasing the absolute gap from 33.9 to 41.2 pp. For the same item, Qwen 3.5 27B and Llama show smaller absolute gaps, whereas OLMo maintains an SRB rate of 100% under both conditions.

Underproduction also persists. For COVID-19 vaccination (SHTCVD191_A), Qwen 3.5 9B increases from 7.8% to 25.5% but remains 55.1 pp below the human reference value of 80.6%. Conversely, Llama's influenza vaccination rate (SHTFLU12M_A) increases from 87.0% to 92.0%, widening the signed gap under the specialized condition to +43.4 pp. A reduction in the mean signed gap therefore does not imply consistent agreement with human references for individual items.

Bidirectional changes are also observed in generated responses. For the same respondent, Qwen 3.5 9B's nonprofit donation response changes from "donate" under the common condition to "do not donate" under the specialized condition, while one influenza vaccination response (FLUSHOT7) changes from "did not receive" to "got". These examples are coded in the contexts of "I [BLANK] money or goods to nonprofit organizations." and "I [BLANK] a flu shot.", respectively. They illustrate label transitions and are not representative samples of their within-item frequencies.

E Direction Pool Composition and Weighting

This appendix describes the response sets used to construct the shared directions in Section 3.3. We specify model-specific pool inclusion criteria, report activation counts by persona and class, and define the weights used in the pooled class means.

E.1 Direction Pool Composition

Pools comprise baseline cells that meet the inclusion criterion for each model. Responses are labeled using the rule-based parser described in Appendix B.

Each pool comprises cells whose baseline uncodable response rate was at most 25% when the pool was constructed. Only pre-response activations associated with SRB and non-SRB responses in those cells enter the class means. Uncodable responses are reported to document sample composition but are excluded from direction estimation. Both classes need not be present within an individual cell: a cell whose codable responses are all SRB still contributes to the overall SRB mean. Items without human reference values are also included if they satisfy this criterion, so the number of direction-estimation cells differs from the number used for reference-based coefficient selection and test evaluation. Sample sizes here count activations associated with prompt–response pairs, rather than unique individuals. Common and specialized conditions for the same respondent are separate observations, and respondents may be reused across items.

Let $n_{o,S}$ and $n_{o,N}$ denote the class-specific observation counts for item o, and n_S and n_N the corresponding counts in the full pool. The item's share of all codable responses and its weights in the class means are

$$w_o = \frac{n_{o,S} + n_{o,N}}{n_S + n_N}, \qquad w_{o,S} = \frac{n_{o,S}}{n_S}, \qquad w_{o,N} = \frac{n_{o,N}}{n_N}. \tag{5}$$

Writing the item means for classes with observed responses as $\boldsymbol{\mu}_{o,S,\ell}$ and $\boldsymbol{\mu}_{o,N,\ell}$, the pooled contrast in Equation (3) is $\mathbf{d}_\ell = \sum_o w_{o,S}\boldsymbol{\mu}_{o,S,\ell} - \sum_o w_{o,N}\boldsymbol{\mu}_{o,N,\ell}$. Terms for classes with zero observations are omitted from the sums. This differs from equally averaging item-level unit vectors or cell-level contrasts. Because item weights may differ between classes, the pooled contrast may reflect differences in item and persona composition as well as behavioral labels. An item's contribution to each class mean is determined by its class-specific weight and mean activation; the overall response share w_o alone does not characterize its contribution to the pooled contrast.

Table 9 reports activation counts by persona and response class.

The table summarizes pool composition by model, persona, and response class, while Equation (5) defines item weights in the pooled means. The behavioral effects of item-excluded directions are evaluated in Section 5.2 and Appendix G.3.

F Layer Selection, Coefficient Selection, and Dose–Response

F.1 Layer Bundle Selection

Layer bundles are selected using the separation between SRB and non-SRB activations in the coefficient-selection data and intervention results for candidate bundles. The separation score is the squared difference between the class means divided by within-class variance. We construct three- and five-layer candidates using contiguous groups of high-scoring layers and sparse sets of top-scoring layers. Table 10 reports the candidate layers and mean separation scores for all four models.

Table 9: Activation counts by persona and class in the direction-estimation pools. Uncodable counts responses excluded from direction means within the included cells.

Model	Persona	Cells	SRB	non-SRB	Uncodable
Qwen 3.5 9B	common	35	10,555	6,321	624
	specialized	25	6,781	4,846	873
Qwen 3.5 27B	common	35	15,605	1,632	263
	specialized	37	14,188	3,767	545
Llama 3.1 8B	common	30	13,668	752	580
	specialized	30	11,841	2,146	1,013
OLMo 3 7B	common	21	8,957	650	893
	specialized	19	8,660	187	653

The final bundles are $\{21, 22, 23\}$ for Qwen 9B, $\{49, 50, 51, 52, 53\}$ for Qwen 27B, $\{15, 17, 18\}$ for Llama, and $\{27, 28, 29, 30, 31\}$ for OLMo. The separation score is used to construct candidates; the final bundle is not always the candidate with the highest score.

Table 10: Candidate layer bundles and mean separation scores for all four models. Contiguous/sparse denotes cases where the two construction methods yield the same layer set.

Model	Candidate	Layers	$\overline{D}$
Qwen 9B	contiguous/sparse-3	21,22,23	0.05047
Qwen 9B	contiguous/sparse-5	20,21,22,23,24	0.04969
Qwen 27B	contiguous-3	49,50,51	0.24748
Qwen 27B	sparse-3	49,50,52	0.25232
Qwen 27B	contiguous-5	49,50,51,52,53	0.23961
Qwen 27B	sparse-5	49,50,52,54,59	0.24277
Llama	contiguous-3	16,17,18	0.19540
Llama	sparse-3	15,17,18	0.19938
Llama	contiguous-5	15,16,17,18,19	0.18807
Llama	sparse-5	2,3,15,17,18	0.19294
OLMo	contiguous-3	29,30,31	0.12110
OLMo	sparse-3	28,29,31	0.12170
OLMo	contiguous-5	27,28,29,30,31	0.12061
OLMo	sparse-5	24,28,29,30,31	0.12078

F.2 Coefficient Selection

With the bundle and directions fixed, we search for layer-specific coefficients using 500 respondents in the coefficient-selection set per cell. For each uncodable-response cutoff, we select the coefficient vector with the smallest absolute gap from the human reference value among all evaluated settings that satisfy that cutoff. Observations are shared across cutoff-specific searches, and final selection uses the complete set of evaluated settings. The main analysis uses a cutoff of 25%. Selected coefficients are not readjusted on the test set.

All coefficients are zero in 7 cells for Qwen 9B, 12 for Qwen 27B, 8 for Llama, and 6 for OLMo. These include cases in which the baseline is already close to the target and cases in which no valid improvement is found; they are therefore not uniformly classified as successes or failures. These cells are excluded from effect- and improvement-retention denominators for nonzero interventions.

F.3 Layer Sweep Protocol

We vary the coefficient of each bundle layer individually while fixing the coefficients of all other layers at zero. Both Qwen models and OLMo are evaluated at $\alpha = -12, -10, \ldots, 12$, and Llama at

$-4, -2, \ldots, 12$. This test dose–response experiment is distinct from coefficient searches in which other layers may have nonzero coefficients. We include cells with and without human reference values and average conditional SRB rates with equal cell weights over the same set of cells whose uncodable response rates are at most 25% at every displayed coefficient for a given layer.

Table 11 reports the cells retained for each layer sweep.

Table 11: Total sweep coverage and numbers of jointly valid cells by layer. Entries in the last two columns correspond in order.

Model	Total	Layers	Jointly valid cells
Qwen 3.5 9B	60	21, 22, 23	59, 59, 59
Qwen 3.5 27B	72	49, 50, 51, 52, 53	72, 72, 72, 72, 72
Llama 3.1 8B	60	15, 17, 18	42, 45, 43
OLMo 3 7B	40	27, 28, 29, 30, 31	35, 33, 33, 33, 38

F.4 Cell-Level Behavioral Consistency

For each cell–layer pair, we compare valid observations with the baseline. Responses are classified as flat if all changes are within 1 pp. Otherwise, increases at positive coefficients or decreases at negative coefficients are classified as expected responses, and changes in the reverse direction as opposite responses. Pairs exhibiting both directions are classified as mixed. The response range is the difference between the maximum and minimum conditional SRB rates among valid observations. Monotonicity is measured by Kendall's τ_b between α and the SRB rate over valid observations. Cases in which τ_b is undefined because all observations are tied are excluded from its mean.

Table 12 summarizes response consistency across the individual-layer sweeps.

Table 12: Responses by cell–layer pair. Categories give counts of expected/flat/opposite/mixed responses; the median response range is in pp. τ_b is averaged over pairs for which it is defined.

Model	Pairs	Response categories	Range	τ_b
Qwen 3.5 9B	180	162/12/4/2	31.19	0.906
Qwen 3.5 27B	360	256/99/0/5	5.69	0.922
Llama 3.1 8B	180	102/60/10/8	7.75	0.647
OLMo 3 7B	200	74/115/11/0	0.25	0.610

F.5 Sweep Validity and Aggregation

We use the jointly valid set across the full grid to prevent the cell composition of the mean curve from changing across coefficients. Because this restriction excludes cells with unstable outputs, the results do not establish unconditional controllability across all items. The minimum value across the four main-text panels is 36.71%; all panels use the same 30–100% scale, omitting only the 0–30% interval containing no observations.

G Steering Comparisons and Robustness

G.1 Held-Out Results by Cell

Tables 13–16 report cell-level conditional SRB rates for the baseline, single-layer intervention, multi-layer intervention, and item-excluded directions, evaluated on the same test respondents within each model, alongside human references. All rates are percentages. The suffix C or S in a cell identifier denotes the common or specialized persona, respectively. The result columns B, S, M, and E denote the baseline, single-layer steering, multi-layer steering, and steering with the target-item-excluded direction, respectively. Steering settings are selected under the 25% validity criterion for

coefficient selection. A dagger (†) marks a test uncodable response rate exceeding 25% in the corresponding condition. Uncodable responses are not counted as non-SRB.

Long item identifiers are abbreviated as WHO-blood (`WHO_whole_blood_per_1000`), Q74-paper (`Q74_recycled_paper`), Q74-bag (`Q74_reusable_bag`), and Q74-refill (`Q74_refill`). The tables retain all 48, 61, 53, and 36 behavioral evaluation cells for Qwen 3.5 9B, Qwen 3.5 27B, Llama 3.1 8B, and OLMo 3 7B, respectively. Values that fail validity are shown but are excluded from comparisons requiring that condition to be valid. Human references and their measurement conventions are described in Appendix A.2; the WHO value is a donation count per population expressed on the percentage scale, rather than an individual participation proportion. Conditional SRB rates are shown to two decimal places; aggregates use unrounded values.

Cells exceeding the validity threshold after multi-layer intervention are the specialized recycled-paper purchase condition for both Qwen models, the common and specialized conditions for contributions to organizations addressing political or social issues (V242031) for Llama, and the common influenza vaccination condition (FLUSHOT7) for OLMo. These cells are not counted as valid target attainment even when their gaps from human reference values are small.

Table 13: Cell-level conditional SRB rates and human references (%) for Qwen 3.5 9B. B/S/M/E denote baseline, single-layer steering, multi-layer steering, and item-excluded steering. Cell suffixes C/S denote common/specialized personas; † indicates a test uncodable response rate above 25%.

Cell	Human	B	S	M	E
FLUSHOT7/C	51.00	67.75	54.75	57.50	58.00
FLUSHOT7/S	51.00	91.25	50.25	48.25	50.75
FRAUD2/C	7.60	99.75	1.25	7.50	11.00
FRAUD2/S	7.60	85.75	3.75	8.00	10.75
LAST30_PT/C	11.20	5.00	6.75	5.00	5.00
LAST30_PT/S	11.20	0.25	8.50	10.75	11.75
PES18/C	53.40	87.25	56.50	56.50	56.25
PES18/S	53.40	12.25	51.75	57.00	56.25
PES6/C	58.80	40.75	57.75	56.75	56.75
Q40a/C	99.20	100.00	100.00	100.00	100.00
Q40a/S	99.20	92.48	94.50	98.25	98.25
Q40d/C	88.60	98.50	80.75	80.75	80.00
Q40d/S	88.60	85.25	85.25	89.25	90.25
Q40e/C	62.90	100.00	100.00	67.96	67.17
Q50/C	28.40	1.50	17.75	23.25	28.25
Q50/S	28.40	4.00	13.50	15.00	15.25
Q74-paper/C	46.90	100.00	47.50	47.75	49.25
Q74-paper/S	46.90	100.00	100.00	52.36†	53.46†
Q74-refill/C	59.90	64.00	64.00	64.00	64.00
SHTCVD191_A/C	80.60	7.79	81.86	80.45	85.71
SHTCVD191_A/S	80.60	25.50	81.25	76.50	79.00
SHTFLU12M_A/C	48.60	16.00	50.00	47.25	49.25
SHTFLU12M_A/S	48.60	64.00	56.50	48.75	48.25
V242029/C	2.60	37.00	2.25	2.75	3.00
V242029/S	2.60	31.94	8.65	1.51	2.27
V242030/C	22.60	1.75	16.75	16.75	16.25
V242030/S	22.60	29.40	29.40	29.40	29.40
V242031/C	16.00	95.25	37.25	11.42	15.74
V242034/C	11.40	97.75	7.25	13.25	17.50
V242034/S	11.40	87.75	49.25	6.25	11.00
V242035/C	31.40	75.25	29.00	29.00	32.25
V242035/S	31.40	43.75	38.00	37.25	37.00
V242036/C	9.20	47.25	1.75	1.25	1.50
V242036/S	9.20	45.25	12.25	10.75	10.75

Continued on next page

Table 13: Cell-level results for Qwen 3.5 9B (continued).

Cell	Human	B	S	M	E
WHO-blood/C	2.82	77.08	1.76	1.51	1.76
WHO-blood/S	2.82	96.50	0.00	0.50	0.25
v19/C	58.80	2.01	53.75	59.75	61.25
v19/S	58.80	0.50	12.25	18.00	17.79
v52/C	89.80	99.50	95.75	94.50	94.25
v52/S	89.80	90.25	90.25	90.25	90.25
v53/C	82.80	100.00	100.00	100.00	100.00
v53/S	82.80	100.00	96.75	83.50	73.50
v54/C	11.20	6.50	6.50	11.25	11.75
v55/C	30.40	0.50	18.75	32.50	33.25
v56/C	30.00	99.00	23.69	26.74	32.35
v57/C	4.60	12.25	3.25	3.25	3.50
val_turnout24/C	79.40	95.75	85.50	85.25	87.50
val_turnout24/S	79.40	98.50	86.35	85.94	86.82

Table 14: Cell-level conditional SRB rates and human references (%) for Qwen 3.5 27B. B/S/M/E denote baseline, single-layer steering, multi-layer steering, and item-excluded steering. Cell suffixes C/S denote common/specialized personas; † indicates a test uncodable response rate above 25%.

Cell	Human	B	S	M	E
FLUSHOT7/C	51.00	73.75	48.75	53.25	53.25
FLUSHOT7/S	51.00	64.75	45.00	52.50	50.50
FRAUD2/C	7.60	100.00	74.50	7.00	9.00
FRAUD2/S	7.60	100.00	56.25	4.75	5.75
LAST30_PT/C	11.20	32.75	17.00	6.75	7.25
LAST30_PT/S	11.20	7.25	7.25	7.25	7.75
PES18/C	53.40	93.75	76.88	47.61	48.74
PES18/S	53.40	70.25	50.82	49.19	50.27
PES6/C	58.80	100.00	100.00	54.00	59.50
PES6/S	58.80	68.25	58.75	58.50	58.75
PES7/C	21.00	98.00	83.25	24.85	31.69
PES7/S	21.00	56.25	23.00	19.75	20.25
Q40a/C	99.20	100.00	100.00	100.00	100.00
Q40a/S	99.20	99.00	99.00	99.00	99.25
Q40b/S	92.70	82.96	89.29	89.34	88.61
Q40d/C	88.60	100.00	100.00	100.00	100.00
Q40d/S	88.60	87.75	87.75	87.75	87.25
Q50/C	28.40	79.50	39.75	27.25	21.75
Q50/S	28.40	22.25	23.50	22.25	22.00
Q74-paper/C	46.90	100.00	100.00	48.25	40.96
Q74-paper/S	46.90	98.97	93.44	29.15†	31.36†
Q74-refill/C	59.90	96.00	61.50	56.00	53.25
Q74-refill/S	59.90	92.50	65.10	60.78	61.04
Q74-bag/C	83.40	100.00	100.00	100.00	100.00
Q74-bag/S	83.40	95.25	86.43	88.50	88.94
SHTCVD191_A/C	80.60	63.50	73.75	75.50	75.00
SHTCVD191_A/S	80.60	75.75	85.00	85.50	86.00
SHTFLU12M_A/C	48.60	33.00	51.25	53.25	54.25
SHTFLU12M_A/S	48.60	57.25	57.25	57.25	58.00
V242029/C	2.60	91.50	47.50	0.00	0.00
V242029/S	2.60	62.57	30.38	0.00	0.00
V242030/C	22.60	99.50	69.25	18.00	22.00

Continued on next page

Table 14: Cell-level results for Qwen 3.5 27B (continued).

Cell	Human	B	S	M	E
V242030/S	22.60	99.25	82.58	21.55	23.93
V242031/C	16.00	98.50	79.00	12.50	11.25
V242031/S	16.00	83.25	58.75	17.25	15.75
V242034/C	11.40	100.00	100.00	6.75	11.75
V242034/S	11.40	99.25	89.32	12.31	19.87
V242035/C	31.40	99.00	95.00	25.00	26.01
V242035/S	31.40	80.25	60.86	29.77	30.03
V242036/C	9.20	99.50	89.75	7.75	12.75
V242036/S	9.20	90.75	71.25	4.00	10.25
V242037/C	35.00	100.00	100.00	98.75	97.75
V242037/S	35.00	99.25	98.70	97.47	95.96
WHO-blood/C	2.82	100.00	10.25	0.00	0.00
WHO-blood/S	2.82	20.00	20.00	0.00	0.00
v19/C	58.80	100.00	52.00	53.00	52.50
v19/S	58.80	53.75	53.75	53.75	54.00
v52/C	89.80	100.00	100.00	95.25	97.00
v52/S	89.80	91.50	91.50	91.50	91.75
v53/C	82.80	100.00	100.00	100.00	100.00
v53/S	82.80	94.50	92.00	82.50	82.25
v54/C	11.20	43.50	19.00	19.00	18.25
v54/S	11.20	57.00	14.00	10.25	9.50
v55/C	30.40	99.25	32.50	31.75	30.50
v55/S	30.40	83.97	53.09	0.00	0.00
v56/C	30.00	98.50	75.50	29.50	31.75
v56/S	30.00	64.50	49.75	36.50	36.50
v57/C	4.60	97.50	43.00	5.75	7.75
v57/S	4.60	62.75	36.75	0.00	0.00
val_turnout24/C	79.40	100.00	98.75	76.00	77.25
val_turnout24/S	79.40	99.75	99.75	82.99	83.50

Table 15: Cell-level conditional SRB rates and human references (%) for Llama 3.1 8B. B/S/M/E denote baseline, single-layer steering, multi-layer steering, and item-excluded steering. Cell suffixes C/S denote common/specialized personas; † indicates a test uncodable response rate above 25%.

Cell	Human	B	S	M	E
FLUSHOT7/C	51.00	100.00	58.80†	58.51	60.51
FLUSHOT7/S	51.00	96.25	40.50	51.75	56.75
FRAUD2/C	7.60	100.00	100.00	14.55	27.64
FRAUD2/S	7.60	100.00	100.00	7.65	11.34
LAST30_PT/C	11.20	65.00	6.50	11.50	7.39
LAST30_PT/S	11.20	29.00	23.00	9.75	1.25
PES18/C	53.40	77.89	52.64	53.25	54.25
PES18/S	53.40	48.91	48.91	48.91	48.91
PES6/C	58.80	100.00	100.00	100.00	100.00
PES7/C	21.00	100.00	91.52†	35.83	39.44
Q40a/C	99.20	100.00	100.00	100.00	100.00
Q40a/S	99.20	98.94	98.94	98.94	98.94
Q40b/S	92.70	99.73	97.95	97.24	97.24
Q40d/C	88.60	100.00	100.00	100.00	100.00
Q40d/S	88.60	95.25	95.25	95.25	95.25
Q40e/C	62.90	100.00	100.00	100.00	100.00
Q50/C	28.40	98.00	34.50	28.25	26.25

Continued on next page

Table 15: Cell-level results for Llama 3.1 8B (continued).

Cell	Human	B	S	M	E
Q50/S	28.40	88.25	19.50	20.75	20.75
Q74-bag/C	83.40	100.00	100.00	90.19	92.02
Q74-bag/S	83.40	99.50	82.50	84.50	84.00
SHTCVD191_A/C	80.60	97.75	83.21	71.18	74.00
SHTCVD191_A/S	80.60	91.50	83.29	77.89	80.65
SHTFLU12M_A/C	48.60	87.00	43.50	53.25	53.25
SHTFLU12M_A/S	48.60	92.00	55.50	47.75	49.50
V242029/C	2.60	100.00	7.49	0.00	0.00
V242029/S	2.60	100.00	72.38	6.32	7.28
V242030/C	22.60	100.00	18.43	18.43	22.36
V242030/S	22.60	94.75	26.99	26.99	28.21
V242031/C	16.00	99.75	1.37†	2.89†	2.10†
V242031/S	16.00	91.21	45.49†	17.54†	21.11†
V242034/C	11.40	100.00	76.58	19.06	33.51
V242034/S	11.40	99.75	70.87	16.00	31.75
V242035/C	31.40	100.00	100.00	84.71	95.25
V242035/S	31.40	99.43	99.43	99.43	99.43
V242036/C	9.20	100.00	100.00	4.15	9.63
V242036/S	9.20	100.00	100.00	8.54	24.50
V242037/C	35.00	100.00	100.00	13.53	9.33
V242037/S	35.00	100.00	100.00	80.13	73.58
v19/C	58.80	79.25	61.50	68.75	65.50
v19/S	58.80	0.00	63.25	59.50	67.25
v52/C	89.80	100.00	100.00	85.40	84.98
v52/S	89.80	89.25	89.25	89.25	89.25
v53/C	82.80	100.00	86.82	84.08	85.15
v53/S	82.80	100.00	89.22	78.35	75.58
v54/S	11.20	0.00	8.75	11.29	27.60†
v55/C	30.40	89.72	33.33	25.13	25.39
v55/S	30.40	52.00	38.50	31.66	27.96
v56/C	30.00	98.75	87.86	27.96	29.47
v56/S	30.00	81.75	36.91†	26.32	25.69
v57/C	4.60	98.50	0.00	1.52	1.27
v57/S	4.60	95.00	0.00	1.50	1.75
val_turnout24/C	79.40	100.00	88.99	84.65	84.91
val_turnout24/S	79.40	100.00	82.49	85.47	86.08

Table 16: Cell-level conditional SRB rates and human references (%) for OLMo 3 7B. B/S/M/E denote baseline, single-layer steering, multi-layer steering, and item-excluded steering. Cell suffixes C/S denote common/specialized personas; † indicates a test uncodable response rate above 25%.

Cell	Human	B	S	M	E
FLUSHOT7/C	51.00	96.24	96.24	87.32†	86.43†
FLUSHOT7/S	51.00	97.22	95.24	94.57	94.82
FRAUD2/C	7.60	100.00	100.00	100.00	100.00
FRAUD2/S	7.60	100.00	100.00	100.00	100.00
LAST30_PT/C	11.20	76.92	13.08	10.38	10.36
LAST30_PT/S	11.20	92.78	46.37†	10.47	11.04
PES18/C	53.40	100.00	100.00	100.00	100.00
PES18/S	53.40	100.00	100.00	94.20	94.68
PES6/C	58.80	100.00	100.00	100.00	100.00
PES7/C	21.00	100.00	100.00	100.00	100.00

Continued on next page

Table 16: Cell-level results for OLMo 3 7B (continued).

Cell	Human	B	S	M	E
Q40a/C	99.20	100.00	100.00	100.00	100.00
Q40a/S	99.20	99.49	100.00	99.49	99.49
Q40d/S	88.60	100.00	100.00	100.00	100.00
Q50/C	28.40	100.00	100.00	26.33	78.77
Q50/S	28.40	100.00	100.00	0.00	0.00†
Q74-refill/C	59.90	100.00	100.00	100.00	100.00
SHTCVD191_A/C	80.60	99.25	84.00	83.50	85.00
SHTCVD191_A/S	80.60	100.00	98.50	78.77	82.44
V242029/C	2.60	89.27	67.28	14.44	17.86
V242034/C	11.40	97.99	86.65	7.06	24.82†
V242034/S	11.40	87.68	81.25	14.58	34.52
V242035/C	31.40	100.00	100.00	100.00	100.00
V242035/S	31.40	100.00	100.00	100.00	100.00
V242036/S	9.20	89.59	84.33	0.00	0.00
WHO-blood/C	2.82	100.00	100.00	100.00	100.00
v19/C	58.80	100.00	100.00	59.82	81.00
v19/S	58.80	100.00	100.00	60.83	83.82
v52/C	89.80	100.00	100.00	100.00	100.00
v53/C	82.80	71.11	86.40	80.37	73.01
v53/S	82.80	96.25	95.00	82.97	23.74
v54/C	11.20	100.00	100.00	10.72	12.47
v54/S	11.20	89.75	84.50	51.22	52.35
v57/C	4.60	34.29	5.96†	3.16	3.10
v57/S	4.60	99.49	95.05	83.73	95.55
val_turnout24/C	79.40	100.00	100.00	100.00	100.00
val_turnout24/S	79.40	100.00	100.00	100.00	100.00

G.2 Single-Layer versus Multi-Layer Steering

The single layer is selected as the layer with the highest separation score within the fixed bundle: layer 22 for Qwen 9B, 50 for Qwen 27B, 17 for Llama, and 31 for OLMo. Both methods select coefficients on the coefficient-selection set and apply them to the same test respondents. The 25% rows in Table 17 represent the main analysis; the other rows assess sensitivity to the permitted uncodable response rate. The multi-layer intervention has a smaller mean gap than the single-layer intervention for all four models. However, because comparison sets depend on method-specific validity, these means should not be directly subtracted from the baseline–multi-layer means in the main text.

G.3 Steering with Item-Excluded Directions

We exclude both persona conditions of the target item from direction estimation while retaining the original bundle and coefficients. Effect retention is the proportion of cells showing an SRB change under a nonzero intervention with the full direction that retain the same sign of change relative to baseline after exclusion. Improvement retention is the proportion of nonzero-intervention cells whose baseline gap is reduced by the full direction that retain a gap reduction after exclusion. These metrics have different denominators and evaluation sets, and exclude cells with all-zero coefficients. Numerical ties are defined as unrounded differences of at most 10^{-12}.

Among the 11 all-zero-coefficient behavioral cells for Qwen 27B, seven show differences of 1–3 SRB responses out of 400 respondents across separate runs. We do not interpret these differences as effects of replacing the direction. Additional validity failures and deterioration relative to baseline are reported in Table 19.

G.4 Validity Sensitivity and Failure Cases

For the single-layer versus multi-layer comparison, coefficients selected on the coefficient-selection set at each cutoff are applied unchanged to the test set. In contrast, the item-exclusion analysis retains the coefficients selected at the 25% criterion and their test outputs, varying only the evaluation cutoff. In the former, both coefficient selection and the valid set change; in the latter, only the valid set changes. These sensitivity analyses should not be interpreted as equivalent reoptimization results. Tables 17 and 18 report these two sensitivity analyses. Below, S, M, and E denote the single-layer intervention, multi-layer intervention, and item-excluded direction, respectively, and B denotes the baseline. The mean gain of the multi-layer intervention over the single-layer intervention is positive at every cutoff. However, because the composition of valid cells varies, differences in means across cutoffs should not be interpreted as improvements in the intervention itself. Cases with large residual errors despite no uncodable responses, such as fraud reporting for OLMo, should be distinguished from cases that lose validity because uncodable responses increase after item exclusion. Retaining the sign of behavioral change, achieving a small gap from the human reference, and producing codable outputs are distinct evaluation criteria.

Table 17: Single-layer versus multi-layer comparisons using coefficients selected at each cutoff. Valid counts are cells with selected settings that pass test validity at the corresponding cutoff (M/S); gaps and gains are computed over the same n cells. Cutoffs are in % and gaps in pp.

Model	τ	Valid M/S	n	S gap	M gap	M gain
Qwen 3.5 9B	10	47/47	47	7.55	4.94	2.61
Qwen 3.5 9B	25	47/48	47	7.55	4.27	3.28
Qwen 3.5 9B	30	48/48	48	8.50	4.29	4.21
Qwen 3.5 9B	50	48/48	48	8.50	4.29	4.21
Qwen 3.5 27B	10	59/61	59	24.45	6.56	17.89
Qwen 3.5 27B	25	60/61	60	25.03	6.24	18.79
Qwen 3.5 27B	30	61/61	61	25.38	6.19	19.19
Qwen 3.5 27B	50	61/61	61	25.38	6.19	19.19
Llama 3.1 8B	10	51/44	44	23.91	13.17	10.74
Llama 3.1 8B	25	51/48	48	23.90	8.70	15.20
Llama 3.1 8B	30	51/50	48	20.55	7.67	12.88
Llama 3.1 8B	50	53/53	53	20.33	7.46	12.87
OLMo 3 7B	10	33/32	30	44.81	32.04	12.77
OLMo 3 7B	25	35/34	33	47.98	29.22	18.76
OLMo 3 7B	30	35/36	35	45.54	27.53	18.01
OLMo 3 7B	50	35/36	35	45.54	26.28	19.27

Table 18: Item-exclusion sensitivity with coefficients and outputs fixed at the 25% criterion and only the evaluation cutoff varied. Gaps are means (pp) over the n cells jointly valid under B, M, and E.

Model	τ	E valid/total	n	M gap	E gap	Effect retention	Improvement retention
Qwen 3.5 9B	10	46/48	46	4.25	4.46	40/40	40/40
Qwen 3.5 9B	25	47/48	47	4.27	4.45	41/41	41/41
Qwen 3.5 9B	30	47/48	47	4.27	4.45	41/41	41/41
Qwen 3.5 9B	50	48/48	48	4.29	4.50	42/42	42/42
Qwen 3.5 27B	10	56/61	56	6.03	5.92	45/45	44/44
Qwen 3.5 27B	25	60/61	60	6.24	6.45	49/49	48/48
Qwen 3.5 27B	30	61/61	61	6.43	6.60	50/50	49/49
Qwen 3.5 27B	50	61/61	61	6.43	6.60	50/50	49/49
Llama 3.1 8B	10	40/53	40	6.76	8.73	32/32	32/32
Llama 3.1 8B	25	50/53	50	8.87	10.76	41/41	41/41
Llama 3.1 8B	30	51/53	50	8.87	10.76	41/41	41/41
Llama 3.1 8B	50	52/53	52	8.81	10.71	43/43	43/43
OLMo 3 7B	10	27/36	22	29.89	33.19	9/9	8/9
OLMo 3 7B	25	33/36	33	28.29	34.28	18/18	17/18
OLMo 3 7B	30	34/36	33	28.29	34.28	18/18	17/18
OLMo 3 7B	50	35/36	35	27.84	33.72	20/20	19/20

Table 19: Additional validity failures after item exclusion or deterioration relative to baseline. Gaps are in pp and E Uncodable in %. C/S denotes common/specialized personas.

Model	Cell	B gap	M gap	E gap	E Uncodable	Type
Qwen 27B	Q40d/S	0.85	0.85	1.35	0.00	Deterioration
Qwen 27B	Q50/S	6.15	6.15	6.40	0.00	Deterioration
Qwen 27B	SHTCVD191_A/S	4.85	4.90	5.40	0.00	Deterioration
Qwen 27B	SHTFLU12M_A/S	8.65	8.65	9.40	0.00	Deterioration
Qwen 27B	v52/S	1.70	1.70	1.95	0.00	Deterioration
Llama	v54/S	11.20	0.09	16.40	52.00	Validity failure
OLMo	Q50/S	71.60	28.40	28.40	78.25	Validity failure
OLMo	V242034/C	86.59	4.34	13.42	30.50	Validity failure
OLMo	v53/S	13.45	0.17	59.06	1.00	Deterioration

G.5 Comparison with a Supervised Fine-Tuning Baseline

We evaluate QLoRA-SFT baselines (Dettmers et al., 2023) for Qwen 3.5 9B and Llama 3.1 8B trained on 24 items with individual-level binary labels under two persona conditions (48 cells). The training set retains the intention item `v26`; human-reference comparisons exclude it and use the remaining 23 behavioral items (46 cells). For each item, the original 500 respondents in the coefficient-selection set are divided into 400 training and 100 development respondents, and evaluation uses the same 400 test respondents as steering. Training uses the same persona-conditioned natural-language completion task, with short target completions corresponding to each individual's observed binary behavioral label. These targets verbalize behavioral labels rather than observed free-text answers. Evaluation uses the same greedy decoding and response-coding rules.

Hyperparameters are selected on the development set based on output validity and mean absolute gap from the human reference SRB rate. Both models use a learning rate of 4×10^{-4}, three training epochs, an effective batch size of 16, and LoRA adapters (Hu et al., 2022) with rank 16, scaling parameter 32, and dropout 0.05. We target all linear modules using 4-bit NF4 with double quantization and bfloat16 computation, and apply the training loss only to completion tokens. The maximum training sequence length is 512 tokens, generation is limited to 18 new tokens, and the random seed is 42. We use linear learning-rate decay with zero warmup and weight decay.

Table 20 compares the same behavioral SFT-eligible cells with a steering setting selected under the 25% uncodable-response criterion and valid test outputs under all three methods. For Qwen, settings are available for 36 of the 46 behavioral cells, all of which remain valid. For Llama, settings are available for 43 cells, but two exceed the test validity threshold, leaving 41 comparison cells. SFT is valid in all 46 behavioral cells for both models, with 0/18,400 uncodable responses for Qwen and 5/18,400 for Llama. Table entries report uncodable rates recomputed over matched cells.

Table 20: Original-model (B), SFT, and multi-layer steering (M) results on matched valid cells. Gap is the mean absolute difference from the human reference (pp), and Uncodable is the mean uncodable-response rate (%). Extreme cells have conditional SRB rates of exactly zero or one; counts and percentages are reported. All metrics within each model use the same n cells.

Model	Method	n	Gap	Uncodable	Extreme cells (%)
Qwen 3.5 9B	B	36	35.65	0.26	2 (5.6)
Qwen 3.5 9B	SFT	36	11.77	0.00	10 (27.8)
Qwen 3.5 9B	M	36	4.39	0.38	1 (2.8)
Llama 3.1 8B	B	41	52.24	1.46	21 (51.2)
Llama 3.1 8B	SFT	41	11.90	0.02	9 (22.0)
Llama 3.1 8B	M	41	8.96	5.45	2 (4.9)

SFT reduces the mean gap relative to the original model, but steering achieves a smaller gap than SFT on the same cells. All extreme SFT cells have conditional SRB rates of zero, indicating that exclusive generation of one behavioral class within some persona–item cells helps characterize the residual gaps. However, an extreme rate does not necessarily imply a large gap: it can be close to a human reference near an endpoint. We therefore treat extreme-cell frequency as a supplementary descriptive statistic rather than an independent failure criterion. The smaller gaps under steering should also be considered alongside its higher uncodable-response rates and the validity exclusions described above.

H Extension to European Populations

This appendix presents a supplementary evaluation extending the main U.S. analysis to respondent data from Germany, Spain, and France. We examine SRB over-selection in persona-conditioned behavioral completion and response-rate adjustment toward human references using the same evaluation framework, but re-estimate directions and intervention coefficients from country-specific data. Thus, this analysis evaluates country-specific re-estimation and coefficient reselection with the layer bundles selected in the United States held fixed, rather than direct transfer of U.S. vectors or coeffi-

cients to other populations. We report only Qwen 3.5 9B and Llama 3.1 8B, for which experiments are complete.

H.1 Data and Experimental Protocol

We follow the behavioral completion task, response coding, definitions of conditional SRB rate and absolute gap, and coefficient-selection and test evaluation procedure in Sections 3.2–3.3. The European extension uses ESS Round 11 (European Social Survey European Research Infrastructure (ESS ERIC), 2026), ISSP Environment IV, and Social Networks and Social Resources data (ISSP Research Group, 2023; 2019) from Germany, Spain, and France. The final items are ESS `donprty`, `sgnptit`, `pbldmna`, `bctprd`, `pstplonl`, and `volunfp`, and ISSP `v19`, `v26`, and `v52`–`v57`, yielding 14 items and 28 common/specialized persona cells per country for direction estimation. Human-reference comparisons use 13 behavioral items (26 candidate cells per country), excluding the intention item `v26`. ESS `contplt` is excluded because of structural ambiguity in the abbreviated completion sentence, and `badge` is excluded under the item selection criteria. Human reference values are weighted SRB rates in the country- and source-specific pools of respondents eligible for complete personas. We use the product of `pspwght` and `pweight` for ESS and `WEIGHT` for ISSP.

ESS common personas include household income deciles. Unlike their U.S. counterparts, the European ISSP common personas omit income to obtain complete-response samples across countries. Specialized personas include political interest, ideology, trust, and social values for ESS; environmental perceptions and attitudes for ISSP Environment; and social contact, isolation, and trust for ISSP Social. Neither the target behavior nor closely overlapping past behaviors are included. Inputs and outputs are in English, and respondents are not restricted by survey interview language. Responses are coded using a rule-based parser with item-specific rules for the ESS and ISSP items. Item sentences preserve the behavioral targets and directions of the original data while simplifying recall-period and frequency conditions; correspondence with human references is approximate.

For each country and item, we use 500 respondents in the coefficient-selection set and 400 test respondents who do not overlap within the same item. Experiments are complete for Qwen 3.5 9B and Llama 3.1 8B. We retain the U.S.-selected layer bundles, $\{21, 22, 23\}$ for Qwen and $\{15, 17, 18\}$ for Llama, but do not reuse U.S. vectors or coefficients. For each country and model, we independently estimate shared directions from valid baseline cells in the coefficient-selection set and select each cell's coefficients using the corresponding country's human reference value. Validity requires an uncodable response rate of at most 25%, and selected directions and coefficients are neither reestimated nor reselected on the test set. The reference-based test evaluation sets contain 22, 20, and 22 selected cells for Qwen and 24, 23, and 24 for Llama in Germany, Spain, and France, respectively: 135 of the 156 behavioral candidate cells. The results therefore assess the application of country-specific coefficient reselection to new respondents, rather than direct cross-country transfer of U.S. settings.

H.2 ESS Completion Sentences and Coding Rules

Following the presentation in Appendices A.3 and B.1, Table 21 lists the exact completion sentences for the six ESS items. All three countries use the same English sentences and the system/user instruction format in Appendix A.3; only the respondent profile and item sentence change. The model generates the phrase replacing `[BLANK]`, without seeing a human reference value or response label. The ISSP items reuse the sentences in Table 6.

For human references, ESS responses coded 1 (Yes) and 2 (No) are mapped to the affirmative and negative behavioral classes, respectively; other codes are excluded from the item-specific denominator. The affirmative proportion is weighted using the ESS weights specified above. Generated completions are classified separately by the ESS-specific rules in the parser. Table 22 summarizes representative accepted expressions in the context of each fixed sentence. For donation/participation, petition signing, demonstration participation, online posting/sharing, and volunteering, direct negation is checked before affirmative action expressions. Product boycotting has the opposite purchase polarity: avoidance or explicit nonpurchase/nonuse is SRB, whereas purchase or use is non-SRB.

The parser removes protocol tokens and extracts the completion when an output exactly reproduces the sentence frame. Empty or malformed outputs, repeated tokens, multiple substantive lines, per-

Table 21: English completion sentences for the ESS items used in the European extension. Sentences are reproduced as supplied to the models.

Item	Completion sentence
donprty	I [BLANK] money to a political party or pressure group, or participated in one.
sgnptit	I [BLANK] a petition.
pbldmna	I [BLANK] part in a public demonstration.
bctprd	I [BLANK] certain products.
pstplonl	I [BLANK] or shared something about politics online.
volunfp	I [BLANK] for a not-for-profit or charitable organisation.

Table 22: Representative ESS completion-coding rules. Examples are interpreted with the fixed sentences in Table 21; they are not context-free keywords. Uncodable responses are excluded from the conditional SRB denominator.

Item	SRB examples	non-SRB examples
donprty	donate, gave, contribute	did not donate, never give
sgnptit	signed, add my name	did not sign, never sign
pbldmna	took, take	did not take, never take
bctprd	boycott, avoid, do not buy	buy, purchase, use
pstplonl	posted, shared	did not post, never share
volunfp	volunteered, work unpaid	did not volunteer, never help

sona copying, and outputs without a recognized target-action expression are uncodable. Uncertainty or intention expressions such as "might" and "would", and attitude expressions such as "support" and "believe", are also uncodable rather than non-SRB. Frame-specific exceptions recognize bare negatives such as "never" or "have not" for the five non-boycott items. For donprty, pbldmna, and pstplonl, specified frequency completions such as "occasionally", "rarely", and "regularly" are coded as SRB because they indicate some occurrence of the action. These exceptions do not extend to arbitrary incomplete phrases or other items. The same coding rules apply across countries, persona conditions, and baseline/steering outputs; the validity criterion remains an uncodable-response rate of at most 25%.

H.3 Baseline SRB Responses

Table 23 reports signed gaps among baseline-valid cells selected for actual test evaluation. The direction of the discrepancy is assessed using the signed gap, and common–specialized comparisons are restricted to paired items with valid baselines under both conditions. Persona-specific means therefore have different denominators from means over all selected cells. These results should not be generalized to response tendencies across all 26 behavioral cells, including conditions excluded during coefficient selection.

Table 23: Test baseline signed gaps in selected cells. n is the number of baseline-valid cells, and Above counts cells exceeding the human reference value. Common and Specialized report mean signed gaps over items valid under both persona conditions. Gaps are in pp.

Model	Country	n	Signed gap	Above	Common	Specialized
Qwen 3.5 9B	Germany	22	25.67	16/22	28.04	21.02
Llama 3.1 8B	Germany	24	28.59	18/24	33.29	25.83
Qwen 3.5 9B	Spain	19	25.25	14/19	31.27	16.31
Llama 3.1 8B	Spain	23	31.29	17/23	38.15	26.16
Qwen 3.5 9B	France	22	23.80	16/22	27.60	19.53
Llama 3.1 8B	France	24	26.25	17/24	31.17	22.22

The mean signed gap is positive for every country–model combination, and most cells exceed their human reference values. On items valid under both persona conditions, specialized personas yield a smaller mean signed gap than common personas, but the mean signed gap remains positive. These results are consistent with the main finding that behavior-related information can change responses without ensuring agreement between SRB rates and human reference values. However, a reduction in the mean signed gap does not imply that every item's absolute gap decreases. The effects of additional information should therefore be interpreted separately in terms of average direction and item-level error.

H.4 Held-Out Steering Results

Table 24 compares mean absolute gaps and uncodable response rates over the same cells valid under both baseline and steering. Validity and target attainment are reported using all selected cells in each group as the denominator; target attainment requires both a steering uncodable response rate of at most 25% and an absolute gap of at most 7 pp. All means weight cells equally. Errors on the coefficient-selection and test sets are also computed over the same comparison set.

Table 24: Held-out steering performance in Europe. B denotes baseline, S denotes steering, and Sel. S denotes steering on the coefficient-selection set. The cell count n, test gaps (pp), and test uncodable response rates (%) refer to the same cells valid under both test conditions. Gaps on the coefficient-selection set use these same cells.

Model	Country	n	B gap	S gap	Sel. S gap	B Uncodable	S Uncodable
Qwen 3.5 9B	Germany	22	34.62	7.15	6.63	0.51	1.77
Llama 3.1 8B	Germany	24	38.54	8.55	8.53	1.41	5.74
Qwen 3.5 9B	Spain	19	32.94	10.18	9.21	0.84	3.07
Llama 3.1 8B	Spain	23	38.88	8.57	8.33	1.93	2.98
Qwen 3.5 9B	France	21	30.46	4.63	3.96	1.02	1.54
Llama 3.1 8B	France	23	34.99	7.36	6.87	1.39	2.08

The mean test absolute gap decreases in all six country–model combinations, and the increase in error from the coefficient-selection set to the test set is less than 1 pp on the same cells.

Steering retains test validity in 63/64 cells for Qwen and 70/71 for Llama, and meets the 7 pp target in 45/64 and 49/71 cells, respectively. However, Qwen's common recycling condition in Spain has a baseline uncodable response rate of 34.5% and is excluded from the jointly valid baseline–steering comparison. The comparison therefore includes 62 cells for Qwen and 70 for Llama, totaling 132, rather than the 133 cells valid under steering alone. Distinguishing validity and target attainment over all originally selected cells from means over the jointly valid set is necessary to identify selection effects associated with changes in validity.

These results provide supplementary evidence that interventions with country-specific coefficients can adjust aggregate SRB distributions in three European countries. However, because the same human reference values are used as targets for coefficient selection and test benchmarks, the analysis does not predict new individuals' behavior or estimate unknown country-level behavioral rates. Valid item sets differ across countries, and coefficients are selected against each country's reference values; cross-country rankings of residual error should therefore not be interpreted as cultural characteristics or inherent differences in simulation difficulty.

H.5 Detailed Results

Country–model aggregate results are reported in Tables 23 and 24. Table 25 presents validity failures and large residual errors that are not apparent from aggregate means. The gap is the absolute difference between the conditional SRB rate and the country-specific human reference value; B and S denote the test baseline and steering with fixed coefficients, respectively. The two cells with uncodable response rates exceeding 25% are not counted as valid target attainment even when their errors are small.

Table 25: Representative exceptions in the European test results. C/S following an item denotes common/specialized personas; B gap and S gap denote baseline and steering absolute gaps (pp), respectively, and S Uncodable denotes the steering uncodable response rate. Validity failure indicates a steering uncodable response rate exceeding 25%.

Model	Country	Item/condition	B gap	S gap	S Uncodable (%)	Category
Qwen 3.5 9B	France	v56 C	85.73	42.22	26.0	Validity failure
Llama 3.1 8B	France	v56 S	47.94	8.34	30.3	Validity failure
Llama 3.1 8B	Germany	v56 C	77.69	69.24	20.8	Large residual error
Llama 3.1 8B	Spain	v56 C	85.47	60.94	12.0	Large residual error
Qwen 3.5 9B	Spain	v55 S	14.66	15.49	9.3	Increased error

Donations to environmental organizations (v56) retain large errors or exceed the output validity threshold after coefficient selection across multiple countries and models. A reduction in mean error therefore does not imply success for every item. Among cells valid under both conditions, the only increase in test error occurs for Qwen's specialized environmental petition condition in Spain, by approximately 0.8 pp. Some cells whose baselines are already close to their reference values receive zero coefficients; therefore, validity alone does not imply successful nonzero steering.