HUMAN NON-CHOICE IN LLMS: INTERNAL SUPPORT, LIMITED EXPRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

LLMs are used to simulate human perspectives, yet people can have little interest in an issue or no settled opinion to report. Supplying a definite position in these circumstances can misrepresent how widely an opinion is held. Using 42 survey statements across six countries and three 7–9B models, we relate this discrepancy to response concentration and trace how internal computation shapes its correction, with explicit non-choice as an imperfect indicator and midpoint responses as a separate outcome. All three models underproduce non-choice despite different midpoint biases, yet retain internal activity capable of increasing it. Rescaling scores to human directional-answer entropy, without fitting non-choice or midpoint rates, closes 58–91% of the mean non-choice gap, yet its changes do not track the correction each question needs. Internal controls produce related changes, which we trace by holding an edit fixed while allowing or preventing later modules from responding. Later computation produces most of a state-derived correction in two models and partly counteracts a head edit in another. Components that support ordinary non-choice can themselves oppose an upstream correction. These results connect scarce non-choice to probability allocation and distinguish available internal support from its realized effect. They motivate assessing non-choice allocation across questions and how subsequent computation shapes its correction.

1 INTRODUCTION

Representing human opinions requires accounting for issues on which people have no settled view. A person may have given little thought to a topic, have little interest in it, or construct an answer only when asked (Converse, 2006; Zaller & Feldman, 1992; Krosnick, 1991). These possibilities matter when large language models (LLMs) simulate human perspectives (Argyle et al., 2023; Aher et al., 2023; Santurkar et al., 2023): a plausible position need not faithfully reflect whether a person has formed a view. When limited engagement is expressed as a definite preference, the simulation can overstate the strength and prevalence of human opinions. Aggregate agreement can also mask distorted response variation (Bisbee et al., 2024). Assessing whether a view is expressed therefore complements the fidelity of its content.

Explaining this aspect of fidelity requires distinguishing available opinion information from its expression in an answer. Opinion probes recover information from hidden states, and activation interventions alter expressed perspectives (Kim et al., 2025; Jahanparast et al., 2026; Rimsky et al., 2024). These findings leave open how internal computation contributes to scarce human-like non-choice and shapes attempts to correct it. Confidence regulation and compensatory responses suggest two relevant processes: allocating probability across answers and modifying an edit through subsequent layers (Stolfo et al., 2024; McGrath et al., 2023). We ask how these processes connect the non-choice deficit to the effect of an internal correction.

We examine 42 statements from the International Social Survey Programme (ISSP), comparing human response distributions with persona-conditioned responses from Qwen3.5-9B, Llama-3.1-8B-Instruct, and OLMo-3-7B-Instruct. *Can't choose* (CC) is our primary, imperfect indicator of opinion absence; *Neither agree nor disagree* (MID) is a separate auxiliary outcome. Both can reflect several human states, so we study their expression without assigning a unique psychological meaning to either category (Gilljam & Granberg, 1993; Sturgis et al., 2014). Six steering operators provide

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.

alternative ways to access response-related activity, with the operator selected separately for each model. Each controller fixes an operator, sites, and per-question strengths on one respondent pool; another pool tests the same questions under four answer orders.

We first find shared CC scarcity despite different MID biases, motivating an analysis of response allocation. Entropy-fitted rescaling closes 58–91% of the mean CC gap without fitting CC or MID targets, yet leaves its allocation across questions inaccurate. Internal controls access related changes. We then test how a fixed edit reaches the output: later computation produces most of a state-derived correction in Llama and OLMo and partly counteracts a head edit in Qwen. Restoring selected components further shows that ordinary CC support can oppose an upstream correction. Together, these findings connect non-choice scarcity to response allocation and model-specific downstream responses.

Our contributions follow this explanatory sequence:

- We establish shared explicit non-choice scarcity across three models and distinguish it from their different midpoint biases (Section 5.1).
- We connect this scarcity to response concentration and show that model-specific internal controls access a related, partially predictable correction (Sections 5.2–5.3).
- We trace how subsequent computation realizes or opposes corrections, and show that a component's ordinary support for CC does not determine its response to an upstream edit (Sections 5.4–5.5).

2 Related Work

2.1 Human Non-Attitudes

Humans do not necessarily hold a settled opinion on every issue. Belief-system research questions whether survey answers reflect stable preferences, while response-construction accounts explain how judgments depend on accessible considerations (Converse, 2006; Zaller & Feldman, 1992). Interest, motivation, and cognitive effort further shape whether and how an attitude is expressed (Krosnick, 1991). Attitude-strength research distinguishes an evaluation's durability and behavioral impact from attributes such as certainty, importance, and ambivalence (Luttrell & Sawicki, 2020). People can therefore report the same position while differing in its stability and influence on behavior. Opinion absence differs from uncertainty, ambivalence, or considered neutrality, and these states do not map cleanly onto response categories: some "don't know" responses conceal attitudes predictive of behavior (Gilljam & Granberg, 1993), whereas midpoint responses can convey a lack of opinion (Sturgis et al., 2014). For LLMs used to represent human perspectives, reproducing this variability is part of behavioral fidelity. Evaluating only which position a model takes can miss whether a person has formed or is willing to express a view at all, potentially overstating human commitment or engagement.

2.2 Interpreting Opinion Expression in Language Models

Survey comparisons already report rare model refusals (Santurkar et al., 2023). Interpretability research identifies attention-head representations of political perspective (Kim et al., 2025) and shows that hidden-state probes can recover human opinion distributions more accurately than model outputs (Jahanparast et al., 2026). Demographic information likewise can be encoded at layers distinct from those that causally influence responses (Sharma & Le, 2026). Fine-grained political vectors distinguish ideological dimensions; a qualitative layerwise readout shows a steered token preference changing near the output (Hu et al., 2026). Complementary methods steer behavioral tendencies through contrastive activation directions (Rimsky et al., 2024), political polarity through sparse feature subspaces (He et al., 2025), and assistant traits through persona vectors (Chen et al., 2025). Value-structure analyses use causal graphs to predict cross-value changes under prompting or sparse-autoencoder steering (Kang et al., 2025). These studies establish accessible opinion-related information and behavioral control, leaving open how computation contributes to scarce human non-choice and shapes its correction. We examine response concentration and downstream computation as a shared, partial account.

3 STUDY SETUP

3.1 SURVEY TASK

"Don't know" can reflect absent views or limited effort (Krosnick, 1991), yet can conceal attitudes predictive of behavior (Gilljam & Granberg, 1993); midpoints can express non-opinion or considered neutrality (Sturgis et al., 2014). We therefore examine CC primarily and MID separately to distinguish these forms of non-directional expression. ISSP pairs both response options with demographics for constructing personas. We select 42 statements with a **five-point Likert agreement scale plus CC**: 15 from Environment 2020, 14 from Family 2022, and 13 from Health 2021. These domains vary the subject matter while preserving the response format and available persona fields. We require a neither-agree-nor-disagree midpoint, excluding middle categories denoting adequacy or equality. Responses 1–5 denote strongly agree, agree, MID, disagree, and strongly disagree; $D = \{1, 2, 4, 5\}$ denotes directional answers. No-answer, missing, and invalid codes are excluded. Appendix A.5 lists all items; Figure 4 shows an actual prompt.

To hold demographic composition constant, we balance 96 cells across **Australia, Germany, Japan, the Philippines, Russia, and the United States**: six countries, two education bands, four age bands, and two recorded sex categories. Each domain/split has three prompted people per cell (288 personas). Prompts supply exact age and education, country, sex, and valid marital-status, activity, and residence fields. Three similarly sized model families test recurrence across implementations: frozen Qwen3.5-9B, Llama-3.1-8B-Instruct, and OLMo-3-7B-Instruct in bfloat16 (BF16). Answer letters are assigned to displayed positions, mapped back to semantic categories, and their token probabilities pooled and normalized over the six legal choices. Each prompt thus yields a distribution over responses.

3.2 EVALUATION PROTOCOL

Controller *development* and held-out *test* use disjoint people and rendered backgrounds on the same questions/cells, testing item-specific control on new personas. Four answer layouts cross scale reversal with CC placement: A $(1, 2, 3, 4, 5, CC)$, B $(CC, 1, 2, 3, 4, 5)$, C $(5, 4, 3, 2, 1, CC)$, D $(CC, 5, 4, 3, 2, 1)$. Fitting uses A; frozen test under A measures new-person transfer, while B–D assess order sensitivity. For question q and order o, human vector H_q averages proportions from all eligible respondents in each split cell; model vector P_{qo} averages its three persona distributions per cell. Both weight cells equally. We score each question/order, then average orders, questions within domains, and domains equally. Test aggregates use all four layouts unless specified.

We optimize CC with binary cross-entropy (CE), a proper scoring rule for non-choice probability:

$$\mathrm{CE}_{CC}(h, p) = -h \log p - (1 - h) \log(1 - p), \quad h = H_{q,CC}, \quad p = P_{qo,CC}. \tag{1}$$

CE and Bernoulli Kullback–Leibler (KL) divergence give identical rankings for fixed targets; ideal CE is target entropy. CC/MID mean absolute errors (MAEs) average $|H_{qk} - P_{qok}|$, and total variation $\mathrm{TV}(H, P) = \frac{1}{2} \sum_k |H_k - P_k|$ checks redistribution across all six responses. Unfitted MID and TV reveal accompanying effects. Rates are percentages; probability changes and MAEs use percentage points (pp). Intervals resample questions within domains 2,000 times, preserving paired methods and orders, conditional on the evaluated people and human targets (Appendix A).

4 ANALYSIS

We test whether redistributing probabilities reduces the CC deficit, whether internal edits access that correction, and how later computation changes a fixed edit's effect.

4.1 RESPONSE REDISTRIBUTION

Let p denote a prompt's six-choice probability vector. We compare its conditional entropy over D with the corresponding human-cell entropy, measuring concentration at these two levels. Rescaling $S_\beta(p) = \mathrm{softmax}(\beta \log p)$ reduces concentration for $0 < \beta < 1$ while preserving answer rankings. One β_H per model matches mean development persona entropy to mean human-cell entropy

over D and is frozen for test. Excluding CC/MID targets tests whether redistribution reduces their discrepancies without fitting those rates.

To ask whether internal edits produce a related change, let x_0, x_a be baseline/intervened log-probabilities over D, centered within each input. A slope fitted on population-balanced development inputs under layout A predicts all six test probabilities:

$$\beta_a = \frac{\mathbb{E}[x_0^\top x_a]}{\mathbb{E}\|x_0\|^2}, \qquad \hat{p}_a = S_{\beta_a}(p_0). \tag{2}$$

The slope excludes intervened CC/MID scores and human rates, testing whether directional changes predict these outcomes. We report $R_6 = 1 - \mathrm{MSE}(\hat{p}_a, p_a)/\mathrm{MSE}(p_0, p_a)$, the reduction in six-category mean squared error (MSE) relative to assuming no change. This evaluates response prediction; Equation (1) measures human fidelity.

4.2 Internal Interventions

Teacher forcing (TF) supplies response-related candidates even when spontaneous CC is scarce. For each development prompt x, we supply all six answer continuations and read activations $h_k(x)$ at the last answer token:

$$\bar{h}_D = \tfrac{1}{4}\sum_{k\in D} h_k(x), \qquad c(x) = h_{CC}(x) - \bar{h}_D, \qquad m(x) = h_3(x) - \bar{h}_D. \tag{3}$$

Paired continuations hold question/persona fixed and contrast CC/MID with expressed agreement/disagreement. Balanced cell/domain aggregation yields candidates shared across questions and personas within each model. Extraction follows the forced answer; intervention at the last prompt token tests control of a future choice. Disjoint reference pairs and inverse-covariance cosine assess shared-reference dependence and activation geometry (Appendix C).

Controller fitting. Six operators provide different access to the candidates: Head scales attention-head activity; Neuron scales selected multilayer perceptron (MLP) coordinates; Additive shifts residual activations; PCA-additive uses principal component analysis of the response contrasts; Angular and Spherical change direction while preserving norm in exact arithmetic. Layers and heads are indexed from zero; L31H14 denotes head 14 in layer 31. All development people supply extraction; within each cell, two people fit coefficients and the third selects them. Screening retains two sets of layers (bundles) for question-specific coefficient fitting; the lower-CC-CE bundle is primary and the other tests sensitivity. We require finite, nonzero edits with bounded answer-token mass loss and freeze each model's lowest development-loss operator for test. Random directions or matched-size component sets undergo the same search (Appendix C).

State-derived interventions. We trace an opinion-state contrast through later computation. Matched development prompts specify an unformed opinion (U: "You have not yet formed an opinion on that statement") or considered, settled neutrality (N). Both can elicit MID. Before any answer, we record each attention or MLP contribution u to the residual stream at the last prompt token. Swapping updates between matched U/N runs selects one site per module type by development CC log-odds effects, without human targets. There, the unit-normalized mean U–N difference defines d; μ_U, μ_N are the two states' mean projections. On ordinary prompts containing neither description, U sets only this projection to μ_U: $u' = u + (\mu_U - d^\top u)d$; N uses μ_N. Site, direction, and projection targets remain fixed. Opposite, random, and orthogonal displacements test direction specificity (Appendix F).

4.3 Downstream Contributions

Internal edits also affect subsequent computation. Activation restoration separates these effects by reinstating reference-run states (Meng et al., 2022; Zhang & Nanda, 2024). For a fixed upstream edit, a *clamped* branch replays later updates from the same unmodified input; a *live* branch recomputes them. Cumulative release recomputes updates through successively later layers and clamps the remainder. Every branch reaches the same final readout, locating where the effect develops with output measurement fixed. Contributions are intervention-specific: representations can depend on the chosen manipulation (Grant et al., 2026; Opielka et al., 2026).

We then test whether components supporting ordinary CC also support a correction. *Native gain* scales ordinary activity; *restoration* retains the upstream edit and returns only the selected component to its same-input baseline value. For MLP activations z, output projection W, and coordinates S, gain g adds $gW_S z_S^0$; restoration removes $W_S(z_S^a - z_S^0)$, where $0/a$ denote baseline/edited runs. OLMo uses the state-derived MLP19 coordinate $d^\top u$. For CC log odds $s = \log[p(CC)/(1-p(CC))]$, we measure the component's induced contribution:

$$R(a, S) = \mathbb{E}[s_{\text{live}} - s_{\text{restored}(S)}]. \tag{4}$$

Positive/negative R means the induced component change raises/lowers CC relative to restoration. We report mean per-input log odds and mean probabilities separately; sites, doses, and controls appear in Results and Appendix G.

5 Results

5.1 A shared non-choice deficit accompanies different midpoint biases

Across 42 ISSP statements, humans select *Can't choose* (CC) at a balanced rate of 3.57%. Qwen, Llama, and OLMo assign only 1.79%, 0.68%, and 1.06% (Figure 1). Their midpoint behavior differs: human *Neither agree nor disagree* (MID) is 22.18%, compared with 11.29%, 15.06%, and 41.41% (Figure 1). OLMo thus couples scarce CC with abundant non-directional responses.

The discrepancy extends beyond the overall mean without holding uniformly in every group. Human country-level CC ranges from 0.61% to 6.24%, and each model underpredicts CC in five of six country means. Age and sex averages also underpredict it, while the 96 joint cells show greater variation (Appendix B). These patterns motivate examining how probability is allocated across answers.

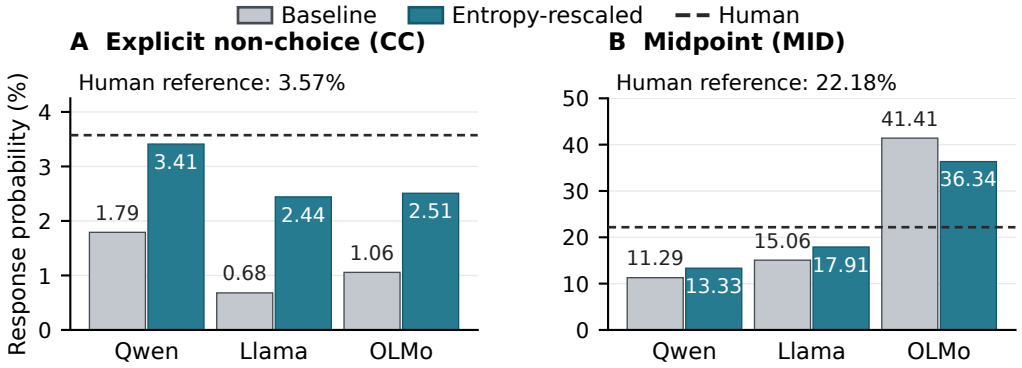


Figure 1: **Shared CC scarcity coexists with opposite MID biases.** Bars show absolute probabilities before and after entropy-fitted output rescaling; dashed lines mark the human rates. CC and MID have separately labeled vertical scales. Means cover 42 items and four test layouts. Fitting uses directional entropy alone. Both mean gaps shrink, while item-level allocation remains inaccurate (Section 5.2).

5.2 Redistributing concentrated probabilities recovers non-choice mass

The models' conditional entropy over the four directional answers is below the corresponding human-cell entropy on 36, 39, and 38 of 42 items. Across items, Pearson correlations between entropy and CC log-odds gaps are .508/.637/.231 for Qwen/Llama/OLMo. Within question–order strata, model entropy and CC log odds correlate across personas (.508/.508/.245; Appendix E). Human cells and persona-conditioned models contain different information; their entropy difference measures distributional concentration, rather than individual psychological confidence.

Reducing concentration changes the CC discrepancy even when CC and MID are excluded from fitting. Entropy-fitted rescaling closes 91/61/58% of the mean test CC gap (Figure 1; intervals and target sensitivity in Appendix Table 18). It also moves MID toward the human target in all

three models, including OLMo's decrease (Figure 1). The same redistribution thus accommodates different MID biases.

Lower average error leaves a distinct allocation problem. Its itemwise CC changes have negative Pearson correlations with the required corrections ($-.32/-.21/-.57$). Thus CC CE and MAE improve, but the change does not track how much correction each item needs. We next test whether internal controls access a related correction, using a common diagnostic of their response changes.

5.3 Internal controls access the corrective response change

The aggregated CC and MID TF contrasts remain aligned with disjoint directional references: median cosines across 32 layers are .338/.377/.313 (Appendix C). This alignment nominates overlapping response-related candidates; ordinary-prompt interventions test their function.

The development-loss criterion selects Neuron for Qwen and Llama and Head for OLMo. Under the fitted layout A, baseline-relative CC CE reductions are .0141/.0496/.0319 on development people and .0141/.0520/.0329 on held-out people (Appendix D.1). The benefit thus persists across people within the same cells and questions; averaging the four test layouts reduces it. The selected controls lower test CC CE in every model (Table 1) and domain mean (Appendix Table 8). They also lower CC and MID MAE, although OLMo Head worsens TV, chiefly in Environment and Family. Joint-cell results retain this distinction between CC improvement and full-distribution fidelity (Appendix B).

Table 1: **Available correction and its distributional costs.** Bold identifies development-selected methods and within-model minimum errors. CE is nats, MAEs pp, TV on $[0, 1]$. Human CE is entropy; MID is 22.18%. Random searches six operators/two seeds; rescaling fits directional entropy.

Model	Condition	CC (%)	CC CE ↓	CC MAE ↓	MID MAE ↓	TV ↓
Human	Target	3.57	0.1515	0.00	0.00	0.000
Qwen	Baseline	1.79	0.1663	1.96	11.25	0.303
	Neuron	3.43	0.1570	**1.43**	10.06	0.282
	Random Spherical	2.00	0.1635	1.69	10.53	0.286
	Rescaling	3.41	**0.1568**	1.46	**9.27**	**0.234**
Llama	Baseline	0.68	0.2052	2.98	9.09	0.312
	Neuron	2.80	0.1756	2.48	8.85	0.262
	Random Head	2.25	0.1751	1.94	9.13	0.271
	Rescaling	2.44	**0.1630**	**1.90**	**6.65**	**0.223**
OLMo	Baseline	1.06	0.1931	2.79	24.53	0.413
	Head	1.70	0.1753	2.23	21.56	0.422
	Random Head	1.72	0.1746	2.20	23.45	0.443
	Rescaling	2.51	**0.1656**	**2.04**	**18.48**	**0.332**

The practical alternatives separate access to correction from superiority of the extracted candidates. Primary-minus-retuned-random CC CE favors Qwen ($-.00641$); Llama/OLMo differences ($+.00057/+.00074$) have intervals including zero. Random has lower CC MAE in Llama and OLMo, while the selected controls have lower MID MAE and TV point estimates in all three models. Entropy-fitted rescaling has lower CC CE, MID MAE, and TV than the selected controls in all three models. Rescaling is an output reference; internal edits locate related computation.

The directional-only forecast tests continuity with the output-level concentration result (Equation (2)). The 72 conditions comprise three models, six operators, and four candidates per operator: two retained layer bundles and two random seeds. All 32 with $|1 - \beta_a| \geq .01$ flatten scores and improve CC CE. The threshold and accounting appear in Appendix E; this is an exploratory association among dependent conditions. For nine displayed Head/Neuron/Spherical controls, the frozen forecast reduces six-category change MSE by 7.7–45.9% and predicts MID-change signs in 27/36 order cells (Appendix E).

Different operators thus access a repeated, partially predictable redistribution beyond the fitted CC outcome. Substantial residuals remain, and uniform-mixture forecasts are competitive, so the slope serves as a shared diagnostic rather than a unique physical mechanism. The optimized controllers establish access to correction. To locate how the network processes an opinion-related change, we next hold an edit fixed and vary only the later computation.

5.4 Downstream computation realizes or opposes the redistribution

Later computation can produce or counteract the same edit's effect (Table 2). We use the state-derived edits from Section 4.2: U sets one update projection to its level under an unformed opinion; N uses considered neutrality. Neither description is present at test.

At MLP19, U improves CC CE, MID MAE, and TV in Llama/OLMo; N worsens all three. CC CE falls from same-run baselines .2051/.1931 to .1902/.1897, respectively; CC rises .27/.10 pp and MID moves toward human targets. Its CC gain closes 9.5/3.8% of the mean deficit; Qwen's U edit at MLP31 changes CC by .02 pp.

Llama U outperforms the fixed direction controls on all three losses; their displacement norms match U at both module output and residual addition on every input after BF16 rounding. OLMo shows the same loss-comparison signs with 95–96% of inputs matched, but its CC-CE advantage over a temperature fitted to U's development mean CC is unestablished (difference $-.00032$, interval $[-.00095, .00028]$).

We then keep each edit fixed and vary only the subsequent computation. For U, releasing the remaining layers after MLP19 (the suffix) produces most of the CC effect and lowers CC CE by .0143/.0031 in Llama/OLMo relative to clamping. In an exploratory TF-based follow-up, Qwen's L31H14 provides an opposing case: its following MLP buffers a beneficial reduction and a harmful increase in head activity (Table 2). H14 and H12 were selected by development contrast magnitude and sign consistency; H12 has opposite signed effects that are also buffered (Appendix Table 23).

Table 2: **Later computation changes the same edit's effect.** Mean CC log-odds changes per input from baseline. Live releases Qwen's next MLP or Llama/OLMo's suffix after MLP19.

Model	Fixed edit	Clamped	Live	Later response
Qwen	H14 $\times.75$	$+.299$	$+.135$	Opposes CC increase
Qwen	H14 $\times1.25$	$-.304$	$-.107$	Opposes CC decrease
Llama	MLP19 U	$+.020$	$+.603$	Realizes CC increase
OLMo	MLP19 U	$+.014$	$+.133$	Realizes CC increase

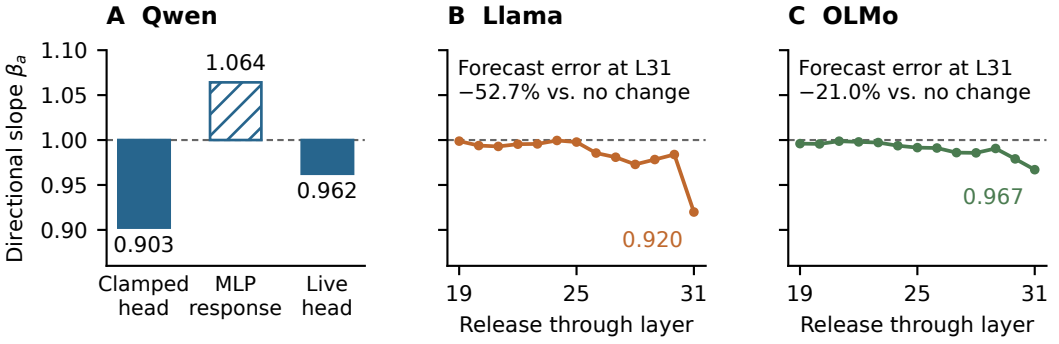


Figure 2: **The concentration diagnostic follows downstream computation.** A: Qwen's H14 $\times.75$. The MLP slope compares live with clamped outputs; both head slopes use baseline. B–C: cumulative release after U, with remaining updates clamped. Slopes β_a are development-fitted point estimates; dashed lines mark one. Final-release labels give test six-category change-MSE reductions versus no change, not fidelity gains or mediated fractions.

Directional forecasts connect these path effects to response concentration (Figure 2). Qwen's head reduction flattens scores ($\beta_a = .903$); the following MLP response sharpens them (1.064), leaving a live slope of .962 and reducing the head edit's fidelity benefit. In Llama, U's slope is near one through early release boundaries and reaches .920 after the final layer; its six-category forecast reduces change MSE by 52.7% there. OLMo reaches .967 with a smaller 21.0% reduction. OLMo's final single release step (layer 30 to 31) increases conditional MID log odds, $\log[p(MID)/\sum_{k\in D} p(k)]$, although the scaling diagnostic predicts a decrease. These residuals distinguish model-specific responses within the shared concentration pattern. On six additional Environment items, the fixed U slopes also predict CC log-odds changes in Llama/OLMo; Qwen remains poorly predicted (Appendix Table 21). Update-norm controls retain most of U's live CC effect, and final-normalization accounting separates the suffix response from denominator changes (Appendix F).

We next test whether these responses involve activity already supporting ordinary CC, comparing gain and restoration on the same components.

5.5 Ordinary CC support can counteract a correction

We hold the same components fixed across gain and restoration: Qwen's 256 and Llama's 64 TF-selected MLP31 coordinates, and OLMo's state-derived MLP19 coordinate $d^\top u$. For OLMo, this coordinate is itself the U axis, so restoration tests its response to an upstream Head correction. Without an upstream edit or state instruction, we first scale their existing activity by 1.25. Mean CC log odds rise by .186/.069/.057 (CC by .311/.052/.040 pp), and CC CE improves in all three models, every order mean, and every domain mean. Negative gain reverses the mean effects (Figure 3A). Thus scarce CC coexists with an active internal contribution that can improve its expression. MID MAE improves in Qwen/OLMo but slightly worsens in Llama.

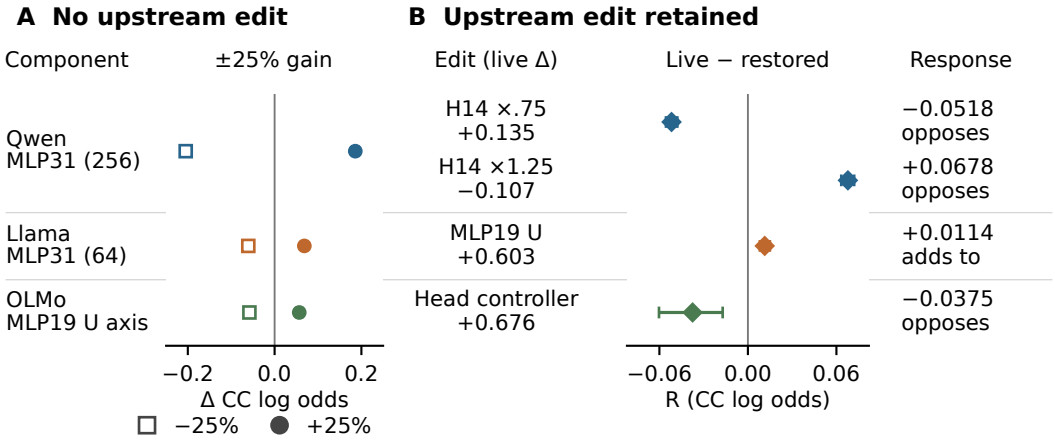


Figure 3: **Ordinary CC support need not reinforce an upstream correction.** A: ±25% gains on ordinary activity (parentheses: coordinate counts). B: retain the edit and restore the component to baseline. Edit gives the total live Δ; Response compares $R = \text{live}-\text{restored}$ with it. Values are four-order test means of per-input CC log-odds changes; doses and axis ranges differ. Bars: pointwise 95% question-bootstrap intervals. Controls and probability changes appear in Appendix G.

With upstream edits retained, we restore the selected components to their same-input baseline values. Live minus restored R measures their induced CC log-odds contribution (Equation (4)). Comparing its sign with the retained edit's total effect tests whether the component adds to or opposes that change. Qwen's coordinates buffer both head perturbations: $R = -.0518$ $[-.0555, -.0478]$ at multiplier .75 and $+.0678$ $[.0633, .0720]$ at 1.25. These changes connect MLP compensation to the extracted support. Llama's coordinates carry a small positive part of U's effect, $R = .0114$ $[.0081, .0149]$, compared with the total .603 log-odds change; its mean-probability difference is approximately zero (Table 26). OLMo's coordinate opposes its Head controller (total $+.676$), with $R = -.0375$ $[-.0603, -.0171]$; restoring it improves CC CE, MID MAE, and TV.

Gain scales baseline activity, whereas restoration removes its induced change. The controls limit uniqueness: a fixed Qwen control slightly exceeds the selected gain on CC CE. Only 7–61% of inputs pass both module-output and residual-addition norm checks jointly for the selected restoration and both control seeds (Appendix G). Ordinary CC support therefore does not determine the sign of a component's contribution under a correction.

6 Discussion

Low non-choice probability relative to humans coexists with usable internal support in all three models. Response concentration links this discrepancy to a recurring probability redistribution, while later computation determines how a fixed internal correction reaches the output. The account accommodates different effective operators and opposite MID biases. Logit scaling, confidence-guided factual abstention, and downstream compensation have precedents in other tasks (Stolfo et al., 2024; Kumaran et al., 2026; McGrath et al., 2023). Our evidence connects these processes to a human-response discrepancy: recovering mean CC mass does not recover its allocation across questions, and downstream computation can limit a correction even through components that support ordinary CC.

Intervention assessment needs to distinguish available support from realized benefit: clamping preserves more of Qwen's useful head effect but removes most of Llama/OLMo's state-derived effect. Neither an edit's immediate effect nor a component's ordinary CC contribution determines the edit's final benefit. The signed downstream response locates where a correction develops or is counteracted. Output rescaling supplies a practical alternative; internal tests locate the computation shaping specified corrections.

For human simulation, matching mean non-choice leaves its allocation unresolved. Entropy-fitted rescaling reduces average error, yet its changes do not track the correction each item needs; population slices expose further variation. A closer aggregate can therefore misrepresent which questions or populations elicit non-choice. Keeping MID separate reveals another boundary: OLMo's abundant midpoint responses coexist with scarce CC. Merging the categories would obscure this difference and the distributional costs of correction (Appendix H).

CC and MID remain partial behavioral indicators. Joint human measurements of opinion presence, familiarity, and ambivalence would connect the computational account more directly to opinion absence. The selected sites and doses establish conditional effects, with competitive controls, forecast residuals, and incomplete norm matching restricting component specificity. The next explanatory test is to predict whether later computation helps a correction at new sites or doses; failure would bound the shared account to its tested operating points. The training origin of baseline concentration remains open. The English, known-question design and three 7–9B models delimit the evidence; new questions and larger models across training stages can test whether these allocation and downstream patterns persist.

7 Conclusion

Across the tested models, limited explicit non-choice coexists with internal activity that can increase its expression. Response concentration connects this discrepancy to a partially predictable redistribution, and fixed-path tests show how later computation realizes or counteracts that change. A component's ordinary CC contribution therefore gives an incomplete account of its effect during correction. Evaluating human opinion simulation requires distinguishing non-choice from midpoint responding, overall frequency from its allocation across questions, and an internal edit from the computation that follows it. These distinctions connect response fidelity to intervention design. Future controls can address both the probability assigned to non-choice and the downstream computation that realizes it.

AI USE STATEMENT

Large language model (LLM) tools were utilized to assist with translation, polish academic phrasing, and enhance the clarity, conciseness, and terminological consistency of the manuscript. All AI-assisted revisions were thoroughly reviewed and edited by the authors to guarantee accuracy and fidelity to the original research findings.

ETHICS STATEMENT

This study analyzes previously released ISSP survey records and recruits no new participants. Persona prompts use demographic background fields without names, respondent identifiers, or human answers; reported outcomes are aggregated response distributions. Simulations that overstate engagement or preferences can distort conclusions about human populations. Separating explicit non-choice from midpoint responding and reporting population variation exposes some of these distortions. The observed categories provide partial behavioral indicators of opinion absence, with the construct limitations discussed in Section 6.

REPRODUCIBILITY STATEMENT

The anonymous supplementary archive `nonchoice-code.zip` provides analysis scripts, intervention operators, frozen controller coefficients, input schemas, and execution instructions. Survey records, model weights, activations, and prediction caches are excluded; the scripts operate on documented saved-input formats through configurable paths. Sections 3–4 define the evaluation, response probabilities, objectives, and interventions. Appendix A documents survey sources, population construction, respondent splits, scoring, and every survey item. Appendix C specifies contrast aggregation, operators, search ranges, and controller selection; Appendix D separates development, held-out-person, and answer-order results. Appendices E–G report the response forecasts, fixed-path procedures, component interventions, and numerical controls used to interpret them.

REFERENCES

Gati V Aher, Rosa I. Arriaga, and Adam Tauman Kalai. Using large language models to simulate multiple humans and replicate human subject studies. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 337–371. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/aher23a.html.

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

James Bisbee, Joshua D. Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer M. Larson. Synthetic replacements for human survey data? The perils of large language models. *Political Analysis*, 32(4):401–416, 2024. doi: 10.1017/pan.2024.5. URL https://doi.org/10.1017/pan.2024.5.

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. URL https://arxiv.org/abs/2507.21509.

Philip E Converse. The nature of belief systems in mass publics (1964). *Critical review*, 18(1-3): 1–74, 2006.

Mikael Gilljam and Donald Granberg. Should we take don't know for an answer? *Public Opinion Quarterly*, 57(3):348–357, 1993.

Satchel Grant, Simon Jerome Han, Alexa Tartaglini, and Christopher Potts. Addressing divergent representations from causal interventions on neural networks. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 11330–11356, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/133e588e1429f9f1e25b215da145580e-Paper-Conference.pdf.

Zirui He, Mingyu Jin, Bo Shen, Ali Payani, Yongfeng Zhang, and Mengnan Du. SAE-SSV: Supervised steering in sparse representation spaces for reliable control of language models. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 2207–2236, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.112. URL https://aclanthology.org/2025.emnlp-main.112/.

Jingyu Hu, Mengyue Yang, Mengnan Du, and Weiru Liu. Fine-grained interpretation of political opinions in large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(45):38570–38579, 2026. doi: 10.1609/aaai.v40i45.41199. URL https://ojs.aaai.org/index.php/AAAI/article/view/41199.

Erfan Jahanparast, Zhiqing Hong, and Serina Chang. What do large language models know about opinions? In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 146105–146128, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/ebfd0d632e950922baad6ecb64cdc407-Paper-Conference.pdf.

Yipeng Kang, Junqi Wang, Yexin Li, Mengmeng Wang, Wenming Tu, Quansen Wang, Hengli Li, Tingjun Wu, Xue Feng, Fangwei Zhong, and Zilong Zheng. Are the values of LLMs structurally aligned with humans? A causal perspective. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 23147–23161, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1188. URL https://aclanthology.org/2025.findings-acl.1188/.

Junsol Kim, James Evans, and Aaron Schein. Linear representations of political perspective emerge in large language models. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 7180–7211, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/1557cc3e1a67f985dbdb162db9963dd0-Paper-Conference.pdf.

Jon A. Krosnick. Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, 5(3):213–236, 1991. doi: 10.1002/acp.2350050305. URL https://doi.org/10.1002/acp.2350050305.

Dharshan Kumaran, Nathaniel Daw, Simon Osindero, Petar Veličković, and Viorica Patraucean. Causal evidence that language models use confidence to drive behaviour. *Nature Machine Intelligence*, 8(9):1411–1425, 2026. doi: 10.1038/s42256-026-01293-x. URL https://www.nature.com/articles/s42256-026-01293-x.

Andrew Luttrell and Vanessa Sawicki. Attitude strength: Distinguishing predictors versus defining features. *Social and Personality Psychology Compass*, 14(8):e12555, 2020. doi: 10.1111/spc3.12555. URL https://doi.org/10.1111/spc3.12555.

Thomas McGrath, Matthew Rahtz, János Kramár, Vladimir Mikulik, and Shane Legg. The hydra effect: Emergent self-repair in language model computations. *arXiv preprint arXiv:2307.15771*, 2023. URL https://arxiv.org/abs/2307.15771.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 17359–17372. Curran Associates, Inc., 2022. doi: 10.52202/068431-1262. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/6f1d43d5a82a37e89b0665b33bf3a182-Paper-Conference.pdf.

Gustaw Opielka, Hannes Rosenbusch, and Claire Stevenson. Causality $\neq$ invariance: Function and concept vectors in LLMs. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 83393–83417, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/86b3697c4eb7792c951831636bfdacd5-Paper-Conference.pdf.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering Llama 2 via contrastive activation addition. 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. 15504–15522, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.828. URL https://aclanthology.org/2024.acl-long.828/.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 29971–30004. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/santurkar23a.html.

Aarushi Sharma and Phong Le. Encoding without influence: Dissociating demographic representation from causal effect in large language models. *Transactions on Machine Learning Research*, 2026.

Alessandro Stolfo, Ben Wu, Wes Gurnee, Yonatan Belinkov, Xingyi Song, Mrinmaya Sachan, and Neel Nanda. Confidence regulation neurons in language models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 125019–125049. Curran Associates, Inc., 2024. doi: 10.52202/079017-3970. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/e21955c93dede886af1d0d362c756757-Paper-Conference.pdf.

Patrick Sturgis, Caroline Roberts, and Patten Smith. Middle alternatives revisited: How the neither/nor response acts as a way of saying "I don't know"? *Sociological Methods & Research*, 43(1):15–38, 2014. doi: 10.1177/0049124112452527. URL https://doi.org/10.1177/0049124112452527.

John Zaller and Stanley Feldman. A simple theory of the survey response: Answering questions versus revealing preferences. *American journal of political science*, pp. 579–616, 1992.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 1651–1678, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/06a52a54c8ee03cd86771136bc91eb1f-Paper-Conference.pdf.

A STUDY POPULATION

Table 3 summarizes the comparisons connecting response behavior to internal computation.

Table 3: Comparisons supporting the explanatory sequence. Each internal comparison uses the same model and input. Fixed-path effects are conditional on the retained edit, site, and dose.

Comparison	**Changed / held fixed**	**Question answered**
Output rescaling	Change scores; keep model computation fixed	Can redistribution reduce the discrepancy?
Internal control	Fit alternative operators on shared candidates; freeze for test	Is related correction internally accessible?
Clamped/live paths	Change later updates; keep upstream edit fixed	Does later computation realize or oppose it?
Component tests	Scale ordinary activity or restore its induced change; keep component fixed	How does ordinary support relate to a correction?

A.1 WHAT THE HUMAN RESPONSES MEASURE

The source benchmark contains 15 Environment 2020, 14 Family 2022, and 13 Health 2021 agreement-scale statements. Each has five ordered responses with an explicit midpoint and a separate CC option. Responses 1/2 express agreement, 3 is MID, 4/5 disagreement, and CC is separately coded as -8. No answer (-9), invalid codes, and missing responses are excluded from each item's denominator. CC may reflect uncertainty, unwillingness, ambivalence, or lack of an opinion; MID may reflect considered neutrality or a weak or unformed opinion. Neither is a validated individual label for opinion absence, and MID under an unformed-state instruction is not counted as an error by definition.

Data come from the GESIS ISSP Environment, Family, and Health modules for the stated years. The public study records are `https://search.gesis.org/research_data/ZA8793`, `https://search.gesis.org/research_data/ZA10000`, and `https://search.gesis.org/research_data/ZA8794`. Section A.5 lists every survey statement used in the study, retaining introductory stems for fragmentary Health items. The balanced study population is not a nationally representative survey estimate.

A.2 RESPONDENT SPLITS WITHIN BALANCED CELLS

Each domain has 96 cells crossing six countries (Australia, Germany, Japan, the Philippines, Russia, United States), two education bands (0–12 and 13–30 years), four age bands (18–34, 35–49, 50–64, 65–99), and two recorded sex categories. Three development and three test people are selected per cell: 288 per domain/split, or 1,728 domain-specific records overall. Personas contain exact age and education, country, recorded sex, and valid marital-status, activity, and residence fields. Names, IDs, and human answers are absent. The split uses fixed hashed identities and prevents an identical rendered background from crossing development and test. Within-split duplicates keep their multiplicity; distinct development/test backgrounds number 287/288, 286/287, and 287/285 in Environment, Family, and Health. All source questions occur in both splits, so the primary generalization is to new people on known items. The controller-development pool supplies extraction, coefficient fitting, and selection. Respondent-held-out evaluation uses different people within the same demographic cells; it does not test new demographic groups.

The human target uses all eligible humans in the corresponding split cell, not only its three prompted people. For human counts n_{qgk}, valid denominator N_{qg}, question q, cell g, order o, and semantic response k,

$$H_{qk} = \frac{1}{96} \sum_{g} \frac{n_{qgk}}{N_{qg}}, \qquad P_{qok} = \frac{1}{96} \sum_{g} \frac{1}{3} \sum_{i=1}^{3} p_k(x_{qogi}). \tag{5}$$

Raw pooled counts therefore need not have the same percentages as the balanced target. For an item-level metric f, the main aggregate is

$$\mathcal{A}[f] = \frac{1}{3}\sum_{d}\frac{1}{|Q_d|}\sum_{q\in Q_d}\frac{1}{4}\sum_{o} f(H_q, P_{qo}). \tag{6}$$

People are averaged before the item/order loss is evaluated. Order-specific losses are averaged after scoring; this differs from scoring a mixture of four order distributions. The additional Government comparison in Appendix D.3 instead scores a mixture of the four order distributions. No individual-answer accuracy is used.

A.3 Response-scoring protocol

The four layouts are A $(1, 2, 3, 4, 5, CC)$, B $(CC, 1, 2, 3, 4, 5)$, C $(5, 4, 3, 2, 1, CC)$, and D $(CC, 5, 4, 3, 2, 1)$. They cross scale reversal with CC placement rather than exhausting permutations. Displayed answer letters are mapped back to semantic categories. Validation fitting uses A; frozen test uses all four. Figure 4 shows an actual prompt with the source's midpoint wording.

Persona

You are a 72-year-old man in Germany. You have 11 years of formal education. Marital status: Married, Civil Partnership. Main activity: Retired. Place of residence: A town or a small city.

Question

How much do you agree or disagree with each of these statements? Modern science will solve our environmental problems with little change to our way of life.

Options

A. Strongly agree
B. Agree
C. Neither agree or disagree
D. Disagree
E. Strongly disagree
F. Can't choose

Response format

Answer with the letter only.
Answer:

Figure 4: **Example survey prompt in layout A.** The four boxes reproduce the input wording in order, including the original midpoint label. Side labels identify the prompt's parts and are not included in the model input. The first answer-token probabilities are read immediately after the final *Answer:*.

At the first answer position, single-token aliases for each letter are pooled by log-sum-exp, then normalized by a six-way softmax. Legal mass is the original vocabulary probability assigned to those aliases. The reported probabilities describe this conditional multiple-choice distribution, not unconstrained open-ended generation. Weights are frozen in BF16; scoring uses float32 and report aggregation float64. The three checkpoints are Qwen/Qwen3.5-9B, meta-llama/Llama-3.1-8B-Instruct, and allenai/Olmo-3-7B-Instruct, using their model-specific chat templates. Each intervention is compared with a baseline computed using the same model and inference procedure.

A.4 Evaluation details

For $h = H_{q,CC}$ and $p = P_{qo,CC}$,

$$\mathrm{CE}_{CC} = -h \log p - (1-h)\log(1-p), \tag{7}$$

$$\mathrm{KL}_{CC} = \mathrm{CE}_{CC} - [-h \log h - (1-h)\log(1-h)], \tag{8}$$

$$\mathrm{MAE}_{CC} = |h-p|, \qquad \mathrm{MAE}_{MID} = |H_{q,3} - P_{qo,3}|, \tag{9}$$

$$\mathrm{TV}(H,P) = \tfrac{1}{2}\sum_{k} |H_k - P_k|. \tag{10}$$

Logarithms are natural; probabilities are clipped away from zero where required. CC CE is the optimized population objective, with the same ranking as binary KL for fixed targets. MAE supplies an interpretable marginal error; TV checks all six responses. The low empirical CC prevalence is preserved rather than replaced by a class-balanced target.

Person-level intervals resample prompted people within cells while preserving method and order pairing. Question-level intervals resample questions within domains 2,000 times, keeping all orders and paired conditions together. The latter condition on evaluated people and empirical human targets; neither incorporates the survey's complete sampling uncertainty. Controller parameters are fixed before held-out evaluation. Mechanistic follow-ups and model-specific method reporting are exploratory and were developed after earlier results were examined. Pointwise intervals describe uncertainty within those analyses; they do not imply preregistration or adjustment for all model/method comparisons.

A.5 Complete question inventory

Tables 4–6 contain the 42 main statements: 15 Environment 2020, 14 Family 2022, and 13 Health 2021 items. Table 7 contains the six Environment 2010 items used in additional-question and fixed-path analyses, and the six Government 2006 items used in external-domain transfer. These additional cohorts do not enter the 42-item main averages. Supplied-state conditions reuse survey items rather than adding different survey questions. The inventory therefore contains 54 survey items in total.

Statements retain the English wording presented to the models, including introductory stems where needed. Country placeholders are replaced by the persona's country during inference. Each item has five agreement-scale responses plus CC; the four layouts reorder these semantic choices as defined above. Government 2006 uses the Role of Government IV module (GESIS study ZA4747).

Table 4: Exact English statements for the 15 Environment 2020 main items.

Item	Statement
Env20__v1	Private enterprise is the best way to solve [COUNTRY'S] economic problems.
Env20__v2	It is the responsibility of the government to reduce the differences in income between people with high incomes and those with low incomes.
Env20__v11	Modern science will solve our environmental problems with little change to our way of life.
Env20__v12	We worry too much about the future of the environment and not enough about prices and jobs today.
Env20__v13	Almost everything we do in modern life harms the environment.
Env20__v14	People worry too much about human progress harming the environment.
Env20__v15	In order to protect the environment [COUNTRY] needs economic growth.
Env20__v17	Economic growth always harms the environment.
Env20__v23	It is just too difficult for someone like me to do much about the environment.
Env20__v24	I do what is right for the environment, even when it costs more money or takes more time.
Env20__v25	There are more important things to do in life than protect the environment.
Env20__v26	There is no point in doing what I can for the environment unless others do the same.
Env20__v27	Many of the claims about environmental threats are exaggerated.
Env20__v28	I find it hard to know whether the way I live is helpful or harmful to the environment.
Env20__v35	Environmental problems have a direct effect on my everyday life.

Table 5: Exact English statements for the 14 Family 2022 main items.

Item	Statement
Fam22__v1	A working mother can establish just as warm and secure a relationship with her children as a mother who does not work.
Fam22__v2	A pre-school child is likely to suffer if his or her mother works.
Fam22__v3	All in all, family life suffers when the woman has a full-time job.
Fam22__v4	A job is all right, but what most women really want is a home and children.
Fam22__v5	Being a housewife is just as fulfilling as working for pay.
Fam22__v6	A man's job is to earn money; a woman's job is to look after the home and family.
Fam22__v11	Having children interferes too much with the freedom of parents.
Fam22__v12	Having children increases people's social standing in society.
Fam22__v13	People who want children ought to get married.
Fam22__v14	It is all right for a couple to live together without intending to get married.
Fam22__v15	A single mother can bring up her child as well as a married couple.
Fam22__v16	A single father can bring up his child as well as a married couple.
Fam22__v17	A same sex female couple can bring up a child as well as a male-female couple.
Fam22__v18	A same sex male couple can bring up a child as well as a male-female couple.

Table 6: Exact English statements for the 13 Health 2021 main items.

Item	Statement
Health21__v4	People use health care services more than necessary.
Health21__v5	The government should provide only limited health care services.
Health21__v6	In general, the health care system in [COUNTRY] is inefficient.
Health21__v8	How much do you agree or disagree with the following statements? People should have access to publicly funded health care even if they Do not hold citizenship of [COUNTRY].
Health21__v9	How much do you agree or disagree with the following statements? People should have access to publicly funded health care even if they Behave in ways that damage their health.
Health21__v10	Severe health problems may have many causes. How much do you agree or disagree with the following statements? People suffer from severe health problems Because they behaved in ways that damaged their health.
Health21__v11	Severe health problems may have many causes. How much do you agree or disagree with the following statements? People suffer from severe health problems Because of the environment they are exposed to at work or where they live.
Health21__v12	Severe health problems may have many causes. How much do you agree or disagree with the following statements? People suffer from severe health problems Because of their genes.
Health21__v13	Severe health problems may have many causes. How much do you agree or disagree with the following statements? People suffer from severe health problems Because they are poor.
Health21__v14	Alternative medicine provides better solutions for health problems than conventional medicine.
Health21__v15	All things considered, doctors can be trusted.
Health21__v16	The medical skills of doctors are not as good as they should be.
Health21__v17	Doctors care more about their earnings than about their patients.

Table 7: Additional survey items used in transfer and fixed-path analyses. These are separate from the 42 main items.

Item	Statement
Environment 2010	How much do you agree or disagree with each of these statements?
Env10__v9	We believe too often in science, and not enough in feelings and faith.
Env10__v10	Overall, modern science does more harm than good.
Env10__v18	The earth simply cannot continue to support population growth at its present rate.
Env10__v47	For environmental problems, there should be international agreements that [COUNTRY] and other countries should be made to follow.
Env10__v48	Poorer countries should be expected to make less effort than richer countries to protect the environment.
Env10__v49	Economic progress in [COUNTRY] will slow down unless we look after the environment better.
Government 2006	Please tick one box on each line to show how much you agree or disagree with each of the following statements.
Gov06__v61	People like me don't have any say about what the government does
Gov06__v62	The average citizen has considerable influence on politics
Gov06__v63	I feel that I have a pretty good understanding of the important political issues facing our country
Gov06__v64	I think most of the people are better informed about politics and government than I am
Gov06__v65	People we elect as MPs try to keep the promises they have made during the election
Gov06__v66	Most civil servants can be trusted to do what is best for the country

B POPULATION VARIATION

The following slices evaluate the frozen controllers against the corresponding human targets. They keep models and operators separate. The displayed method is Neuron for Qwen/Llama and Head for OLMo, fixed by the same development criterion used in Results, rather than selected anew in each group. For a marginal group, we average its cells before scoring each question/order and then apply equal-domain weighting. For a joint cell, we score its own human target against its three-person model average before averaging questions, orders, and domains. Consequently, mean cell CE differs from CE of the whole population. Cell estimates are descriptive and can be noisy, especially with rare CC and only three model personas per cell; no demographic causal effect or subgroup significance is inferred.

B.1 AGGREGATED POPULATION COMPARISONS

Table 8 shows the human rates, before/after rates, and distribution errors by domain. Every selected operator improves CC CE in all three domains. The associated costs differ: Llama's Family MID MAE slightly worsens, and OLMo's TV cost is concentrated in Environment and Family. This locates the aggregate trade-off instead of treating a larger CC rate as an automatic full-distribution improvement.

Table 8: Domain-level human targets and selected-controller effects. Rates are percentages; error changes are after minus before. Negative loss changes improve fidelity. CC CE is nats, MID MAE is pp, and TV is on $[0, 1]$. Human targets are repeated to permit direct comparison. Bold highlights the primary CC CE change; underlined positive secondary changes mark increased error.

	CC (%)		**MID (%)**		**Error change (after − before)**		
Model **Domain**	Human	Before → after	Human	Before → after	CC CE ↓	MID MAE ↓	TV ↓
Qwen Environment	3.51	2.19 → 3.52	25.42	11.01 → 12.63	**-0.0050**	-1.68	-0.025
Qwen Family	3.26	1.32 → 3.14	17.24	9.03 → 10.83	**-0.0102**	-0.81	-0.019
Qwen Health	3.95	1.86 → 3.64	23.88	13.82 → 15.30	**-0.0126**	-1.08	-0.020
Llama Environment	3.51	0.96 → 3.09	25.42	13.78 → 14.01	**-0.0157**	-0.36	-0.036
Llama Family	3.26	0.67 → 2.69	17.24	15.58 → 16.55	**-0.0309**	<u>+0.10</u>	-0.064
Llama Health	3.95	0.41 → 2.61	23.88	15.84 → 16.05	**-0.0420**	-0.46	-0.049
OLMo Environment	3.51	0.51 → 1.11	25.42	52.16 → 48.44	**-0.0185**	-3.19	<u>+0.034</u>
OLMo Family	3.26	1.03 → 2.15	17.24	31.15 → 30.63	**-0.0165**	-0.52	<u>+0.013</u>
OLMo Health	3.95	1.63 → 1.84	23.88	40.91 → 34.88	**-0.0181**	-5.20	-0.021

Country means in Figure 5 show both the human variation and its incomplete reproduction. CC CE improves in four of six countries for Qwen and five of six for Llama/OLMo. The human country CC range is 0.61–6.24%; this variation is hidden by the 3.57% global target. Country-level MID MAE improves in all six countries for Qwen/OLMo and five for Llama, but that does not imply improvement in the complete distribution. For example, OLMo improves country TV in only two of six countries.

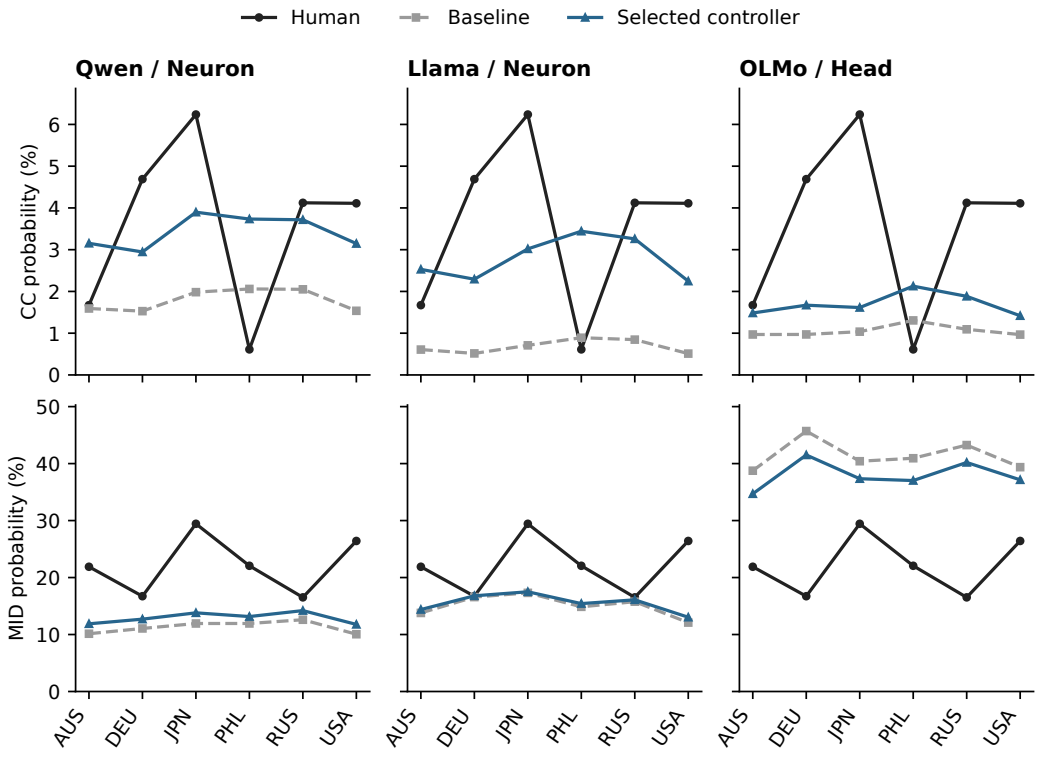


Figure 5: **Country variation, with each model evaluated separately.** Circles: human; squares: baseline; triangles: selected controller. CC and MID rates are averaged over domains and orders. Columns identify each model and its fixed operator. The controller is fixed before slicing. Rates visualize the direction of a change; the accompanying itemwise losses determine fidelity. These balanced-cell comparisons are descriptive rather than nationally representative.

B.2 Demographic variation

Table 9 retains each marginal group's human and model CC rates and the accompanying MID/TV error changes. CC CE improves in all four age bands, both sex categories, and both education bands in each model. These positive marginal results coexist with non-improving joint cells: 30/23/18 of the 96 cells do not improve CC CE in Qwen/Llama/OLMo (Table 10). The much lower OLMo count for TV improvement reflects its distributional trade-off rather than an absence of CC control. The 96 cells cross the same categories across domains; humans in different survey modules are not treated as the same individuals.

C Response-based interventions

C.1 Teacher forcing nominates candidates

For every development input we force each of six continuations, "letter. option text," regardless of the person's observed response. The contrast uses activations at the last forced-answer token; steering acts at the last prompt token. Because the model is causal, the same prefix has the same pre-answer activation under different later forced answers. Pre-answer activations provide the intervention input but cannot distinguish forced continuations that have not been read.

For $D = \{1, 2, 4, 5\}$ and forced-answer activation $h_k(x)$,

$$\bar{h}_D(x) = \tfrac{1}{4} \sum_{k \in D} h_k(x), \qquad c(x) = h_{CC}(x) - \bar{h}_D(x), \qquad m(x) = h_3(x) - \bar{h}_D(x). \tag{11}$$

Table 9: Marginal persona variation under the selected controller. CC rates are percentages. Error changes are after minus before; negative improves fidelity. MID MAE is pp, CC CE nats, TV on $[0, 1]$. Bold highlights primary CC CE changes; underlined positive secondary changes mark increased error. These are descriptive balanced-group comparisons.

Model	Group	Human CC (%)	Model CC (%) Before → after	ΔCC CE	ΔMID MAE	ΔTV
Qwen	Age: 18–34	4.05	2.00 → 3.75	**-0.0087**	-0.84	-0.022
Qwen	Age: 35–49	2.77	1.74 → 3.37	**-0.0039**	-1.26	-0.024
Qwen	Age: 50–64	3.42	1.62 → 3.19	**-0.0104**	-1.26	-0.019
Qwen	Age: 65–99	4.06	1.80 → 3.42	**-0.0147**	-1.24	-0.019
Qwen	Sex: Male	3.49	1.77 → 3.38	**-0.0092**	-1.31	-0.024
Qwen	Sex: Female	3.66	1.81 → 3.49	**-0.0093**	-1.10	-0.018
Qwen	Education: 0–12 years	4.43	2.03 → 3.77	**-0.0123**	-1.05	-0.022
Qwen	Education: 13–30 years	2.71	1.55 → 3.09	**-0.0060**	-1.28	-0.020
Llama	Age: 18–34	4.05	0.64 → 2.64	**-0.0376**	-0.75	-0.055
Llama	Age: 35–49	2.77	0.55 → 2.41	**-0.0244**	-0.51	-0.051
Llama	Age: 50–64	3.42	0.62 → 2.69	**-0.0313**	-0.21	-0.049
Llama	Age: 65–99	4.06	0.91 → 3.46	**-0.0289**	<u>+0.25</u>	-0.042
Llama	Sex: Male	3.49	0.67 → 2.78	**-0.0282**	-0.29	-0.050
Llama	Sex: Female	3.66	0.69 → 2.81	**-0.0312**	-0.26	-0.050
Llama	Education: 0–12 years	4.43	0.74 → 2.89	**-0.0379**	-0.06	-0.050
Llama	Education: 13–30 years	2.71	0.62 → 2.71	**-0.0209**	-0.39	-0.048
OLMo	Age: 18–34	4.05	0.92 → 1.54	**-0.0236**	-1.79	<u>+0.008</u>
OLMo	Age: 35–49	2.77	1.05 → 1.72	**-0.0118**	-2.19	<u>+0.006</u>
OLMo	Age: 50–64	3.42	1.09 → 1.72	**-0.0161**	-3.15	<u>+0.009</u>
OLMo	Age: 65–99	4.06	1.16 → 1.82	**-0.0204**	-4.95	<u>+0.012</u>
OLMo	Sex: Male	3.49	1.08 → 1.71	**-0.0173**	-2.80	<u>+0.006</u>
OLMo	Sex: Female	3.66	1.03 → 1.69	**-0.0182**	-3.11	<u>+0.011</u>
OLMo	Education: 0–12 years	4.43	1.08 → 1.74	**-0.0235**	-3.40	<u>+0.015</u>
OLMo	Education: 13–30 years	2.71	1.04 → 1.66	**-0.0121**	-2.40	<u>+0.004</u>

Table 10: Descriptive variation across the 96 joint cells. Counts use each cell's equally domain-weighted means. The first count compares mean baseline CC with the human rate; the other counts require a strictly lower after-intervention loss. Bold highlights the primary CC CE improvement count. Counts are not significance tests.

Model	Method	CC deficit	CC CE improves	MID MAE improves	TV improves
Qwen	Neuron	70/96	**66/96**	89/96	65/96
Llama	Neuron	84/96	**73/96**	79/96	93/96
OLMo	Head	77/96	**78/96**	91/96	26/96

MID is excluded from the directional reference. Shared directions average contrasts within each domain–cell, normalize each cell vector, average and normalize within domains, then average and normalize across domains. All 288 development people per domain contribute; repeated persona descriptions retain their multiplicity. These paired response contrasts can contain answer identity, lexical effects, and common-reference geometry.

Disjoint references assign one pair of directional choices to the CC contrast and the complementary pair to MID, then average across the saved pair assignments. Inverse-covariance cosine uses a shrinkage estimate from domain-balanced development pre-answer question–cell mean activations:

$$\cos_{\Sigma_\lambda^{-1}}(c, m) = \frac{c^\top \Sigma_\lambda^{-1} m}{\sqrt{c^\top \Sigma_\lambda^{-1} c}\sqrt{m^\top \Sigma_\lambda^{-1} m}}, \quad \Sigma_\lambda = (1-\lambda)\Sigma + \lambda \frac{\operatorname{tr}\Sigma}{4096} I. \tag{12}$$

Here $\lambda \in \{.1, .5\}$ stabilizes the inverse, and 4096 is the activation width. The positive alignment survives both checks (Table 11). No cosine threshold establishes a pure semantic feature or a shared human mental state.

Table 11: Median CC–MID similarity over 32 layers. The last two columns use inverse-covariance weighting. Measurements follow forced answers.

	Cosine		Inverse-covariance cosine	
Model	Shared reference	Disjoint reference	$\lambda = .1$	$\lambda = .5$
Qwen	0.432	0.338	0.435	0.430
Llama	0.460	0.377	0.457	0.457
OLMo	0.413	0.313	0.413	0.414

C.2 Frozen controller selection

Head and Neuron scale selected native coordinates by $1 + g$. Head selects two or four attention-output input slices; Neuron selects 64 or 256 gated coordinates entering the MLP down projection. Coordinate ranking multiplies absolute shared contrast by sign consistency across domain–cell vectors; Head sums these scores within each head. These interventions act in different spaces from a residual-vector addition. Additive applies $h' = h + \alpha\sigma d$, where d is the shared contrast and σ a fixed development projection scale. PCA-additive instead uses the leading direction of centered paired-contrast covariance, oriented toward the mean; rank is one. Angular rotates in a fixed candidate plane while preserving the residual norm in exact arithmetic. Spherical moves a signed fraction of the angular distance toward d while preserving the norm. For an orthonormal candidate plane (a, b) and coordinates $v_a = a^\top h$, $v_b = b^\top h$, Angular applies

$$h' = h + (v_a \cos\theta - v_b \sin\theta - v_a)a + (v_a \sin\theta + v_b \cos\theta - v_b)b. \tag{13}$$

The plane and orientation are fixed from development contrasts; the fitted angle is converted from degrees to radians. For Spherical, set $u = h/\|h\|$, $t = \operatorname{sign}(\gamma)d$, $\phi = \arccos(u^\top t)$, and $v =$

$(u - (u^\top t)t)/\sin\phi$:

$$h' = \|h\|\{\cos[(1-|\gamma|)\phi]t + \sin[(1-|\gamma|)\phi]v\}. \tag{14}$$

Zero strength gives the identity; deterministic tangents handle degenerate angles, and the implementation renormalizes the result for numerical stability. Additive's σ is the square root of the mean development pre-answer projection variance along the CC and MID directions. Qwen Head uses 16 slices of width 256 at full-attention layers and 32 slices of width 128 at the other targeted attention adapters. OLMo's delivered MLP change is measured after its post-feedforward normalization.

Within each development cell, two people form fitting set F and one selection set H; coarse set C takes one F person. Directions use all development people, so H is not an independent discovery holdout. Candidate bundles contain the top one, three, or five layers, and consecutive three/five layers starting at the highest-ranked layer where possible. Coarse optimization evaluates all 42 questions and retains two distinct bundles for fine optimization. Each question receives one coefficient per selected layer, fitted on F; H chooses a multiplier from $\{.25, .5, 1\}$ and the primary bundle. There is no fitted per-person coefficient. The second retained bundle is a sensitivity condition, called the robustness bundle in the 72-condition analysis; it is fixed before test.

Bounded multistart optimization uses two separated starts, radius .125 of the normalized range, radius halving after an unsuccessful sweep, minimum $1/128$, and at most eight rounds. Each question/bundle permits up to 24 additional evaluations on C and 48 on F; previously evaluated candidates remain eligible. Head/Neuron initial gains span $[-.75, 4]$, hard bound $[-.75, 8]$; Additive/PCA-additive $[-4, 4]$ then $[-8, 8]$; Angular $[-60, 60]$ then $[-120, 120]$ degrees; Spherical $[-.25, .25]$ then $[-.9, .9]$. Only finite candidates losing at most .01 legal mass and delivering a perturbation above 10^{-8} are eligible. Steering must remain active; zero is a reference rather than a fallback, so worsening results are possible. CC CE alone is optimized; MID and TV are measured consequences. Search histories and finite budgets differ across methods, so these comparisons establish attainable control rather than global optimality or equal-compute superiority. Table 12 records the frozen primary layer tuples, support sizes, and applied ranges; the code supplement contains the question-specific coefficients.

D Controller evaluation

D.1 Generalization across evaluation settings

Table 13 reports achieved control on the people used for development, including the internal F/H fitting and selection subsets. These in-sample results quantify the correction reached by the bounded search; they are neither an independent generalization estimate nor a global optimum. The same frozen parameters are then applied to held-out people under A, followed by layouts B–D. Table 14 separates these evaluation axes for the selected method, while Table 15 retains all six methods over the four test layouts.

The different human targets contribute to absolute CE differences across splits; baseline-relative changes compare the correction within each split. All three selected methods retain approximately the development CC CE benefit on held-out people under A. Layout changes introduce a separate cost, most visibly for Llama: its held-out CC MAE is .51 pp under A and 2.48 pp averaged over A–D. Secondary outcomes also depend on the evaluation axis. Llama Neuron worsens MID MAE under A on development and held-out people, although its four-order mean improves slightly. OLMo Head improves TV under A on both populations but worsens the four-order test mean.

D.2 Controls under four answer layouts

Figure 6 shows how the selected controls redistribute probability across all six response categories.

Retuned random controllers test whether the selected extraction improves practical optimization relative to alternative supports under the same search procedure. Fixed opposite/random/orthogonal displacements test dependence on direction at a chosen site and dose. Same-input restoration measures a component's conditional contribution. These are different estimands: a trained network's random components can be behaviorally active, while orthogonality to one axis does not imply functional irrelevance. CC/MID construct ambiguity separately limits interpretation as human opinion absence; it does not make alternative internal controls unnecessary.

Table 12: Frozen primary layer tuples, support size k, and applied-coefficient ranges across questions and layers. Angular coefficients are degrees; the other coefficients follow the operator definitions. Bold method names identify the development-selected controller.

				Applied coefficient		
Model	**Method**	**Layers**	k	Min.	Median	Max.
Qwen	Head	7,23,31	2	-0.75	1.8	8
Qwen	Angular	0,1,2,3,4	–	-120	38.9	120
Qwen	Additive	0,1,2	–	-4	-0.119	3.53
Qwen	PCA-add.	0	–	-2	0.242	8
Qwen	**Neuron**	29,30,31	256	-0.713	1.08	4.63
Qwen	Spherical	0,1,2,3,4	–	-0.175	-0.0209	0.225
Llama	Head	5,6,7,8,9	4	-0.75	0.898	8
Llama	Angular	0,1,2	–	-75	-1.95	90
Llama	Additive	0,1,2,3,4	–	-4	-0.0163	4
Llama	PCA-add.	0,1,2	–	-4	0.367	4
Llama	**Neuron**	31	64	-0.75	3.76	8
Llama	Spherical	0,1,2,3,4	–	-0.504	-0.25	0.371
OLMo	**Head**	10,11,12,13,14	2	-0.75	2.15	8
OLMo	Angular	0,1,2,3,4	–	-45.9	-1.88	120
OLMo	Additive	0,1,2	–	-3	-0.25	3
OLMo	PCA-add.	0,1,2,3,4	–	-3.94	-0.172	4
OLMo	Neuron	25,28,29,30,31	64	-0.75	-0.75	5.96
OLMo	Spherical	0,1,2,3,4	–	-0.504	-0.179	0.287

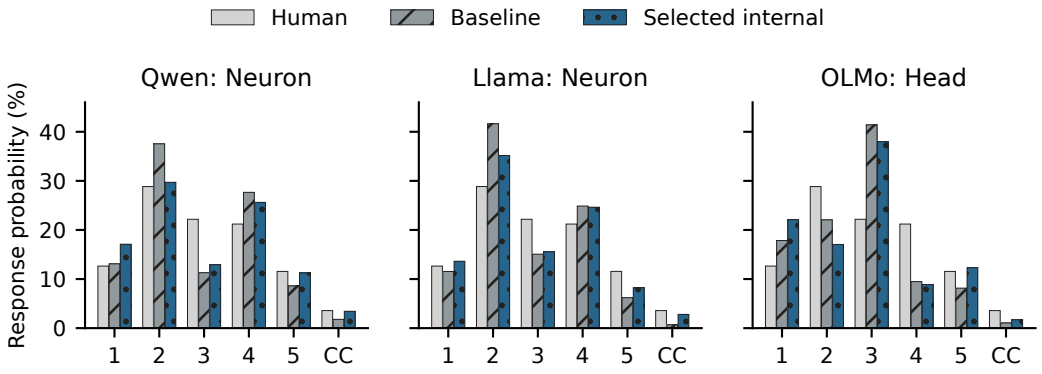


Figure 6: All six response categories for human targets, baseline, and selected internal control. Categories 1–5 are the ordered agreement scale, with 3 MID. Each model's internal method is selected on development CC CE and identified above its panel. Four-order test rates are shown.

Primary-minus-retuned-random CC CE is $-.00641/+.00057/+.00074$ in Qwen/Llama/OLMo. The latter two intervals include zero, establishing neither equivalence nor superiority. Llama random is better on CC MAE by .54 pp, interval [.32,.78], while primary MID MAE and TV point estimates are better. OLMo random also has a slightly lower CC MAE point estimate (2.20 versus 2.23 pp), while the selected controller has lower MID MAE and TV. The random comparison selects its best operator by the same development criterion and uses two fixed seeds; it does not average six operators into one comparator. Only two random seeds were used in these inference comparisons; broad claims about a random-direction distribution are unwarranted.

Table 13: All six methods on controller-development people, order A. Human CC/MID targets are 3.46/21.86%. Rates are percentages, MAEs are pp, CE is nats, and TV is on $[0, 1]$. The human CE row is target entropy. These people contributed to extraction, fitting, and selection. Bold errors mark column minima within each model (ties at displayed precision); bold method names mark the selected controller.

Model	Intervention	CC (%)	MID (%)	CC CE ↓	CC MAE ↓	MID MAE ↓	TV ↓
Human	target	3.46	21.86	0.1473	0.00	0.00	0.000
Qwen	Baseline	1.64	10.63	0.1622	2.01	11.23	0.345
	Head	3.10	10.11	0.1485	0.41	11.85	0.394
	Angular	2.00	10.70	0.1568	1.54	11.17	0.333
	Additive	1.64	10.65	0.1622	2.01	11.22	0.344
	PCA-add.	1.64	10.65	0.1622	2.01	11.22	0.344
	Neuron	3.14	16.54	**0.1481**	**0.34**	**7.17**	**0.324**
	Spherical	1.98	10.11	0.1572	1.60	11.75	0.341
Llama	Baseline	0.89	14.28	0.1972	2.72	8.34	0.310
	Head	2.66	21.82	0.1547	0.85	6.68	**0.239**
	Angular	0.90	14.36	0.1967	2.71	8.27	0.309
	Additive	0.90	14.29	0.1971	2.72	8.33	0.310
	PCA-add.	0.90	14.30	0.1971	2.72	8.32	0.310
	Neuron	3.34	12.17	**0.1476**	**0.22**	9.89	0.274
	Spherical	1.96	19.92	0.1650	1.63	**5.64**	0.259
OLMo	Baseline	0.76	47.15	0.1991	2.95	30.54	0.427
	Head	1.54	46.03	**0.1672**	**1.93**	29.28	0.420
	Angular	0.74	47.25	0.1989	2.93	30.60	0.427
	Additive	0.76	47.14	0.1991	2.95	30.55	0.427
	PCA-add.	0.75	47.15	0.1991	2.95	30.55	0.427
	Neuron	1.12	47.86	0.1779	2.35	30.53	0.418
	Spherical	1.29	23.45	0.1747	2.33	**15.81**	**0.346**

Table 14: Separating transfer to new people from layout robustness. Neuron/Neuron/Head are fixed for Qwen/Llama/OLMo. ΔCE is after minus before, so negative values improve fidelity. CC MAE is in pp. Held-out human CC/MID targets are 3.57/22.18%. A–D averages order-specific losses on the same held-out people. Bold changes highlight the within-row CC CE reduction, not a ranking across populations.

Model	Population / order	CC CE Before	CC CE After	ΔCC CE	CC MAE (pp) After
Qwen	Development A	0.1622	0.1481	**-0.01408**	0.34
	Held-out A	0.1666	0.1525	**-0.01408**	0.57
	Held-out A–D	0.1663	0.1570	**-0.00927**	1.43
Llama	Development A	0.1972	0.1476	**-0.04961**	0.22
	Held-out A	0.2041	0.1521	**-0.05202**	0.51
	Held-out A–D	0.2052	0.1756	**-0.02953**	2.48
OLMo	Development A	0.1991	0.1672	**-0.03192**	1.93
	Held-out A	0.2053	0.1724	**-0.03287**	2.10
	Held-out A–D	0.1931	0.1753	**-0.01772**	2.23

Table 15: All six primary methods on held-out people, averaged across layouts A–D. Rates are percentages, MAEs are pp, CE is nats, and TV is on $[0, 1]$; lower errors are better. The human CE entry is target entropy. Bold errors mark column minima within each model (ties at displayed precision); bold method names mark the development-selected controller.

Model	Intervention	CC (%)	MID (%)	CC CE ↓	CC MAE ↓	MID MAE ↓	TV ↓
Human	target	3.57	22.18	0.1515	0.00	0.00	0.000
Qwen	Baseline	1.79	11.29	0.1663	1.96	11.25	0.303
	Head	4.33	12.59	0.1627	2.14	10.51	0.306
	Angular	1.96	11.78	0.1633	1.74	10.86	0.291
	Additive	1.79	11.29	0.1663	1.96	11.25	0.303
	PCA-add.	1.79	11.29	0.1663	1.96	11.25	0.303
	Neuron	3.43	12.92	**0.1570**	**1.43**	**10.06**	**0.282**
	Spherical	1.95	11.67	0.1638	1.79	11.03	0.294
Llama	Baseline	0.68	15.06	0.2052	2.98	9.09	0.312
	Head	1.87	21.30	0.1789	**2.07**	7.28	**0.247**
	Angular	0.68	15.12	0.2050	2.98	9.05	0.311
	Additive	0.68	15.06	0.2052	2.98	9.09	0.312
	PCA-add.	0.68	15.06	0.2052	2.98	9.09	0.312
	Neuron	2.80	15.54	**0.1756**	2.48	8.85	0.262
	Spherical	1.31	19.72	0.1914	2.48	**6.44**	0.263
OLMo	Baseline	1.06	41.41	0.1931	2.79	24.53	0.413
	Head	1.70	37.99	**0.1753**	**2.23**	21.56	0.422
	Angular	1.05	41.48	0.1930	2.79	24.57	0.413
	Additive	1.06	41.40	0.1931	2.79	24.52	0.413
	PCA-add.	1.06	41.41	0.1931	2.79	24.53	0.413
	Neuron	1.38	39.87	0.1820	2.46	22.84	0.407
	Spherical	2.47	25.27	0.1780	2.74	**14.91**	**0.383**

Entropy-fitted rescaling uses human development directional entropy. It excludes CC/MID fitting targets but remains human-supervised, making it a practical output reference with a different fitting objective from internal controls. The separate source-global Head comparison has mixed new-domain outcomes (Appendix D.3). Table 16 retains each method's order-specific errors; no order is selected after viewing test performance.

D.3 Scope on new questions

A source-global Head controller fits one layer-wise gain tuple on source fitting people and one multiplier on selection people, then freezes them for six Government 2006 items with the same 96 cells and 288 personas. Its supports are Qwen layers 7/23/31 with two heads, Llama 5–9 with four, and OLMo 10–14 with two. This is a separate controller from the question-specific primary methods. The Bias comparator adds one shared offset to the semantic CC log probability and renormalizes the six choices, preserving conditional non-CC probabilities. For each model, the offset minimizes source fitting-set CC CE; the source selection set chooses a multiplier from $\{.25, .5, 1\}$, after which it is frozen for all target items. Table 17 retains both order A and the four-order probability mixture; the latter differs from averaging order-specific losses in the main evaluation. Qwen improves TV but worsens CC CE; Llama improves CC CE slightly while worsening CC MAE; OLMo's order-A CC CE benefit does not persist in the mixed-order comparison. These results bound transfer beyond the known-question design.

Table 16: Four frozen test orders, with CC cross-entropy and the minimum–maximum six-category TV over orders. Bold CE values are minima within each model/order; bold method names identify the development-selected controller. No order-specific retuning is performed.

Model	**Method**	A CE ↓	B CE ↓	C CE ↓	D CE ↓	TV range
Qwen	Head	0.1527	0.1754	**0.1552**	0.1674	0.236–0.400
Qwen	Angular	0.1609	0.1690	0.1577	0.1656	0.246–0.336
Qwen	Additive	0.1666	0.1707	0.1611	0.1670	0.260–0.348
Qwen	PCA-add.	0.1666	0.1707	0.1611	0.1669	0.260–0.348
Qwen	**Neuron**	**0.1525**	**0.1579**	0.1566	**0.1612**	0.247–0.328
Qwen	Spherical	0.1614	0.1696	0.1581	0.1663	0.241–0.344
Llama	Head	0.1586	0.2005	**0.1571**	0.1993	0.243–0.255
Llama	Angular	0.2037	0.2205	0.1849	0.2109	0.286–0.348
Llama	Additive	0.2041	0.2205	0.1852	0.2108	0.286–0.349
Llama	PCA-add.	0.2041	0.2205	0.1853	0.2108	0.286–0.349
Llama	**Neuron**	**0.1521**	**0.1935**	0.1686	**0.1884**	0.252–0.278
Llama	Spherical	0.1700	0.2182	0.1624	0.2151	0.256–0.277
OLMo	**Head**	**0.1724**	0.1847	**0.1615**	0.1827	0.359–0.490
OLMo	Angular	0.2052	0.1873	0.1820	0.1976	0.348–0.493
OLMo	Additive	0.2053	0.1873	0.1821	0.1976	0.348–0.494
OLMo	PCA-add.	0.2053	0.1873	0.1820	0.1976	0.348–0.494
OLMo	Neuron	0.1834	**0.1841**	0.1658	0.1944	0.335–0.485
OLMo	Spherical	0.1804	0.1846	0.1710	**0.1761**	0.299–0.490

Table 17: Government 2006 transfer with frozen source-global Head steering. ABCD denotes scoring the four-order probability mixture, not averaging four losses. CC rates are percentages, MAEs pp, CE nats, and TV is on $[0, 1]$. Bold losses are minima within each model/order block, with ties at displayed precision.

Model	**Order**	**Arm**	CC	CC CE ↓	CC MAE ↓	MID MAE ↓	TV ↓
Qwen	A	baseline	3.02	**0.1768**	**1.47**	8.92	**0.285**
Qwen	A	Head	2.38	0.1780	1.84	**8.65**	0.321
Qwen	A	Bias	6.16	0.1792	2.74	8.95	0.286
Qwen	ABCD	baseline	3.20	**0.1749**	**1.60**	9.40	0.226
Qwen	ABCD	Head	8.84	0.1901	4.74	**7.63**	**0.198**
Qwen	ABCD	Bias	6.52	0.1792	2.42	9.65	0.225
Llama	A	baseline	3.54	**0.1785**	**1.50**	10.02	0.322
Llama	A	Head	6.47	0.1808	3.18	**7.96**	**0.250**
Llama	A	Bias	13.09	0.2256	9.21	10.31	0.330
Llama	ABCD	baseline	2.85	0.1783	**1.64**	6.42	**0.159**
Llama	ABCD	Head	5.08	**0.1753**	1.85	**5.57**	**0.159**
Llama	ABCD	Bias	10.44	0.2032	6.34	6.74	0.180
OLMo	A	baseline	0.15	0.2989	3.95	32.99	0.509
OLMo	A	Head	0.24	0.2678	3.86	**29.41**	**0.493**
OLMo	A	Bias	0.75	**0.2382**	**3.35**	32.78	0.504
OLMo	ABCD	baseline	0.49	0.2541	3.61	32.61	0.422
OLMo	ABCD	Head	0.30	0.2564	3.80	**26.80**	0.423
OLMo	ABCD	Bias	2.11	**0.2067**	**2.86**	31.89	**0.417**

E CONNECTING THE CONTROLLERS TO RESPONSE CONCENTRATION

The diagnostic in Equation (2) regresses intervened centered log probabilities on their baseline values over categories 1/2/4/5, using development layout A. It then predicts all six probabilities from the baseline without using intervened test CC/MID scores. The expectation uses the population-balancing weights; the slope is frozen during test. A CC-only offset changes neither conditional directional odds nor their centered log probabilities, so CC fitting alone does not algebraically force $\beta_a < 1$. Nonetheless, selecting controllers for CC is a substantive source of dependence, and the observed association is exploratory.

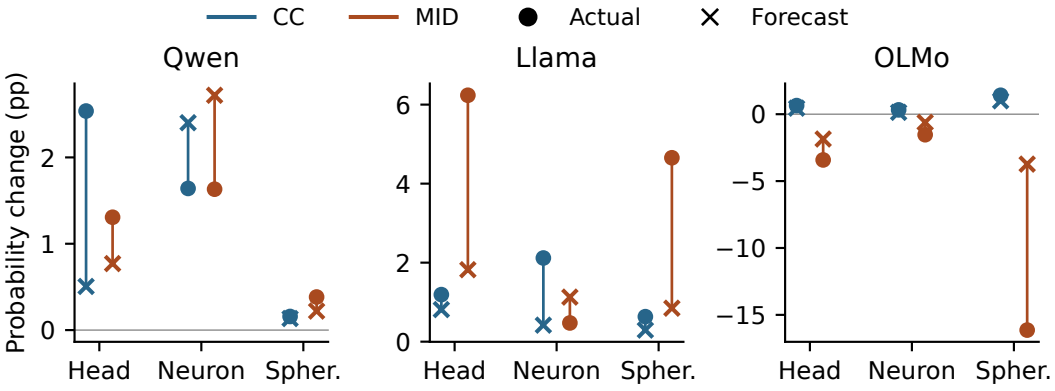


Figure 7: Directional-only forecasts of original controls. Blue/orange denote CC/MID. Circles are actual effects and crosses are development-fitted forecasts, in pp; connecting lines show residuals. Axes differ across models. Head, Neuron, and Spherical are displayed to show contrasting operators; all six remain in the complete tables.

The six-category predictive reduction is $R_6 = 1 - \mathrm{MSE}(\hat{p}, p_a)/\mathrm{MSE}(p_0, p_a)$. It compares the forecast with predicting no probability change. It is neither a fraction of human-fidelity improvement nor a causal mediation estimate. Across the nine controls in Figure 7, R_6 is 7.7–45.9%. The remaining change can be large: Llama Head increases MID by 6.24 pp versus a 1.82-pp forecast, and OLMo Spherical decreases it by 16.14 pp versus 3.73 pp. Qwen Neuron's MID effect changes sign across layouts. Alternative one-parameter forecasts toward a uniform distribution are competitive; an input-specific retrospective fit still leaves residuals. These results support a recurring distributional component without selecting a unique mathematical transformation. The broader accounting includes 72 primary, second-bundle, and retuned-random conditions. All 32 with $|1 - \beta_a| \geq .01$ flatten scores and improve CC CE; 29 raise directional entropy and 13 use random supports. The threshold was chosen retrospectively, and shared models, inputs, and candidates make the conditions dependent.

The native-baseline entropy comparisons use human-cell entropy rather than the entropy of a pooled population. The across-item correlations are unweighted Pearson correlations over 42 questions. For each question, the CC gap is human log odds minus the log odds of the model's probability averaged over personas and four layouts; the entropy gap is mean human-cell entropy minus mean persona-level model entropy. The within-stratum correlations center model entropy and CC log odds separately within each question–layout stratum, then pool covariances and variances with population-balancing weights; they do not average stratum-specific correlations. The correction-allocation comparison correlates rescaled-minus-baseline CC probability with human-minus-baseline CC probability across questions. All entropy comparisons here condition on directional responses. The entropy-fitted rescaling in Results uses one β_H per model so that mean conditional-directional entropy across development prompts matches mean human-cell directional entropy. Both means use the population-balancing weights, and the fitted value is shared across all test questions. This target differs from matching entropy after pooling responses into an item distribution. Table 18 shows the fitted slopes and sensitivity to the entropy target. For a six-choice distribution, $S_\beta(p)_k/p_k = p_k^{\beta-1}/\sum_j p_j^\beta$; when $0 < \beta < 1$ and $p_k < 1/6$, concavity gives $\sum_j p_j^\beta \leq 6^{1-\beta}$ and hence $S_\beta(p)_k > p_k$. Raising

a sufficiently rare response is thus a property of rescaling; the empirical results concern the amount of human-gap recovery, its allocation across items, and its recurrence under internal edits.

Table 18: **Entropy-fitting sensitivity.** CC rates are percentages; gap closure compares corrected and baseline means with the human 3.57%. Brackets: paired-item 95% intervals with scalar refitting. Cell is the main input–cell entropy target; MM adds the Miller–Madow correction, and Agg. uses pooled item entropy. Bold marks the main Cell estimate and mean-gap closure.

Model	β	Base CC	Cell CC	Gap closed (%)	MM CC	Agg. CC
Qwen	0.735	1.79	**3.41**	**91** [60,127]	4.22	6.47
Llama	0.586	0.68	**2.44**	**61** [46,80]	3.25	5.68
OLMo	0.697	1.06	**2.51**	**58** [33,92]	3.41	6.05

Within-question–order associations hold question and layout fixed but remain observational across persona prompts. These concentration comparisons leave the training origin of the deficit and the pre-answer decodability of human CC propensity unresolved.

F FIXED-PATH INTERVENTIONS

F.1 A FIXED STATE CONTRAST WITH AN ORDINARY-PROMPT TEST

State development uses three source statements, two personas, four orders, two phrasings, and four supplied states (unformed, neutral, agree, disagree): 192 inputs per model, including 48 matched U/N pairs. The evaluation in Figure 8 uses six other statements and four personas, giving 768 inputs. The two U phrases are "You have not yet formed an opinion on that statement" and "Your view on that statement is still unformed." N describes a considered, settled position that is neither agreement nor disagreement. A/D supply considered directional positions. Only this paragraph is inserted after the survey statement; the six answers stay available and none is forced.

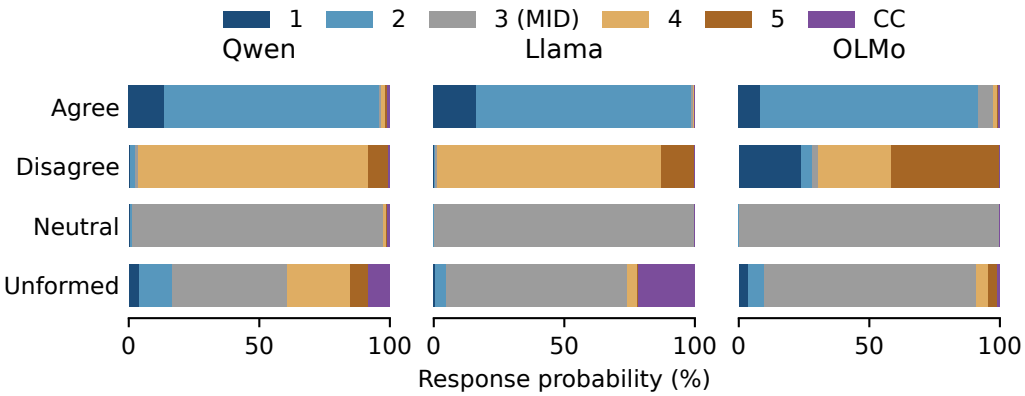


Figure 8: Responses under four supplied opinion states on a separate six-item evaluation cohort. Agree/disagree are directional controls; neutral (N) specifies a settled midpoint and unformed (U) an unsettled opinion. Colors denote semantic responses, with 3 = MID. U can elicit both CC and MID, whereas N mainly elicits MID. These are experimental descriptions, not verified respondent labels. The evaluation averages four answer layouts and two phrasings.

Development full-update interchange measures the paired change in CC log odds when U replaces N and when N replaces U at attention and MLP updates in layers $3, 7, 11, 15, 19, 23, 27, 31$. The maximum absolute mean bidirectional effect selects one site of each component type without human response targets: attention/MLP layers 23/31 in Qwen, 15/19 in Llama, and 15/19 in OLMo. At the selected site, matched pre-answer module updates at the last prompt token define the unit U–N

direction d and projection target μ_U (Table 19). For update u,

$$u' = u + (\mu_U - d^\top u)d, \qquad \mu_U = \mathbb{E}_{\text{development}}[d^\top u_U]. \tag{15}$$

N uses μ_N. At test, the state paragraph is absent and the fixed anchor acts on the ordinary survey input. Its displacement varies with the input, but no human target is used to adjust it. The full set of development-selected attention/MLP U/N outcomes appears in Table 20; it includes weak and contrary cases. The edit transfers a supplied-state contrast into response control; it is not trained to classify survey respondents' latent attitudes. A separate mean-CC-matched temperature uses U's model-produced mean CC on development prompts as its target and is frozen for test. It acts on answer-token alias logits before semantic pooling, whereas β_H and the directional slope act on pooled semantic probabilities.

Table 19: Fixed projection targets from development opinion-state contrasts. Layers are zero-based. The final column is the raw U–N mean-contrast norm, not the delivered ordinary-prompt perturbation.

		Fixed projection target		
Model	**MLP layer**	μ_U	μ_N	$\|\bar{u}_U - \bar{u}_N\|$
Qwen	31	30.541804	11.948063	18.593742
Llama	19	1.514190	-0.932754	2.446944
OLMo	19	2.916283	-1.613725	4.530009

Table 20: **All development-selected state-derived intervention sites.** Changes from the corresponding ordinary-prompt baseline use source test people and four layouts. Negative changes improve fidelity; MID MAE is pp. U/N results retain weak and contrary cases at the same selected sites. Bold marks the lowest change per loss within each model, a descriptive comparison at these fixed sites.

Model	**Site**	**Anchor**	ΔCC CE ↓	ΔMID MAE (pp) ↓	ΔTV ↓
Qwen	Attn 23	U	+0.00060	-0.142	-0.00012
Qwen	Attn 23	N	+0.00339	**-1.272**	**-0.00041**
Qwen	MLP 31	U	**-0.00017**	-0.050	-0.00022
Qwen	MLP 31	N	+0.00158	+0.983	+0.01022
Llama	Attn 15	U	-0.00654	**-0.623**	-0.01300
Llama	Attn 15	N	+0.00836	+0.654	+0.01603
Llama	MLP 19	U	**-0.01486**	-0.597	**-0.02384**
Llama	MLP 19	N	+0.00921	+0.263	+0.01128
OLMo	Attn 15	U	+0.00154	+1.283	+0.00610
OLMo	Attn 15	N	+0.02758	+15.157	+0.08025
OLMo	MLP 19	U	**-0.00334**	**-0.776**	**-0.00615**
OLMo	MLP 19	N	+0.01386	+2.768	+0.02383

F.2 Holding the upstream edit fixed

Let the unmodified input produce residual updates a_l^0, m_l^0. A clamped branch applies the upstream edit while replaying later baseline updates; a live branch recomputes later updates on the changed state. Cumulative release at boundary b recomputes updates through b and clamps the remainder, always evaluating the final readout rather than a raw intermediate logit lens. The contrast $L(p_{\text{live}}) - L(p_{\text{clamped}})$ measures the conditional consequence of allowing that response at the retained edit. It is a counterfactual path intervention, whose clamped branch need not be a naturally occurring state.

Intervention divergence therefore constrains wider interpretation (Grant et al., 2026; Opielka et al., 2026).

The fixed U edit has small direct CC effects but larger live effects in Llama/OLMo. Source and six additional Environment items show the same downstream fidelity signs. The main depth plot uses development-fitted slopes from the cumulative release branches; its test R_6 is evaluated on the corresponding held-out-person probabilities. Llama's final slope is .9199, with $R_6 = 52.7\%$; OLMo's is .9670, with $R_6 = 21.0\%$. Qwen's final U slope is .9862 but $R_6 = -7.7\%$, another limit on this particular route. These outcomes motivate model-specific interventions rather than assuming U must dominate every model. Table 21 retains the fixed U slope's CC log-odds prediction on source and six additional Environment 2010 items, separately under A and B–D. The additional items test output-change prediction under an already fixed edit; they do not predict which new site or dose will induce a helpful downstream response.

Table 21: **Fixed U-slope prediction beyond the fitting inputs.** Entries are percentage CC-log-odds MSE reductions relative to no change. Slopes use source development A. Source columns use held-out people; additional columns use the six Environment 2010 statements. Bold positive values improve on predicting no change; negative values are worse.

	Source questions		**Additional questions**	
Model	A	B–D	A	B–D
Qwen	-22.8	-15.7	-30.3	**1.0**
Llama	**91.0**	**79.5**	**83.2**	**73.8**
OLMo	**45.8**	**55.2**	**59.0**	**63.1**

For Qwen H14, the clamped comparison holds the following MLP update at baseline; the live branch recomputes it. The head-reduction slope .903 describes clamped versus baseline outputs; .962 describes live versus baseline; 1.064 describes live versus clamped. Each slope is fitted separately for its stated comparison. The following MLP worsens the beneficial reduction's CC CE by .0022 while reversing sign under head amplification. The individual-head follow-up is exploratory; H14 is a diagnosed site rather than an asserted complete circuit. Tables 22–23 retain all 16 heads, including H12, whose signed effect opposes H14 and is also buffered by the following MLP.

Table 22: **Qwen L31 heads 0–7.** Both fixed multipliers and live/baseline-clamped following MLP are shown. Probability and MID-MAE changes are pp; CC CE is nats and TV is on $[0, 1]$. Same-run source-test baselines are used. Relative multipliers do not match absolute perturbation norms between heads.

Head	Multiplier	MLP	ΔCC	ΔMID	ΔCE	ΔMID MAE	ΔTV
0	0.75	Live	+0.012	-0.20	-0.00007	+0.129	+0.0006
0	0.75	Clamped	+0.047	-0.30	-0.00037	+0.183	-0.0015
0	1.25	Live	-0.012	+0.20	+0.00008	-0.135	-0.0006
0	1.25	Clamped	-0.046	+0.31	+0.00041	-0.184	+0.0014
1	0.75	Live	-0.023	+0.05	+0.00027	-0.052	+0.0008
1	0.75	Clamped	+0.029	+0.22	-0.00029	-0.202	-0.0018
1	1.25	Live	+0.022	-0.05	-0.00021	+0.049	-0.0007
1	1.25	Clamped	-0.023	-0.21	+0.00037	+0.195	+0.0019
2	0.75	Live	+0.044	+0.07	-0.00048	-0.038	-0.0012
2	0.75	Clamped	-0.016	-0.03	+0.00019	+0.031	+0.0019
2	1.25	Live	-0.044	-0.07	+0.00051	+0.039	+0.0013
2	1.25	Clamped	+0.017	+0.02	-0.00016	-0.032	-0.0018
3	0.75	Live	+0.018	+0.01	-0.00024	-0.016	-0.0009
3	0.75	Clamped	-0.004	-0.09	+0.00008	+0.081	+0.0010
3	1.25	Live	-0.017	-0.01	+0.00027	+0.012	+0.0009
3	1.25	Clamped	+0.005	+0.09	-0.00007	-0.082	-0.0010
4	0.75	Live	+0.045	+0.08	-0.00048	-0.086	-0.0030
4	0.75	Clamped	-0.003	-0.01	+0.00006	-0.009	-0.0002
4	1.25	Live	-0.045	-0.08	+0.00052	+0.086	+0.0032
4	1.25	Clamped	+0.004	+0.01	-0.00006	+0.006	+0.0001
5	0.75	Live	-0.008	+0.03	+0.00009	-0.019	-0.0005
5	0.75	Clamped	-0.031	-0.04	+0.00031	+0.029	+0.0005
5	1.25	Live	+0.009	-0.03	-0.00008	+0.020	+0.0006
5	1.25	Clamped	+0.032	+0.04	-0.00028	-0.035	-0.0005
6	0.75	Live	+0.012	+0.04	-0.00017	-0.035	-0.0009
6	0.75	Clamped	-0.020	-0.02	+0.00022	+0.006	+0.0003
6	1.25	Live	-0.013	-0.04	+0.00019	+0.037	+0.0010
6	1.25	Clamped	+0.022	+0.02	-0.00020	-0.007	-0.0003
7	0.75	Live	+0.044	+0.03	-0.00050	-0.027	-0.0014
7	0.75	Clamped	+0.018	-0.09	-0.00019	+0.078	+0.0002
7	1.25	Live	-0.043	-0.03	+0.00052	+0.022	+0.0015
7	1.25	Clamped	-0.017	+0.09	+0.00020	-0.079	-0.0002

Table 23: **Qwen L31 heads 8–15.** Both fixed multipliers and live/baseline-clamped following MLP are shown. Probability and MID-MAE changes are pp; CC CE is nats and TV is on $[0, 1]$. Same-run source-test baselines are used. Relative multipliers do not match absolute perturbation norms between heads. Bold identifies the H12/H14 cases discussed in Results and their CC changes, including both effect signs.

Head	Multiplier	MLP	ΔCC	ΔMID	ΔCE	ΔMID MAE	ΔTV
8	0.75	Live	+0.012	+0.08	+0.00007	-0.057	-0.0013
8	0.75	Clamped	+0.020	-0.18	-0.00005	+0.122	+0.0017
8	1.25	Live	-0.009	-0.07	-0.00006	+0.048	+0.0011
8	1.25	Clamped	-0.017	+0.18	+0.00008	-0.123	-0.0017
9	0.75	Live	+0.029	+0.17	-0.00028	-0.148	-0.0042
9	0.75	Clamped	-0.061	-0.08	+0.00080	+0.070	+0.0046
9	1.25	Live	-0.028	-0.15	+0.00029	+0.136	+0.0040
9	1.25	Clamped	+0.065	+0.08	-0.00076	-0.072	-0.0046
10	0.75	Live	+0.087	+0.19	-0.00090	-0.160	-0.0044
10	0.75	Clamped	+0.009	-0.15	+0.00008	+0.137	+0.0020
10	1.25	Live	-0.082	-0.19	+0.00097	+0.150	+0.0045
10	1.25	Clamped	-0.007	+0.15	-0.00006	-0.144	-0.0019
11	0.75	Live	-0.005	-0.01	+0.00005	+0.008	+0.0003
11	0.75	Clamped	-0.014	+0.04	+0.00016	-0.026	+0.0007
11	1.25	Live	+0.005	+0.00	-0.00003	-0.006	-0.0002
11	1.25	Clamped	+0.015	-0.04	-0.00015	+0.022	-0.0008
12	0.75	Live	**-0.211**	+0.05	**+0.00305**	+0.053	+0.0039
12	0.75	Clamped	**-0.466**	-0.82	**+0.00683**	+0.826	+0.0306
12	1.25	Live	**+0.280**	+0.08	**-0.00301**	-0.164	-0.0070
12	1.25	Clamped	**+0.616**	+0.83	**-0.00497**	-0.849	-0.0307
13	0.75	Live	+0.057	+0.17	-0.00062	-0.151	-0.0048
13	0.75	Clamped	-0.040	-0.13	+0.00041	+0.105	+0.0036
13	1.25	Live	-0.056	-0.16	+0.00067	+0.147	+0.0048
13	1.25	Clamped	+0.041	+0.13	-0.00037	-0.108	-0.0037
14	0.75	Live	**+0.182**	-0.03	**-0.00218**	-0.073	-0.0056
14	0.75	Clamped	**+0.498**	+0.92	**-0.00436**	-0.896	-0.0324
14	1.25	Live	**-0.122**	+0.18	**+0.00187**	-0.050	+0.0016
14	1.25	Clamped	**-0.389**	-0.84	**+0.00558**	+0.834	+0.0310
15	0.75	Live	-0.049	+0.00	+0.00063	+0.036	+0.0008
15	0.75	Clamped	-0.095	-0.21	+0.00110	+0.239	+0.0078
15	1.25	Live	+0.058	+0.03	-0.00067	-0.067	-0.0015
15	1.25	Clamped	+0.100	+0.22	-0.00103	-0.251	-0.0078

F.3 Response changes beyond norm scaling

For $\ell_M = \log[p(MID)/\sum_{k\in D} p(k)]$, Llama's final single U release step, from layer 30 to 31, changes ℓ_M by +.08652, versus +.07850 predicted from directional scaling. OLMo's same single step changes it by +.05707, versus -.00268 predicted. Its earlier release 22 to 23 changes it by -.03887, versus -.00015 predicted. Thus substantial OLMo MID computation remains outside the shared concentration diagnostic, even when its final average MID moves toward the human target. These contrasts distinguish cumulative release from the contribution of one release step.

To separate the edited update's direction from its norm, let u_U be the U-anchored update in Equation (15) and u its unedited value. The norm-preserving and radial controls respectively use

$$u_{\text{preserve}} = u_U \frac{\|u\|}{\|u_U\|}, \qquad u_{\text{radial}} = u \frac{\|u_U\|}{\|u\|}. \tag{16}$$

Applied at the selected MLP sites with all later computation live, U produces test CC log-odds changes of +.012/+.603/+.133 in Qwen/Llama/OLMo. Preserving the original update norm retains +.011/+.590/+.118, whereas changing only its norm yields +.001/−.006/+.010. These controls test the complete live edit's dependence on update direction. The following decomposition separately measures final-normalization contributions to the suffix response. Let r_0, r_d, r_a be the final pre-normalization residuals for baseline, direct (later updates clamped), and actual (later updates live) computation, with RMS denominator $s(r) = \sqrt{\text{mean}(r^2) + \epsilon}$. Define $Y(v, s)$ as the response log odds computed from numerator v and denominator s, retaining the learned normalization weights, the model's output head, and nonlinear alias pooling. The direct effect compares the clamped edit with baseline; the remaining terms separate the content and normalization changes produced by later computation:

$$\Delta_{\text{direct}} = Y(r_d, s(r_d)) - Y(r_0, s(r_0)), \tag{17}$$

$$\Delta_{\text{content}} = Y(r_a, s(r_d)) - Y(r_d, s(r_d)), \tag{18}$$

$$\Delta_{\text{norm}} = Y(r_d, s(r_a)) - Y(r_d, s(r_d)), \tag{19}$$

$$\begin{aligned}\Delta_{\text{interaction}} = {} & Y(r_a, s(r_a)) - Y(r_a, s(r_d)) \\ & - Y(r_d, s(r_a)) + Y(r_d, s(r_d)).\end{aligned} \tag{20}$$

Their sum is the actual effect $Y(r_a, s(r_a)) - Y(r_0, s(r_0))$. This accounting tests whether the denominator alone reproduces the suffix effect; the terms are not unique causal mediation fractions because numerator and denominator depend on the same residual. Table 24 reports the native U decomposition.

G Component contributions during correction

G.1 Distinguishing gain from restoration

For Qwen/Llama MLP update Wz, selected coordinates S are fixed at 256/64 MLP31 coordinates. A gain uses $z'_S = (1+g)z_S$, $g \in \{-.25, +.25\}$. Restoration retains upstream edit u but sets $z'_S = z^0_S$ and $z'_{\bar{S}} = z^u_{\bar{S}}$, where 0 denotes the same input's unmodified run. Before finite precision, gain changes the update by $gW_S z^0_S$; restoration removes $W_S(z^u_S - z^0_S)$. The vectors differ, permitting opposite CC effects without a contradiction.

For OLMo the operations use the actual post-normalization MLP19 update x and fixed state axis d:

$$x' = Q[x + g(x^\top d)d], \tag{21}$$

$$x' = Q[x^u + ((x^0 - x^u)^\top d)d], \tag{22}$$

where Q is BF16 rounding. The gain scales the existing signed coordinate; it does not set the person to U or guarantee movement toward μ_U. All subsequent computation runs after either operation. The upstream edit is Qwen L31H14 scaling, Llama U at MLP19, or OLMo's question-specific Head controller on layers 10–14. OLMo restoration therefore tests the state coordinate's response to a Head correction, separate from its earlier U-edit path test. Mean input CC log odds and population loss are separate outcomes: their signs need not equal the change in the log odds of the mean probability.

Table 24: **Content and final-normalization accounting for U.** Changes are mean per-input log odds. Direct, suffix content, suffix norm, and suffix interaction sum to the actual change, up to numerical precision. MID is conditional on directional answers. Totals are bold.

Model	Component	Δ CC/non-CC	Δ MID/directional
Qwen	direct	0.01210	0.01054
Qwen	suffix content	0.00000	0.00000
Qwen	suffix norm	0.00000	0.00000
Qwen	suffix interaction	-0.00000	-0.00000
Qwen	**total**	**0.01210**	**0.01054**
Llama	direct	0.01993	0.00098
Llama	suffix content	0.54165	0.13523
Llama	suffix norm	0.04774	0.01667
Llama	suffix interaction	-0.00618	-0.00252
Llama	**total**	**0.60314**	**0.15035**
OLMo	direct	0.01358	0.01507
OLMo	suffix content	0.14324	-0.03817
OLMo	suffix norm	-0.02536	-0.00118
OLMo	suffix interaction	0.00107	0.00042
OLMo	**total**	**0.13253**	**-0.02386**

Tables 25–26 report the values underlying Figure 3. Here $\ell_{CC} = \log[p(CC)/(1 - p(CC))]$ is computed per input before averaging. Llama's positive restoration contribution is specific to mean log odds and CC CE: the mean CC-probability difference is approximately zero. Thus the small positive R should not be read as a comparable increase in population CC.

Table 25: **Scaling ordinary component activity.** Changes are relative to the same unedited inputs, averaged over all four test layouts. Brackets are pointwise 95% within-domain question-bootstrap intervals. Probability and MAE changes are pp; CE is nats. These are fixed $\pm 25\%$ gains, not optimized doses. Bold marks the CC log-odds and CC CE estimates, not significance.

Model Gain	$\Delta\ell_{CC}$	ΔCC (pp)	$\Delta\mathrm{CE}_{CC}$	$\Delta\mathrm{MAE}_{MID}$ (pp)	ΔTV
Qwen	**-.205**	-.290	**.0037**	.516	.0199
-25%	$[-.211, -.198]$	$[-.314, -.269]$	$[.0029, .0044]$	$[.453, .571]$	$[.0184, .0215]$
Qwen	**.186**	.311	**-.0028**	-.463	-.0178
$+25\%$	$[.180, .192]$	$[.290, .333]$	$[-.0035, -.0021]$	$[-.517, -.406]$	$[-.0196, -.0160]$
Llama	**-.061**	-.045	**.0015**	-.048	.0008
-25%	$[-.063, -.059]$	$[-.054, -.035]$	$[.0012, .0018]$	$[-.085, -.011]$	$[.0004, .0011]$
Llama	**.069**	.052	**-.0017**	.045	-.0012
$+25\%$	$[.066, .071]$	$[.042, .063]$	$[-.0019, -.0014]$	$[.010, .081]$	$[-.0015, -.0009]$
OLMo	**-.058**	-.038	**.0016**	.389	.0030
-25%	$[-.065, -.051]$	$[-.043, -.033]$	$[.0012, .0021]$	$[.309, .471]$	$[.0022, .0038]$
OLMo	**.057**	.040	**-.0016**	-.408	-.0031
$+25\%$	$[.051, .063]$	$[.034, .045]$	$[-.0020, -.0012]$	$[-.492, -.326]$	$[-.0039, -.0023]$

G.2 What the fixed controls qualify

Opposite, random, and two orthogonal displacements target the norm of the U-induced displacement $\delta_U = u_U - u$ at each input. The two fixed seeds are 1701/1702; no output or human label enters

Table 26: **A component's response under a retained upstream edit.** Live effects compare the edit with baseline. R compares live with component-restored outputs; its sign is interpreted relative to the live edit. Each column reports its own outcome, rather than converting mean log odds into mean probability. Brackets are pointwise 95% question-bootstrap intervals over the four test layouts. Bold highlights the live and restoration log-odds point estimates for sign comparison.

Fixed upstream edit	Live $\Delta\ell_{CC}$	R (ℓ_{CC})	R (CC, pp)	R (CE_{CC})
Qwen: H14 $\times.75$	**.135** $[.128, .143]$	**-.0518** $[-.0555, -.0478]$	-.118 $[-.141, -.098]$	.0005 $[.0003, .0008]$
Qwen: H14 $\times1.25$	**-.107** $[-.114, -.100]$	**.0678** $[.0633, .0720]$	.108 $[.089, .131]$	-.0011 $[-.0014, -.0008]$
Llama: MLP19 U	**.603** $[.550, .660]$	**.0114** $[.0081, .0149]$	.000 $[-.002, .001]$	-.0002 $[-.0003, -.0001]$
OLMo: Head controller	**.676** $[.552, .799]$	**-.0375** $[-.0603, -.0171]$	-.049 $[-.077, -.023]$	.0010 $[.0004, .0018]$

numerical norm matching. Module-output and post-addition displacement checks require 1% relative or 10^{-7} absolute tolerance. For nonzero displacements, the opposite-direction check requires $\cos(\delta_{\mathrm{opp}}, -\delta_U) \geq .99$; the orthogonal check requires $|\cos(\delta_{\mathrm{orth}}, d)| \leq .02$. Tolerance is unchanged after observing BF16 failures. The full-sample outcome compares delivered operations; geometry-qualified subsets are conditional diagnostics, not reweighted human-fidelity claims. Llama meets full matching; OLMo reaches approximately 95–96% for the native direction comparisons.

Gain/restoration controls use count-matched disjoint coordinate sets in Qwen/Llama and orthogonal axes in OLMo. Positive selected gains improve CC CE across all order/domain means, although Qwen control 1701 has slightly lower CC CE (.16320 versus .16353); the selected gain has lower MID MAE and TV. The negative selected gain reverses the global mean signs; Llama order D is a small CC-CE exception. Selected-versus-control restoration passes joint module/post-addition checks on only 7–61% of inputs, restricting specificity comparisons. This range uses the intersection of passing inputs for the selected restoration and both control seeds, as required for the contrast against their average. It differs from coverage for either seed alone. No such comparison is needed to define the selected live-versus-restored contribution itself. Table 27 reports the fixed-direction outcomes with matching coverage; the restoration-specific 7–61% coverage is a separate comparison.

The restoration signs in Results persist in development A, while other metrics can differ: Qwen restoration's TV effect reverses between development and the four-order test mean. Llama's positive restoration contribution is small relative to U's total effect, and OLMo's contribution has the opposite sign to its upstream Head correction. These cases reject a universal positive-transmission account and support the stated distinction between ordinary contribution and induced response. They do not estimate a natural mediated fraction or exhaust the causes of CC scarcity.

H Detailed Discussion

H.1 A functional account of scarce non-choice

The findings connect scarce explicit non-choice to response concentration and to the computation that realizes its correction. Across three models, a redistribution fitted to human directional entropy, without fitting CC or MID targets, closes 58–91% of the average CC gap. Different internal controls access a related redistribution, part of which can be predicted from changes in directional scores alone. Fixed-path interventions then show how subsequent computation builds or counteracts that change. Together, these results explain a tractable part of non-choice fidelity: its dependence on probability allocation and on the network's response to an internal correction.

The contribution lies in the measured connection between these computations and human non-choice. Score scaling describes an imposed output transformation, while confidence regulation

Table 27: **Fixed U displacement against alternative directions.** Differences are U minus the named control, with paired 95% question-bootstrap intervals. Lower losses favor U. Random and Orth. 1 use seed 1701; Orth. 2 uses 1702. Module/Both give percentages passing module-output and joint module/post-addition norm checks. Full-sample outcomes include every input. Bold highlights the primary CC CE contrast and joint matching coverage, not significance.

		Loss difference (U − control)			Passing (%)	
Model	**Control**	CC CE ↓	MID MAE ↓ (pp)	TV ↓	Module	Both
Qwen	opposite	**-0.0004** $[-0.0005, -0.0002]$	-0.101 $[-0.167, -0.035]$	-0.0003 $[-0.0011, 0.0007]$	99.9	**58.5**
Qwen	Random	**-0.0002** $[-0.0003, -0.0002]$	-0.082 $[-0.113, -0.051]$	-0.0009 $[-0.0013, -0.0004]$	99.7	**67.2**
Qwen	Orth. 1	**-0.0001** $[-0.0002, -0.0001]$	-0.003 $[-0.035, 0.030]$	-0.0007 $[-0.0011, -0.0002]$	98.9	**63.0**
Qwen	Orth. 2	**-0.0003** $[-0.0004, -0.0002]$	-0.027 $[-0.065, 0.013]$	-0.0004 $[-0.0008, 0.0001]$	99.5	**66.9**
Llama	opposite	**-0.0321** $[-0.0373, -0.0271]$	-1.340 $[-1.769, -0.952]$	-0.0486 $[-0.0575, -0.0405]$	100.0	**100.0**
Llama	Random	**-0.0178** $[-0.0209, -0.0149]$	-0.456 $[-0.629, -0.290]$	-0.0244 $[-0.0290, -0.0201]$	100.0	**100.0**
Llama	Orth. 1	**-0.0101** $[-0.0120, -0.0083]$	-0.680 $[-0.945, -0.434]$	-0.0206 $[-0.0249, -0.0167]$	100.0	**100.0**
Llama	Orth. 2	**-0.0201** $[-0.0235, -0.0169]$	-0.699 $[-0.973, -0.443]$	-0.0283 $[-0.0333, -0.0237]$	100.0	**100.0**
OLMo	opposite	**-0.0072** $[-0.0091, -0.0055]$	-1.541 $[-1.900, -1.230]$	-0.0128 $[-0.0168, -0.0092]$	100.0	**94.8**
OLMo	Random	**-0.0038** $[-0.0048, -0.0029]$	-1.034 $[-1.299, -0.812]$	-0.0064 $[-0.0090, -0.0039]$	100.0	**95.9**
OLMo	Orth. 1	**-0.0007** $[-0.0011, -0.0004]$	-0.689 $[-0.842, -0.555]$	-0.0034 $[-0.0049, -0.0020]$	100.0	**96.4**
OLMo	Orth. 2	**-0.0024** $[-0.0031, -0.0018]$	-0.808 $[-1.037, -0.606]$	-0.0041 $[-0.0064, -0.0019]$	99.9	**96.1**

and compensatory responses have established mechanistic precedents (Stolfo et al., 2024; Kumaran et al., 2026; McGrath et al., 2023). Those principles alone leave open how much of a human discrepancy the change recovers, how it affects MID, and how it arises under internal intervention. Here, entropy-fitted rescaling recovers substantial CC mass while moving opposite MID biases toward their human targets; internal controls reproduce part of the same response pattern. The causal tests connect these corrections to activity supporting ordinary CC: it can oppose a correction in Qwen and OLMo, while Llama's contribution is small and positive in mean log odds. This supplies a partial functional explanation of the observed deficit and its correction, with the origin of the baseline concentration still unresolved.

H.2 A SHARED PROCESS WITH DIFFERENT INTERNAL REALIZATIONS

The commonality is the relationship between redistribution and subsequent computation, with different sites and signed effects across models. Qwen's following MLP partly reverses a useful head-induced redistribution. Llama and OLMo instead realize much of a state-derived correction through later computation; OLMo's state coordinate also opposes part of a separate Head correction. These observations specify where the shared account varies with the intervention and operating point. The optimized controllers establish available correction, while fixed state-derived edits isolate conditional contributions along specified paths. Their common response diagnostic connects the experiments without attributing an entire optimized controller's gain to a separately tested edit.

A component's ordinary contribution to CC does not determine its response to a correction. Increasing existing activity and restoring an induced change perturb different quantities, $W_S z_S^0$ and $W_S(z_S^u - z_S^0)$. The gain and restoration results show why that distinction matters for human fidelity: CC-supporting activity can transmit a correction or counteract it. Useful control therefore depends on both the targeted contribution and the computation that follows. Different effective steering operators are compatible with this shared functional relationship; the experiments do not isolate architecture itself as the cause of their differences.

H.3 Implications for human simulation

Recovering non-choice frequency and recovering its allocation across questions are distinct achievements. Entropy-fitted rescaling closes much of the mean CC gap, yet its itemwise changes correlate negatively with the required corrections. Thus a closer overall rate can coexist with an inaccurate pattern of which questions and population cells elicit non-choice. Country and joint-cell variation, together with OLMo's domain-level TV costs, makes the same point at finer resolution. Evaluating synthetic respondents consequently requires CC, MID, and full-distribution outcomes at supported levels of aggregation, alongside the population average. Sparse persona fields and rare CC responses limit the conclusions available from individual matches.

OLMo's midpoint behavior also sharpens the interpretation of the shared deficit. Its abundant MID rules out a uniform shortage of non-directional responses across these models, while its scarce CC retains the explicit non-choice finding. Separating the two outcomes reveals how models allocate probability among different ways of expressing limited directional commitment. Opinion probes characterize information recoverable from hidden states (Jahanparast et al., 2026); our experiments examine how response-related activity affects expression. Low CC alone cannot establish whether a model lacks information about human opinion absence.

The downstream diagnosis gives a concrete criterion for assessing interventions: measure the consequence of allowing later computation to respond. Clamping preserves more of Qwen's useful head effect but removes much of Llama/OLMo's state-derived benefit. This motivates testing control strategies that account for that signed response, whose practical advantage remains to be established. Directional-entropy-fitted rescaling already matches or exceeds the selected internal controls on aggregate test CC CE, MID MAE, and TV, making it a strong practical reference when human development distributions are available. The internal experiments contribute an explanation of how particular corrections are realized and why their local and final effects differ.

H.4 Future research questions

CC and MID only partially identify human opinion absence: choice difficulty, neutrality, ambivalence, and low interest can yield overlapping survey responses. Joint measurements of opinion presence, familiarity, and ambivalence would allow future research to test which of these human states is associated with the computational response identified here.

The traced paths establish conditional effects at selected sites and doses; exploratory selection, competitive random controls, incomplete norm matching, and substantial forecast residuals constrain claims about unique components or a complete natural cause. A next mechanistic test would freeze the proposed path and predict when its subsequent response should aid or oppose CC across new inputs and doses, distinguishing the general redistribution from response-specific residuals.

The main evaluation covers three 7–9B models, English prompts, and new people on known questions: same-layout gains persist across people, whereas changing answer order weakens them, and existing transfer results are limited or mixed. Frozen control on new questions would test the unresolved generalization of the account, especially whether it can correct the allocation of non-choice across items as well as its average frequency.

Broader model scales and training stages would extend the computational scope of these path tests. Comparing matched checkpoints before and after instruction tuning could distinguish when the concentration pattern emerges from how later layers respond to an edit. Larger site and dose ranges could then test whether the observed downstream responses persist across operating points. These extensions would address the origin and reach of the present functional account while retaining its separation of ordinary component activity from the response induced by an intervention.