RIGHT ON AVERAGE, WRONG PER RESPONDENT: A MECHANISTIC DIAGNOSIS OF LLM-SIMULATED CHOICE

Anonymous authors
Paper under double-blind review

ABSTRACT

Language models are increasingly used as synthetic survey respondents, and what those applications need is how a choice moves when an attribute changes, not how plausible a single answer looks. Existing evaluations score the level of an answer and not its response to a change. We introduce MagBench, which perturbs one attribute of a choice situation at two sizes, the smaller half the larger, on five datasets of real human choices, and asks whether the response scales in proportion. In aggregate it does. For an individual respondent it does not, and most of that respondent's movement has nothing to do with the size of the change. The aggregate is right because those respondent-specific parts cancel, and what cancels is not noise but a displacement that reappears at both sizes of change and under a different prompt. Inside the models the size of the change is decodable and reaches the subspace that carries the response, and neither fact predicts behaviour. Writing the displacement one respondent's perturbation produces into another respondent's run moves the response to the donor and not the recipient, at 0.89 against -0.03 over 67 runs and five model families. It holds even in a model whose choices follow the human direction 51 percent of the time, so whether this structure is present and whether the behaviour it should drive is valid are separate questions. Encoding a quantity is not using it.

1 INTRODUCTION

Language models are increasingly used as proxies for human respondents in surveys, marketing, and policy analysis (Argyle et al., 2023; Horton et al., 2026; Aher et al., 2023; Huang & Wang, 2026). What those applications read off a simulated population is usually not the answer to a fixed question, but how that answer changes when an attribute such as price or travel time changes, so a simulated population is useful only if its responses move by the right amount. Current evaluations mostly assess answer levels, for example demographic alignment, calibration, or stability across prompts and models (Tigre & Souto, 2026; Maier et al., 2025; Lin et al., 2025). The one place this has been measured, simulated opinions move much further than human ones after the same information (Wali & Tayyab, 2026), but that result stops at the behaviour and says nothing about where the mismatch comes from inside the model.

Testing whether a simulated respondent moves by the right amount needs a change whose correct response is known for that respondent and not only for the population. We perturb one attribute of a real choice situation at two sizes, the smaller half the larger, so that a response function fitted to the real choices predicts a ratio of one half between the two responses whatever weight a respondent places on that attribute. MagBench applies this to five datasets of real human choices. In aggregate the simulated response is close to the human reference. For an individual respondent it is not, and most of that respondent's movement has nothing to do with the size of the change, so the aggregate agrees because the respondent-specific parts cancel. What cancels is not noise. It displaces the same respondent in the same direction at both sizes and under a different prompt architecture, and its own small-to-large ratio is 0.80 where proportionality gives 0.50 and noise would give 1. An aggregate that matches is not evidence that the respondents match.

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.

We then ask where that part comes from, in open-weight models we can intervene on. Two natural explanations fail. The size of a change is linearly decodable from the residual stream, and it reaches the low-rank subspace that carries the response, and neither fact predicts how much of a respondent's movement the size accounts for. What does account for it can be located. The residual difference a perturbation produces for one respondent, written into the unperturbed run of another, moves the response to the first respondent and not the second, and a single direction inside that difference carries almost all of it. This holds across depths, across perturbed attributes and across model families, including models whose choices barely respond to an attribute at all, which separates whether a model carries this structure from whether the behaviour it should drive is valid. Since the later layers reproduce the response from the difference faithfully and close to linearly, the failure to scale is produced before the layer we probe, in how a number in the prompt becomes a displacement, and not after it.

Our contributions are threefold. We introduce MagBench as a targeted test of response scaling in simulated respondents. We separate a respondent's movement into a part proportional to the size of the change and a reproducible part that is not, and show that aggregate agreement is produced by the second part cancelling rather than by the respondents being right. And we locate that second part causally, in a residual difference that transfers between respondents over 67 runs and five model families, including models whose behaviour is at chance, which places the limitation before the layer we probe and separates the presence of this structure from the validity of the behaviour it should drive.

2 Related work

What is measured in simulated respondents. Work on simulated respondents asks whether a model reproduces who people are and what they answer, through covariate overlap and stability across models (Tigre & Souto, 2026), paired human data (Toubia et al., 2025), elicitation for purchase intent (Maier et al., 2025) and alignment on survey preferences (Lin et al., 2025). Comparisons of the size of a change do exist, and concurrent work reports that simulated opinions move considerably further than human opinions after the same information (Wali & Tayyab, 2026). What has not been done is to treat the response to a numeric attribute as the object of audit, against a human response function estimated from choices in which those attributes were varied by design, and to ask where inside the model the size is lost. Meanwhile the property goes unvalidated where it is used (Huang & Wang, 2026).

Mechanistic background. Interpretability work on personas targets personality, values and style (Courtis & Hu, 2026; Beckmann & Butlin, 2026; Saini et al., 2026; Wang et al., 2025) rather than quantitative decision behaviour, and safety work shows that behaviour can look acceptable while the representation says otherwise (Jiang et al., 2026). Numbers are not embedded as a continuous quantity, comparison running on digit-level and unit-specific heuristics (Davies et al., 2025; Sasaki et al., 2026; Ni, 2026), which supplies a prior for what we find, since a size encoded piecewise can be readable by a probe without being used proportionally. Our method follows the dissociation between where information can be located and where intervention changes behaviour (Hase et al., 2023; Elazar et al., 2021), and the low-rank direction template with an intervention and a control (Arditi et al., 2024) together with the critique that faults arbitrary dimensionality, weak controls and failure across models (Joad et al., 2026). We answer those three in advance. The rank of our subspace is set by the spectrum of the response, our control is a norm-matched orthogonal complement alongside random directions of the same rank, and the one configuration that behaves differently across models fails a scope condition stated before the analysis.

3 Experimental setup

3.1 MagBench

Attributing a change in a simulated choice to the size of an attribute change means holding everything else fixed, which language models make harder than it looks. The same respondent must face the same options with the same values, the answer must not depend on which letter an option was given, and the response must not be one sample. MagBench therefore reads the model's probabilities

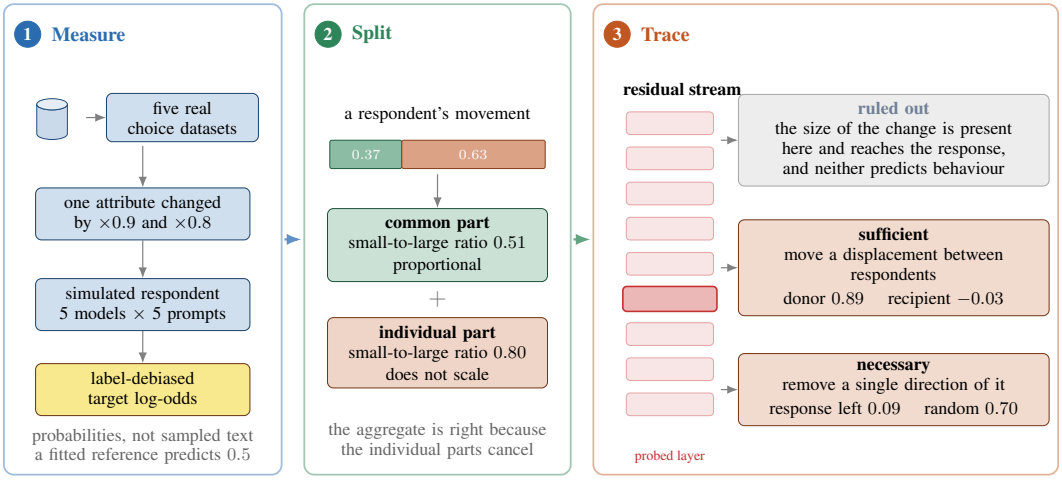


Figure 1: The measurement, the split it produces, and what the split is made of inside the model. One numeric attribute of one option is changed at two sizes, the smaller half the larger, so a response function fitted to the real choices predicts a ratio of one half between the two responses whatever the weight of that attribute. Against that target the part of a respondent's movement that is common to respondents is proportional and the part specific to that respondent is not, and the bar gives their shares. At the probed layer the size of the change is present and reaches the directions that carry the response, and neither fact predicts behaviour, so the displacement a perturbation produces is acted on instead.

over the option labels rather than generating text, asks every choice under each cyclic relabelling of the options, and subtracts a letter offset estimated across relabellings (Appendix B). Between the two responses being compared, one attribute of one option differs and nothing else does, by 10 or 20 percent of its value and by 25 and 50 percent for headway, so the smaller change is always half the larger.

The harder problem is what to compare against, since the human counterfactual for a perturbation of a given size is not observed in any of these datasets. A softmax regression fitted to the real choices in each dataset predicts, for respondent i and attribute a under multiplier m, a change in the target log-odds,

$$\Delta^{\mathrm{ref}}_{i,a}(m) \;=\; \beta_a \frac{x_{i,a}}{\sigma_a}\,(m-1), \qquad \Delta_{i,a}(m) \;=\; y_{i,a}(m) - y_i(1), \tag{1}$$

against which the model's own change is set. What makes this usable is the ratio rather than the level, since under the reference the response to the small change is one half of the response to the large one for every respondent and every β, so the target is identified without trusting the estimated weight. Two sizes rather than three follow from the same argument. The reference is therefore local and operational rather than a measured human counterfactual, and not equally good everywhere, electricity failing both a sign check and a range check (Appendix A). Because a model can also fail this test for uninteresting reasons, configurations enter the analysis through a preregistered ladder fixed before the confirmatory runs: L0 asks whether choices follow an option's content rather than its letter position, L1 whether the change has the sign of the reference in at least 65 percent of perturbations, and L2 whether the ratio of the two mean absolute responses, divided by the same ratio for the reference, is at most 1.25. The unit throughout is the configuration, a combination of dataset, model and prompt architecture, and not the individual response.

3.2 Models and prompts

Five instruction-tuned open-weight models from three families span an order of magnitude in size, Qwen2.5 at 7B, 14B and 32B, Mistral-7B-Instruct-v0.3 and Phi-3-mini, each under five prompt architectures that differ in how the respondent is framed before the choice (Appendix I). Of the 125

Table 1: Human choice datasets. SP: stated preference. RP: revealed preference. Rows per configuration equal respondents × 5 states × options.

domain	type	options	target option	perturbed attributes	respondents	rows
swissmetro	SP	3	Swissmetro	cost, travel time, headway	300	4500
electricity	SP	4	supplier 1	price	100	2000
heating	RP	5	heat pump	installation cost, operating cost	200	5000
catsup	RP	4	Heinz 41 oz	price	100	2000
fishing	RP	4	charter	price, catch rate	200	4000

possible configurations 92 were run, the gaps fixed by prompt availability and memory before any result was read, for 329,500 readouts. Three subsets carry different claims and we name them so the denominators are explicit. The 67 that pass L0 carry the behavioural results. Seventeen were selected for the mechanistic analysis to span the families, the size ladder and three datasets, and also to include the configurations closest to proportionality, which means that sample is not independent of the behaviour it is used to explain (Section 5.3). One of the 17 fails L0, leaving 16 wherever mechanism is related to behaviour.

3.3 Probes and patching

The object of every internal measurement is a difference rather than a state, because the question is about a response to a change. For each respondent we take the residual stream at the last prompt token under the perturbed prompt and subtract the same vector under the unperturbed one, which gives a displacement d_i per respondent, and everything below reads or writes that displacement at one layer. Respondents are split by identity into two thirds for fitting and one third held out, and every number reported is computed on the held-out third, so a direction or a subspace is never evaluated on the respondents that defined it.

What the probe regresses on is unusual and is what makes its failure informative. The target is not a label or a concept but the change in the target log-odds that the human response function of Section 3.1 assigns to that respondent and that perturbation, so a probe that scores well has recovered a behavioural counterfactual and not a category. That target carries the direction of the change as well as its size, and a probe that has learned only whether the attribute improved or worsened would already score well on it, which is why the raw R^2 is reported beside a within-direction R^2, fitted and scored separately within each direction, and an ordering accuracy whose chance level is one half. Probes are fitted at five depths per model and at every second layer in four configurations, and the layer at which the held-out R^2 peaks is the layer used by every intervention.

Reading a quantity off activations does not show that the model uses it, so the causal arms write instead of read. Principal components of d_i are computed on the training respondents, and part of a held-out respondent's own displacement is added back into its unperturbed run at that layer. Recovery is the correlation across held-out respondents between the change in the readout this produces and the model's own change, and three controls decide what it means. The orthogonal complement of a subspace says whether the recovered response is specific to that subspace rather than to the amount of activation injected, random directions of matched rank say whether any subspace of that size would do, and the displacement taken from a different respondent turns the same machinery into an interchange intervention. That last arm is the one the paper leans on, and it differs from the usual form in what is exchanged: not a representation computed on another input, but the displacement another input's perturbation produced, which is why its response can be compared against the donor's own response and against the recipient's and the two can come apart (Appendix E).

4 Results

4.1 Most of a respondent's movement is not about the size of the change

Simulated respondents follow the direction of a change more often than chance and its size almost never. Of the 92 configurations, 67 pass L0. Write a respondent's aligned response as a part common to respondents, which is what an aggregate reads off, plus a part specific to the respondent.

Table 2: The behavioural grid by model. Ladder columns are counts passing each level of Section 3.1, L2 corrected scoring each attribute against its own reference-implied target and excluding electricity. Shading is the systematic share. *Invalid reference.

model	n	L0	L1	L2	L2 corrected	systematic share
Phi-3-mini	15	8	1	0	0	0.17
Mistral-7B	25	17	3	0	0	0.31
Qwen2.5-7B	25	18	0	1*	0	0.34
Qwen2.5-14B	17	14	2	0	0	0.42
Qwen2.5-32B	10	10	5	3	2	0.72

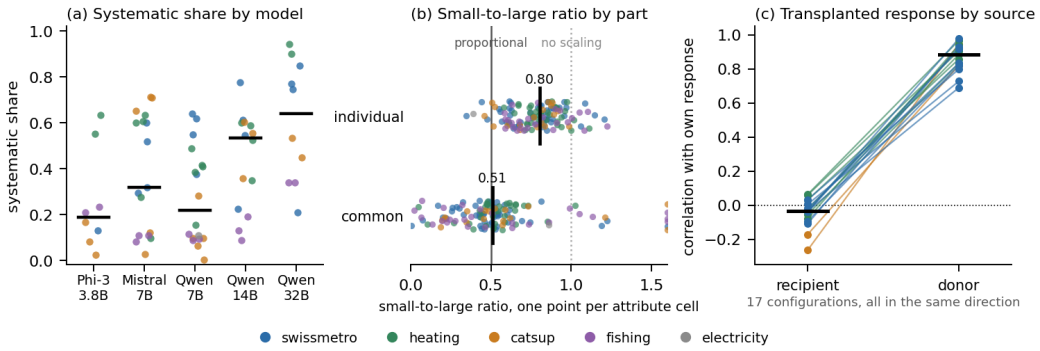


Figure 2: (a) The share of a respondent's movement that the size of the change accounts for, one point per configuration, with each model's median as a bar. (b) The same movement split into a part common to respondents and a part specific to each one. The common part sits on the proportional value of one half, the individual part does not, and neither sits at 1, where pure noise would be. (c) The transplanted response against the recipient's own response and against the donor's, one line per configuration. Colour is the dataset throughout.

The common part is proportional, at a small-to-large ratio of 0.509 against a target of 0.500, and a ratio-free contrast on the same responses has a median of -0.005 with an interval containing zero in 108 of the 132 attribute cells. The respondent-specific part has a ratio of 0.80 and is the larger of the two (Figure 2b). The size of the change accounts for 0.37 of a typical respondent's movement at the large change and 0.22 at the small one, a fraction we call the systematic share. The aggregate is proportional because those respondent-specific parts cancel, so a correct aggregate is not evidence that any individual response is correct.

The respondent-specific part is neither noise nor preference heterogeneity. Noise would give a ratio near 1 and would not reproduce. Heterogeneity in how much a respondent cares about the attribute would give a slope of 2 when the large-change displacement is regressed on the small-change one, and the measured slope is 0.61. The displacement does reproduce, pushing the same respondent in the same direction at both sizes of change (median correlation 0.47) and under a different prompt architecture (median 0.44 over 90 pairs, with Mistral-7B the exception at 0.18), and nothing is sampled. Throughout we first remove what the human reference predicts for that respondent, so heterogeneity the reference itself contains is not counted (Appendix C).

The systematic share rises from 0.17 for Phi-3-mini to 0.72 for Qwen2.5-32B and direction follows it, mean sign agreement rising within Qwen2.5 from 0.533 to 0.690 and exceeding chance in 61 of the 67 configurations (Table 2, Figure 2a). Failures concentrate by attribute more widely than by model, from 0.064 for the fishing catch rate to 0.684 for the heating installation cost. The preregistered linearity index compares how far respondents move without regard to direction, which lets the part that does not scale enter both terms and pushes the index towards 2. Its median is 1.70 and it is close to the inverse of the systematic share ($\rho = -0.80$), so we report it because it was preregistered and recommend the decomposition instead. The same mixing explains why headway, the one attribute perturbed by 25 and 50 percent rather than 10 and 20, has the best index of any attribute. Telling the model the size of the change repairs neither part (Appendices B, G and H).

Table 3: The transplant across 67 runs, medians throughout. Shading is the absolute correlation, so the donor column is dark and the recipient column is not. A bracketed count marks a subset or a re-analysis of the row above, not counted again. Appendix H gives every configuration and the within-attribute comparison.

condition	runs	faithful	donor	recipient	donor > recipient
main set, probed layer	17	0.995	0.885	−0.034	17/17
Qwen2.5	(13)	0.997	0.915	−0.036	13/13
Mistral-7B	(4)	0.906	0.827	−0.028	4/4
donor and recipient differ in attribute	(15)	—	0.898	−0.188	15/15
above the probed layer, Qwen2.5	24	0.999	0.945	−0.059	24/24
above the probed layer, Mistral-7B	8	0.999	0.969	−0.006	8/8
held out, Llama-3.1-8B	3	0.996	0.953	−0.126	3/3
held out, OLMo-2-7B	3	0.971	0.913	−0.078	3/3
Phi-3-mini, two depths	12	0.928	0.882	−0.010	12/12

4.2 The size is represented and reaches the response, and neither tracks behavioural validity

The size of the change is decodable in mid-network and reaches the directions that carry the response, and neither fact separates the configurations that match human magnitude from those that do not. In all 17 configurations a linear probe recovers the reference change on held-out respondents, peak R^2 between 0.56 and 0.90, with shuffled-target controls at or below zero. A direction-only predictor matches or beats that R^2 in 14 of the 17, and after the control the within-direction R^2 has a median of 0.35 and the ordering accuracy a median of 0.71, above chance in 16 of the 17. The size becomes decodable abruptly in mid-network, peaking between 0.56 and 0.79 of relative depth and weaker at the output than at its peak in all four dense sweeps (Figure 3).

Patching the top eight components recovers 0.87 to 0.99 of what the full difference recovers in all 17 configurations, against 0.05 to 0.48 for their orthogonal complement and -0.08 to 0.19 for random directions of the same rank (Table 5). Inside that subspace the size is decodable at 22 to 85 percent of what the full residual stream gives, so part of the representation never enters the directions that drive the response.

What neither of these facts does is track behaviour. Peak R^2 does not predict the index ($\rho = -0.16$, $p = 0.56$, $n = 16$), nor do the two direction-controlled measures ($\rho = -0.40$ and -0.11), and the configuration with the highest R^2 of 0.901 fails L2. Neither how much of the size enters the response-carrying subspace ($\rho = 0.01$) nor what share of it enters ($\rho = -0.01$) predicts it either. At sixteen configurations these are underpowered nulls, so we read them as the absence of any strong relation rather than as evidence of none. Of six summary measures only one tracks behaviour, the share of the size carried by the first component ($\rho = 0.81$), which we report as exploratory and keep out of the argument (Appendix D).

4.3 The rest is one displacement that transfers between respondents

What accounts for an individual respondent's movement is a single displacement at the probed layer that transfers to another respondent and whose one direction is necessary. At the probed layer we replace the last-token residual of the unperturbed prompt by

$$\tilde{h}_i = h_i(\text{base}) + c\, d_{\pi(i)} \tag{2}$$

and leave every earlier position as it was, reading the response under every cyclic relabelling. With $c = 1$ and π the identity this restores the perturbed state at that position alone and reproduces the model's own response at a median correlation of 0.995, which bounds the artefact of injecting rather than prompting. With π a fixed random permutation one respondent receives another's displacement. Correlations are across held-out respondents, so they measure the respondent-specific part and not the common response.

Under the permutation the injected response correlates 0.885 at the median with the donor's own response and -0.034 with the recipient's, the donor ahead in every one of the 17 main configurations

(sign test $p = 1.5 \times 10^{-5}$, Figure 2c). The displacement is therefore already fixed in the residual difference at this layer, and what the later layers do with it does not depend on the state it is added to.

Transplanting shows that the displacement suffices, and removing part of it shows which part is needed. Estimating a single direction u on the training respondents, as their mean aligned displacement, and removing from each held-out displacement its component along u leaves 0.085 of the average response, while removing the same length along a random orthogonal direction leaves 0.703, the first below the second in 15 of the 17 configurations. That component is 0.463 of a typical displacement's length, so less than half the length carries more than nine tenths of the response, in a direction fitted without seeing the respondents it is tested on. It does not separate the common part from the respondent-specific part, since the length along u is itself respondent-specific.

Table 3 collects 67 runs and the split holds in all of them, at three depths above the probed layer, between respondents whose perturbed attributes differ, and in five model families, two of which entered no analysis. That the cross-attribute pairs transfer as well as the rest makes the displacement a property of the respondent rather than of the attribute that was changed (Appendix H). It is also present where the behaviour is not. Phi-3-mini has sign agreement at chance in six of the eight configurations it passes the lock on, and still gives a median donor correlation of 0.882 against -0.010 over twelve runs, while OLMo-2-7B follows the human direction 51 percent of the time and gives 0.913 against -0.078. A model whose choices barely move when an attribute changes therefore still carries a displacement that can be moved from one respondent to another, so whether this structure is present and whether the behaviour it should drive is valid are separate questions.

Scaling the displacement scales the response, at median slopes of 1.07 at $c = 0.5$ and 2.30 at $c = 0.25$ where 1 and 2 are predicted, and the individual response does not follow. The half-scaled large displacement correlates 0.652 with the model's own small-perturbation response, far below the 0.995 the same injection reaches when the displacement is the right one, so halving d_i is not the vector the small perturbation produces. Since the readout is close to linear in the displacement and close to noiseless, the failure to scale cannot be produced after this layer and is produced before it, in how a number in the prompt becomes a displacement. How it is built there we do not show.

One part of this is weaker than it looks. Near-additivity of the residual difference, which the recovery columns of Table 5 measure directly, already implies that writing a donor's displacement into a recipient moves the readout in the donor's direction. Donor-following is therefore close to what additivity alone predicts rather than independent evidence for it. The ablation is what carries the weight, because no rescaling of a near-additive readout produces an asymmetry between one direction and a random direction of the same length.

5 Discussion

5.1 Which explanation survives

The reference does not account for the gap, since scoring each attribute against its own fitted ratio changes only one of the 67 outcomes, though that does not make the benchmark a human counterfactual, which no dataset here identifies. Two mechanistic explanations then fail. The size of the change is represented and it reaches the directions that causally carry the response, and neither fact predicts behaviour, so reporting probe accuracy alone would have supported a conclusion the behaviour does not. What we locate instead is the larger component. One respondent's residual difference, written into another, moves the response to the first and not the second, one direction inside that difference is necessary for almost all of the average response, and the separation holds at three depths above the probed layer, between respondents perturbed on different attributes, in five model families, and in a model whose sign agreement with the human reference is 0.51. That last case is what we would carry away, because whether a model holds this structure and whether the behaviour it should drive is valid turn out to be independent questions. The later layers are not where the failure lies, since they scale the response with the difference as a proportional readout requires, so the failure to scale is produced before the probed layer, in how a number in the prompt becomes a displacement, and how it is built there we do not show. For a user the reading is simpler. An aggregate that matches is not evidence that the respondents match, and the ordering they would act on moves when the size of the change used to measure it moves.

5.2 What scale changes

Scale moves the systematic share from 0.17 for Phi-3-mini and 0.31 for Mistral-7B to 0.34, 0.42 and 0.72 along the Qwen2.5 ladder, and it moves neither of the two things on either side of that. The representation is already present at 7B and strengthens only modestly, while the share of the response the top eight components carry is at its ceiling throughout and the donor and recipient split holds at every size in the ladder and in the two smallest models in the study. What scale improves therefore lies between a representation that is already there and a readout that already works. The only summary measure that moves with it is where the size sits inside the response, which we treat as exploratory (Appendix D). We state this as consistency rather than demonstration, on one size ladder in one family.

5.3 Limitations

In this study we ask for a property of simulated respondents that is rarely tested, and we ask it of models whose weights we can patch. That choice bounds what the results support. The scale ladder is one model family, every model is open weights at 32B or smaller, and Phi-3 has no specialist prompt and so covers 15 of 25 cells, which leaves the comparison across prompt architectures unbalanced. The electricity reference fails its validity checks and supports no conclusion, and five of nine attributes show curvature, which moves one ladder outcome.

Not every result here carries the same weight. The split by direction was added after the preregistration in response to the signed probe target, and the design cannot separate an asymmetry in the model from an asymmetry in the fitted human weights. The relation between the first principal component and the systematic share is exploratory, since the 17 mechanistic configurations were chosen partly on the behavioural outcome, so the sample is not independent of the outcome it is used to predict, and a larger share on the first component cannot be distinguished from less of anything else being placed there. We use principal components rather than a learned dictionary, so the directions are not claimed to be the model's own features, and on five held-out configurations the relation agrees in sign at $\rho = +0.40$ without reaching significance, in a narrow range of response strength.

What we locate we do not explain. A single direction estimated on other respondents is necessary for almost all of the average response, and we do not show what forms it, since injecting below the probed layer is inert, the residual difference there being near zero. That intervention also does not separate the common part of the response from the respondent-specific part, because the length along the direction it removes is itself respondent-specific. The depth check covers two model families, and the extension to the held-out families rests on small responses for OLMo-2-7B, whose aligned response has a spread of 0.06 to 0.09 against 0.29 to 0.32 for Llama. The decline of probe R^2 after its peak is descriptive, since we do not show that later layers discard the size rather than recode it. The mismatched-unit control is unseeded and varies across reruns, so we omit it from Table 5.

6 Conclusion

In this study we ask whether a simulated respondent's choice moves by the right amount when an attribute changes, on five datasets of real choices and against a reference we check before using. In aggregate it does. For an individual respondent it does not, and only about a third of that respondent's movement is about the size of the change, so the aggregate is right because the individual errors cancel. What is left reproduces across sizes of change and across prompt architectures and does not scale with the stimulus, so it is neither measurement noise nor preference heterogeneity. Two mechanistic explanations of it fail, since the size of the change is represented and reaches the subspace carrying the response and neither tracks behavioural validity. The remainder we can locate. Writing the residual difference one perturbation produces for one respondent into the run of another moves the response to the donor rather than the recipient, a single direction inside that difference carries almost all of the average response, and both hold at every depth above the probed layer, between respondents perturbed on different attributes, and in five model families. Both also hold in models whose choices barely respond to an attribute at all, so the presence of this structure and the validity of the behaviour it should drive are separate questions. How the displacement is built, and whether any of this holds at larger scales, we do not establish. For anyone using simulated respondents to produce a quantity that depends on the size of a response, an average that scales correctly

is not a respondent who responds, and evidence that a model encodes a quantity is not evidence that it acts on it.

REPRODUCIBILITY STATEMENT

Everything needed to regenerate the paper is in the supplementary material, the code, the per-configuration outputs and numbered entry scripts for each table and figure. We rebuilt the evaluation ladder from scratch and it reproduces the preregistered table exactly, and we re-ran two mechanistic configurations end to end, where the probe R^2 and every patching arm reproduce exactly. The one exception is the mismatched-unit control, which is unseeded, varies by up to 0.14 across reruns and supports no conclusion here. The probe split is drawn per configuration, so a rerun with a different seed moves the correlation between probe strength and behaviour within the range we report without changing its lack of significance.

USE OF AI ASSISTANTS

The author designed and ran every experiment. An AI assistant was used to draft and polish prose, to retrieve and triage related work, and as a sounding board for analysis choices. Every number reported here was produced by the released code and checked by the author against its output.

REFERENCES

Gati Aher, Rosa I. Arriaga, and Adam Tauman Kalai. Using large language models to simulate multiple humans and replicate human subject studies. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *ICML'23*, pp. 337–371, Honolulu, Hawaii, USA, July 2023. JMLR.org.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in Language Models Is Mediated by a Single Direction, October 2024. URL http://arxiv.org/abs/2406.11717. arXiv:2406.11717 [cs.LG].

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua Gubler, Christopher Rytting, and David Wingate. Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis*, 31(3):337–351, July 2023. ISSN 1047-1987, 1476-4989. doi: 10.1017/pan.2023.2. URL http://arxiv.org/abs/2209.06899. arXiv:2209.06899 [cs.LG].

Pierre Beckmann and Patrick Butlin. Where is the Mind? Persona Vectors and LLM Individuation, September 2026. URL http://arxiv.org/abs/2604.17031. arXiv:2604.17031 [cs.CL].

David Courtis and Ting Hu. Mechanistic Personality Analysis of LLMs Steering Personality via Latent Feature Interventions, June 2026. URL http://arxiv.org/abs/2606.28770. arXiv:2606.28770 [cs.AI].

Alex O. Davies, Roussel Nzoyem, Nirav Ajmeri, and Telmo M. Silva Filho. Language Models Do Not Embed Numbers Continuously, October 2025. URL http://arxiv.org/abs/2510.08009. arXiv:2510.08009 [cs.AI].

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic Probing: Behavioral Explanation with Amnesic Counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 2021. doi: 10.1162/tacl_a_00359. URL https://aclanthology.org/2021.tacl-1.10/.

Peter Hase, Mohit Bansal, Been Kim, and Asma Ghandeharioun. Does Localization Inform Editing? Surprising Differences in Causality-Based Localization vs. Knowledge Editing in Language Models, October 2023. URL http://arxiv.org/abs/2301.04213. arXiv:2301.04213 [cs.LG].

John J. Horton, Apostolos Filippas, and Benjamin S. Manning. Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?, February 2026. URL http://arxiv.org/abs/2301.07543. arXiv:2301.07543 [econ.GN].

Chengpiao Huang and Kaizheng Wang. LLM-Powered Virtual Population for Demand Simulation and Pricing, June 2026. URL http://arxiv.org/abs/2606.16183. arXiv:2606.16183 [cs.LG].

Enyi Jiang, Anders Gjølbye, Yibo Jacky Zhang, and Sanmi Koyejo. When Behavioral Safety Evaluation Fails: A Representation-Level Perspective, August 2026. URL http://arxiv.org/abs/2606.08044. arXiv:2606.08044 [cs.LG].

Faaiz Joad, Majd Hawasly, Sabri Boughorbel, Nadir Durrani, and Husrev Taha Sencar. There Is More to Refusal in Large Language Models than a Single Direction, September 2026. URL http://arxiv.org/abs/2602.02132. arXiv:2602.02132 [cs.CL].

Chenxi Lin, Weikang Yuan, Zhuoren Jiang, Biao Huang, Ruitao Zhang, Jianan Ge, Yueqian Xu, and Jianxing Yu. AlignSurvey: A Comprehensive Benchmark for Human Preferences Alignment in Social Surveys, November 2025. URL http://arxiv.org/abs/2511.07871. arXiv:2511.07871 [cs.CL].

Benjamin F. Maier, Ulf Aslak, Luca Fiaschi, Nina Rismal, Kemble Fletcher, Christian C. Luhmann, Robbie Dow, Kli Pappas, and Thomas V. Wiecki. LLMs Reproduce Human Purchase Intent via Semantic Similarity Elicitation of Likert Ratings, October 2025. URL http://arxiv.org/abs/2510.08338. arXiv:2510.08338 [cs.AI].

Aoxin Ni. Numeracy in Large Language Models: Fundamental Limitations and Paths to Improvement, August 2026. URL http://arxiv.org/abs/2608.13129. arXiv:2608.13129 [cs.AI].

Harshvardhan Saini, Yiming Tang, and Dianbo Liu. Bridging Mechanistic Interpretability and Prompt Engineering with Gradient Ascent for Interpretable Persona Control, June 2026. URL http://arxiv.org/abs/2601.02896. arXiv:2601.02896 [cs.LG].

Mutsumi Sasaki, Go kamoda, Ryosuke Takahashi, Kosuke Sato, Kentaro Inui, Keisuke Sakaguchi, and Benjamin Heinzerling. Language Models Compare Quantities Using Number-specific and Unit-specific Heuristics, August 2026. URL http://arxiv.org/abs/2606.03982. arXiv:2606.03982 [cs.CL].

Robson Tigre and Hugo Gobato Souto. When Can You Trust Your Synthetic Users? Diagnostics and Corrections for LLM Consumer Panels, June 2026. URL http://arxiv.org/abs/2609.13148. arXiv:2609.13148 [cs.HC] version: 1.

Olivier Toubia, George Z. Gui, Tianyi Peng, Daniel J. Merlau, Ang Li, and Haozhe Chen. Twin-2K-500: A dataset for building digital twins of over 2,000 people based on their answers to over 500 questions, May 2025. URL http://arxiv.org/abs/2505.17479. arXiv:2505.17479 [cs.CY].

Ahmed Wali and Hassaan Tayyab. Before You Poll with LLMs: A Deliberative Diagnostic Framework, September 2026. URL http://arxiv.org/abs/2609.15849. arXiv:2609.15849 [cs.CL].

Anyi Wang, Dong Shu, Yifan Wang, Yunpu Ma, and Mengnan Du. Improving LLM Reasoning through Interpretable Role-Playing Steering, September 2025. URL http://arxiv.org/abs/2506.07335. arXiv:2506.07335 [cs.CL].

A VALIDITY OF THE HUMAN REFERENCE

Two assumptions licence the target of one half. The human response must be linear in the attribute over the perturbed range, and the perturbed values must stay inside the range observed in the data.

Table 4 reports both checks for each of the nine dataset and attribute pairs. A quadratic term improves the fit for five of the nine, which is expected with many observed choices and does not by itself invalidate the target, since what matters is the ratio the fitted curve implies over the perturbed range, and that stays between 0.453 and 0.498. The corrected-target analysis of Section 3.1 uses those implied ratios for the five and leaves the target at one half for the rest. Exactly one ladder outcome changes, swissmetro with Qwen2.5-32B and the specialist prompt, whose index moves from 1.244 to 1.263 and crosses the threshold. Electricity is the one reference we discard. Half its perturbed prices fall outside the observed range and its fitted price weight is positive, so that reference says a higher price makes an option more attractive. It is retained under the preregistered target only so that the preregistered numbers are reported as preregistered.

Table 4: The two validity checks, by dataset and attribute. An implied ratio is given only where a quadratic term improves the fit, and the target stays at one half elsewhere. *Positive fitted price weight, reference treated as invalid.

dataset	attribute	curvature	implied ratio	values outside observed range
swissmetro	cost	yes	0.485	$\leq 10\%$
swissmetro	travel time	yes	0.493	$\leq 10\%$
swissmetro	headway	no	—	22%
heating	installation cost	no	—	$\leq 10\%$
heating	operating cost	no	—	$\leq 10\%$
catsup	price	yes	0.453	$\leq 10\%$
fishing	price	yes	0.488	$\leq 10\%$
fishing	catch rate	yes	0.498	$\leq 10\%$
electricity	price*	no	—	$\approx 50\%$

B MEASURING PROPORTIONALITY

Write Π for the K cyclic relabellings of the K options, $p_\pi(\ell)$ for the probability the model puts on letter ℓ under relabelling π, and $b(\ell)$ for the offset of letter ℓ, estimated as the mean centred value of that letter over all states and relabellings. For option o,

$$\tilde{z}_\pi(o) \;=\; \log p_\pi(\pi(o)) - \frac{1}{K}\sum_{o'} \log p_\pi(\pi(o')) - b(\pi(o)), \qquad q(o) \;=\; \mathrm{softmax}_o\Big(\tfrac{1}{K}\textstyle\sum_{\pi\in\Pi} \tilde{z}_\pi(o)\Big), \tag{3}$$

and $y = \log q(t)/(1-q(t))$ for the target option t is the readout used throughout. Between the two responses being compared, the respondent, the choice situation, every attribute of every other option, every other attribute of the target option, the model, the prompt architecture, the decoding and the label positions are all identical, and the ratio of small to large is fixed at one half.

The preregistered index compares the two mean absolute responses against the same comparison for the reference,

$$\lambda_a \;=\; \frac{\mathrm{E}_i\big|\Delta_{i,a}(m_s)\big|}{\mathrm{E}_i\big|\Delta_{i,a}(m_l)\big|} \Big/ \frac{\mathrm{E}_i\big|\Delta^{\mathrm{ref}}_{i,a}(m_s)\big|}{\mathrm{E}_i\big|\Delta^{\mathrm{ref}}_{i,a}(m_l)\big|}, \qquad \frac{\mathrm{E}_i|\Delta^{\mathrm{ref}}_{i,a}(m_s)|}{\mathrm{E}_i|\Delta^{\mathrm{ref}}_{i,a}(m_l)|} = \frac{|m_s-1|}{|m_l-1|} = \frac{1}{2}, \tag{4}$$

and an absolute mean keeps both parts of the response. Over the 132 attribute cells the common part has a small-to-large ratio of 0.509, the respondent-specific part one of 0.803, and the mean absolute response one of 0.851, which is the index value of 1.70. Proportionality can instead be tested without forming a ratio. With k the reference ratio and $A_i(m)$ the response of respondent i aligned to the direction the reference predicts, $c_i = A_i(m_s) - k\,A_i(m_l)$ has expectation zero under proportionality, is a linear combination of means, and so needs no selection and has no floor. A cluster bootstrap over respondents puts the interval above zero in 10 cells and below it in 14, with a median contrast of -0.005. The same contrast on absolute responses has a median of $+0.536$, and that difference is the measurement effect rather than a property of the model.

The systematic share reports the common part as a fraction of a typical respondent's movement, $|\mathrm{E}_i A_i|/\mathrm{E}_i|A_i|$, with a median of 0.365 at the large change and 0.223 at the small one. Every breakdown of it in this paper is a median over attribute cells and not over configurations, and it is not the

share of a pooled response, which would weight a cell by how far its respondents move. Table 7 gives the share of every configuration so that any of them can be recomputed. Perturbation size matters for the contrast as well as for the index. Swissmetro headway, perturbed by 25 and 50 percent rather than 10 and 20, gives a median contrast of -0.025 or -0.17 relative to its large-change response, and the fishing catch rate, the one attribute perturbed upward, gives $+0.03$ or $+0.57$ relative, so the deviations that survive the contrast test are attribute-specific and go in both directions.

C THE RESPONDENT-SPECIFIC PART

Every number reported for this part is computed after residualising each response on what the human reference predicts for that respondent, so that heterogeneity the reference itself contains is not counted. Without that step the correlation across the two sizes of change, the slope of the large on the small displacement, the correlation with the unperturbed target log-odds and the cross-prompt correlation are 0.51, 0.64, -0.08 and 0.42, against 0.47, 0.61, -0.06 and 0.44 with it, so the control is not what produces them. The cross-prompt correlation is taken over 90 pairs of prompt architectures, 43 of which exceed 0.5, and by model its median is 0.76 for Phi-3-mini, 0.75 for Qwen2.5-32B, 0.53 for Qwen2.5-14B, 0.51 for Qwen2.5-7B and 0.18 for Mistral-7B.

D WHERE THE SIZE SITS INSIDE THE RESPONSE

The first principal component of the response carries little of the size in most configurations, with R^2 at most 0.1 in 12 of 17 and below that of components 2 to 8 in 13 of 17. The exceptions include the two Qwen2.5-32B heating configurations, the only ones that match human magnitude under the corrected target, where it carries 0.53 against at most 0.07 for components 2 to 8. Across the 16 configurations that pass L0 that R^2 predicts the systematic share at Spearman $\rho = 0.81$ and the preregistered index at $\rho = -0.74$ (bootstrap interval $[-0.91, -0.36]$, permutation $p = 0.002$), the only one of six summary measures that predicts behaviour at all, and it survives a Bonferroni correction for those six ($p \approx 0.007$). The 16 share four models, three datasets and three prompt architectures, so we checked whether any of those carries it. Leaving out any one model, dataset or prompt architecture leaves ρ between -0.50 and -0.83. Permuting the first component R^2 within model gives $p = 0.008$ and within dataset $p = 0.002$, permuting whole model blocks gives $p = 0.04$, the smallest value four blocks allow, and removing model and dataset means from both variables leaves $\rho = -0.67$ ($p = 0.005$). Response strength is the obvious confound, correlating with both at $|\rho| = 0.58$, and removing its rank leaves $\rho = -0.60$ ($p = 0.009$).

It stays out of the argument for three reasons. The 17 configurations were selected partly on the behavioural outcome, so the sample is not independent of what it is used to predict. Two attempts to turn the relation into a manipulation gave differences in the predicted direction and neither reached significance. And on five configurations from the held-out families it agrees in sign at $\rho = +0.40$ with $p = 0.50$. A larger share on the first component also cannot be distinguished from less of anything else being placed there, and along the heating ladder the 14B configurations improve on 7B without any such shift, so it is not the only route. We report it because it was run and because it is the only summary measure that tracks behaviour, and we draw no conclusion from it.

E PROBE AND PATCHING DETAIL

The machinery is standard: linear probes scored on held-out data, activation patching, an interchange intervention, and a low-rank arm with a norm-matched control. Two things differ from the usual form. The probe target is the change in the target log-odds that a response function fitted to real choices assigns to that respondent and perturbation, rather than a label or a concept, which is what makes the direction control below necessary. And the transplant applies the interchange intervention to a residual difference rather than to a representation, read out under every cyclic relabelling, so the correlation it reports is between respondent-specific deviations rather than between raw responses.

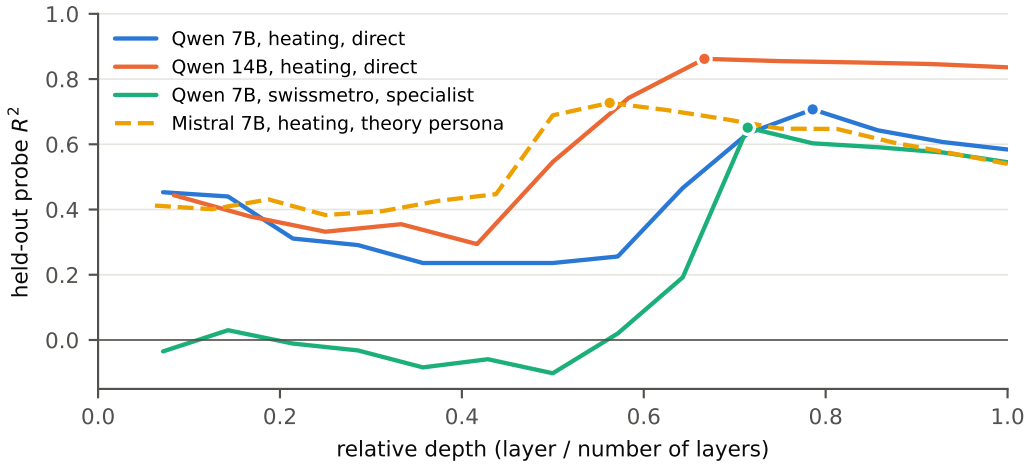


Figure 3: Held-out probe R^2 for the size of the applied change, by relative depth, in four configurations probed every second (Qwen2.5-14B every fourth) layer. Dots mark the peak.

Writing h_i for the residual stream at the probed layer, $u_i = \Delta^{\mathrm{ref}}_{i,a}(m)$ for the target and $\mathcal{T}$ for the training respondents,

$$\hat{w} \;=\; \arg\min_{w} \sum_{i \in \mathcal{T}} \left(h_i^{\top} w - u_i\right)^2 + \alpha \|w\|^2, \qquad R^2 \;=\; 1 - \frac{\sum_{i \notin \mathcal{T}} (u_i - h_i^{\top}\hat{w})^2}{\sum_{i \notin \mathcal{T}} (u_i - \bar{u})^2}, \tag{5}$$

with $\alpha = 10^3$, $\bar{u}$ over the held-out respondents, and a control refitted on a randomly permuted target. Since u_i is signed, writing $s_i = \mathrm{sign}(u_i)$ and $\hat{u}^{(s)}$ for the probe fitted within that sign, the two direction-controlled quantities of Section 3.3 are

$$R^2_{\mathrm{within}} = \frac{1}{2} \sum_{s \in \{-1,+1\}} \left[1 - \frac{\sum_{i \notin \mathcal{T},\, s_i = s} (u_i - \hat{u}^{(s)}_i)^2}{\sum_{i \notin \mathcal{T},\, s_i = s} (u_i - \bar{u}^{(s)})^2}\right], \qquad \mathrm{acc} = \Pr\left[\, s\, \hat{u}^{(s)}_{\mathrm{large}} > s\, \hat{u}^{(s)}_{\mathrm{small}}\right], \tag{6}$$

the probability taken over held-out triples of respondent, attribute and direction. We also report the R^2 of the best predictor that knows only the direction, as an upper bound on what direction alone explains.

For patching, with h^{b}_i and h^{p}_i the residual stream in the unperturbed and perturbed runs, Π_V the projection onto the subspace an arm uses, and $y(\cdot)$ the readout of Equation 3 when the run continues from a given activation,

$$\tilde{h}_i = h^{\mathrm{b}}_i + \Pi_V\left(h^{\mathrm{p}}_i - h^{\mathrm{b}}_i\right), \qquad \mathrm{recovery} = \mathrm{corr}_i\left(y(\tilde{h}_i) - y(h^{\mathrm{b}}_i),\; y(h^{\mathrm{p}}_i) - y(h^{\mathrm{b}}_i)\right), \tag{7}$$

the correlation running over held-out respondents.

F What the individual-level failure costs a user

For respondent i at multiplier m the implied sensitivity is $e_i(m) = \Delta_i(m)/(m-1)$, which for the fitted reference equals $\beta_a x_{i,a}/\sigma_a$ and is therefore the same at both sizes of change by construction. This is the decomposition of Section 4.1 in the units a practitioner reads rather than independent evidence. Over 171 attribute cells the mean of e_i at the small change divided by its mean at the large change is 1.063 for heating, 0.962 for swissmetro and 0.902 for catsup, and 0.634 for fishing and 0.730 for electricity, the two datasets Appendix G already singles out.

Which respondents are sensitive is not stable. The rank correlation of e_i between the two sizes has a median of 0.400 and ranges from -0.235 to 0.928, exceeding 0.5 in 56 of the 171 cells where the reference predicts 1, and the most sensitive quarter of respondents is the same quarter at the other

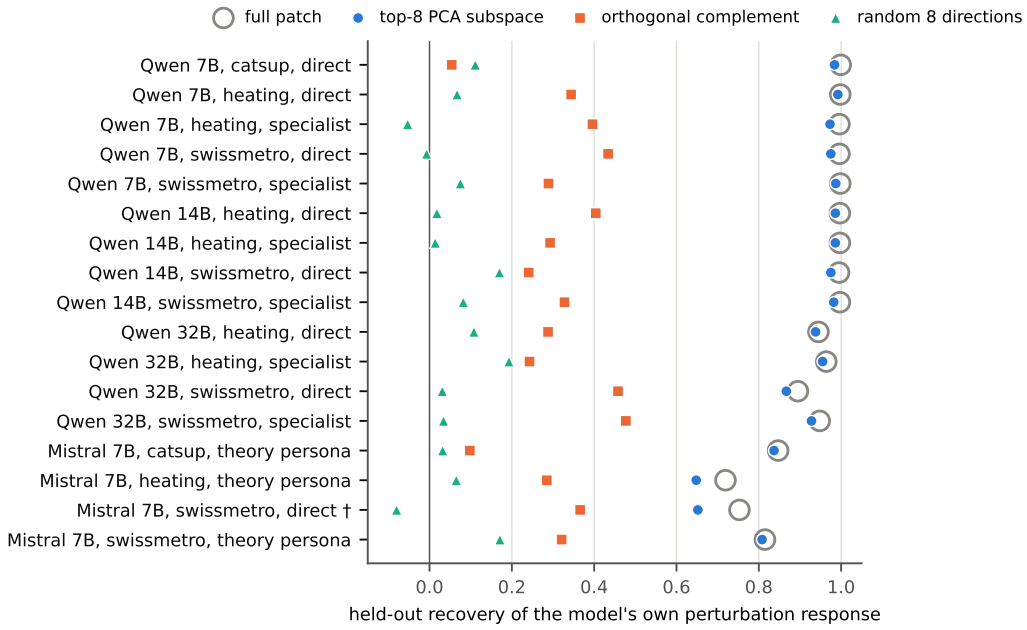


Figure 4: Held-out recovery of the model's own perturbation response when patching the full difference, the top-8 PCA subspace, its orthogonal complement, or 8 random directions. † fails L0.

Table 5: Probe and patching, the 17 mechanistic configurations. Layer is the depth at which the held-out probe peaks and frac. its relative depth. R^2 columns are the held-out probe at that layer and at the output. Recovery columns are the correlation across held-out respondents between the patched change in the readout and the model's own, for the full residual difference, its first component, its top eight, the orthogonal complement of those eight, and eight random directions.

			layer		probe R^2		recovery				
model	domain	prompt	at	frac.	peak	last	full	PC1	top 8	compl.	random
Mistral-7B	catsup	theory persona	16	0.50	0.901	0.801	0.847	0.681	0.837	0.098	0.032
Mistral-7B	heating	theory persona	16	0.50	0.680	0.548	0.719	0.278	0.648	0.285	0.065
Mistral-7B	swissmetro	direct	16	0.50	0.647	0.458	0.753	−0.050	0.652	0.366	−0.080
Mistral-7B	swissmetro	theory persona	16	0.50	0.637	0.468	0.815	0.172	0.808	0.321	0.171
Qwen2.5-14B	heating	direct	36	0.75	0.855	0.836	0.997	0.967	0.986	0.404	0.018
Qwen2.5-14B	heating	specialist	36	0.75	0.848	0.817	0.997	0.956	0.986	0.293	0.014
Qwen2.5-14B	swissmetro	direct	36	0.75	0.628	0.544	0.995	0.717	0.975	0.241	0.170
Qwen2.5-14B	swissmetro	specialist	36	0.75	0.648	0.628	0.997	0.865	0.982	0.328	0.082
Qwen2.5-32B	heating	direct	48	0.75	0.859	0.778	0.945	0.817	0.938	0.288	0.108
Qwen2.5-32B	heating	specialist	48	0.75	0.894	0.802	0.964	0.849	0.955	0.243	0.193
Qwen2.5-32B	swissmetro	direct	48	0.75	0.730	0.660	0.895	0.406	0.867	0.458	0.031
Qwen2.5-32B	swissmetro	specialist	48	0.75	0.717	0.654	0.948	0.359	0.928	0.477	0.034
Qwen2.5-7B	catsup	direct	21	0.75	0.842	0.792	0.999	0.882	0.984	0.054	0.111
Qwen2.5-7B	heating	direct	21	0.75	0.740	0.629	0.998	0.974	0.992	0.344	0.067
Qwen2.5-7B	heating	specialist	21	0.75	0.791	0.739	0.996	0.935	0.973	0.396	−0.053
Qwen2.5-7B	swissmetro	direct	21	0.75	0.564	0.504	0.996	0.908	0.975	0.434	−0.007
Qwen2.5-7B	swissmetro	specialist	21	0.75	0.611	0.545	0.998	0.858	0.987	0.289	0.075

size 0.480 of the time against 0.250 by chance. The coefficient of variation of e_i is 3.32 at the small change and 1.97 at the large one, so even the dispersion a user would integrate over depends on the size used to measure it. Scale helps and does not fix it: Qwen2.5-32B reaches a rank correlation of 0.570 and a quartile overlap of 0.64, against 0.34 to 0.40 and 0.44 for the smaller models. The cells are not independent, sharing five models, five datasets and five prompt architectures, so we report medians and counts rather than a test.

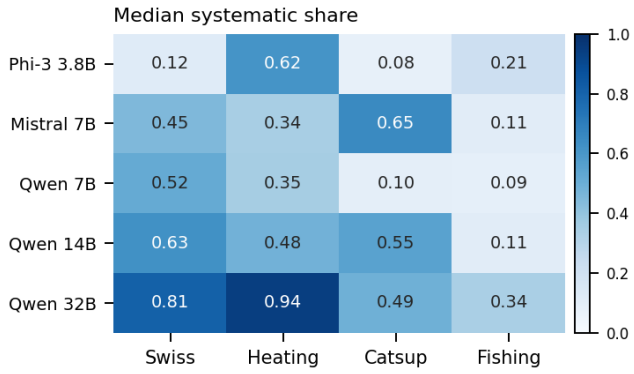


Figure 5: Median systematic share by model and dataset. The rise with model size is clean on swissmetro and heating and not elsewhere, and the electricity column is omitted because its reference is invalid.

G WHERE THE FAILURES CONCENTRATE

Table 6 gives the systematic share and the small-to-large ratio of the common part for each perturbed attribute. Seven of the nine have a common part proportional to within 0.055 of the target. Both exceptions were predictable from the design. The fishing catch rate is the only attribute made better by raising it, and it has the lowest share of any attribute and a common ratio of 0.859. Electricity price is the attribute whose fitted weight is positive and whose reference we treat as invalid, and it appears in one cell only.

The direction split is computed from the same runs over all 92 configurations. At the large change the mean absolute response is 1.331 for improvements against 1.214 for worsenings, larger for improvements in 73 of the 92 with a Wilcoxon signed-rank test at $p < 0.001$, and the systematic share is 0.453 against 0.291. At the small change the values are 1.037 and 0.991, larger for improvements in 56 of 92, with shares of 0.324 and 0.219. The configurations share five models and five datasets, so we report the counts rather than treating the test as a population statement.

Table 6: Systematic share and small-to-large ratios by perturbed attribute, over the attribute cells of the configurations that pass L0. A common part that is proportional has a ratio of one half.

			systematic share		small-to-large ratio		
dataset	attribute	cells	large	small	common	individual	index
catsup	price	18	0.224	0.107	0.501	0.800	1.716
electricity	price	1	0.110	0.180	0.858	0.387	1.050
fishing	catch rate	13	0.064	0.052	0.859	0.899	1.908
fishing	price	13	0.163	0.073	0.335	0.663	1.753
heating	installation cost	18	0.684	0.477	0.517	0.796	1.555
heating	operating cost	18	0.319	0.187	0.555	0.826	1.806
swissmetro	cost	17	0.579	0.349	0.484	0.821	1.712
swissmetro	headway	17	0.537	0.250	0.508	0.781	1.363
swissmetro	travel time	17	0.488	0.370	0.513	0.822	1.622

H THE FULL GRID

Table 7 lists every configuration that was run. Its share column is computed once per configuration, while the medians quoted in Sections 4.1 and 5.2 are over attribute cells, of which a configuration has one per perturbed attribute, so neither an average nor a median of this column reproduces them. The cell-level values are in the released decomposition.

Table 7: Every configuration that was run. Sign is agreement with the reference on direction, index is the preregistered linearity index and corr. its corrected-target value, and share is the systematic share of that configuration, computed once per configuration rather than as a median over attribute cells.

domain	model	prompt	sign	index	corr.	L0	L1	L2	L2c	share
catsup	Mistral-7B	actor-reviewer	0.495	2.08	2.30	✓				0.03
catsup	Mistral-7B	direct	0.689	1.37	1.51	✓	✓			0.71
catsup	Mistral-7B	functional delib.	0.537	2.00	2.20	✓				0.12
catsup	Mistral-7B	specialist	0.689	1.35	1.49	✓	✓			0.65
catsup	Mistral-7B	theory persona	0.708	1.55	1.71	✓	✓			0.71
catsup	Phi-3-mini	actor-reviewer	0.513	1.90	2.09	✓				0.02
catsup	Phi-3-mini	direct	0.512	2.12	2.34	✓				0.08
catsup	Phi-3-mini	theory persona	0.511	2.03	2.24	✓				0.17
catsup	Qwen2.5-14B	actor-reviewer	0.485	1.48	1.63	✓				0.36
catsup	Qwen2.5-14B	direct	0.604	1.36	1.50	✓				0.60
catsup	Qwen2.5-14B	theory persona	0.616	1.47	1.62	✓				0.55
catsup	Qwen2.5-32B	direct	0.646	1.50	1.66	✓				0.45
catsup	Qwen2.5-32B	specialist	0.630	1.39	1.54	✓				0.53
catsup	Qwen2.5-7B	actor-reviewer	0.529	1.87	2.07	✓				0.10
catsup	Qwen2.5-7B	direct	0.489	1.78	1.97	✓				0.00
catsup	Qwen2.5-7B	functional delib.	0.479	1.88	2.07	✓				0.10
catsup	Qwen2.5-7B	specialist	0.407	1.65	1.82	✓				0.28
catsup	Qwen2.5-7B	theory persona	0.504	1.79	1.98	✓				0.06
electricity	Mistral-7B	actor-reviewer	—	—	—					—
electricity	Mistral-7B	direct	—	—	—					—
electricity	Mistral-7B	functional delib.	—	—	—					—
electricity	Mistral-7B	specialist	—	—	—					—
electricity	Mistral-7B	theory persona	—	—	—					—
electricity	Phi-3-mini	actor-reviewer	—	—	—					—
electricity	Phi-3-mini	direct	—	—	—					—
electricity	Phi-3-mini	theory persona	—	—	—					—
electricity	Qwen2.5-14B	actor-reviewer	—	—	—					—
electricity	Qwen2.5-14B	direct	—	—	—					—
electricity	Qwen2.5-14B	theory persona	—	—	—					—
electricity	Qwen2.5-7B	actor-reviewer	—	—	—					—
electricity	Qwen2.5-7B	direct	0.489	1.05	1.05	✓		✓	✓	0.11
electricity	Qwen2.5-7B	functional delib.	—	—	—					—
electricity	Qwen2.5-7B	specialist	—	—	—					—
electricity	Qwen2.5-7B	theory persona	—	—	—					—
fishing	Mistral-7B	actor-reviewer	—	—	—					—
fishing	Mistral-7B	direct	0.528	1.88	1.91	✓				0.08
fishing	Mistral-7B	functional delib.	—	—	—					—
fishing	Mistral-7B	specialist	0.529	1.78	1.81	✓				0.11
fishing	Mistral-7B	theory persona	0.538	1.86	1.88	✓				0.11
fishing	Phi-3-mini	actor-reviewer	—	—	—					—
fishing	Phi-3-mini	direct	0.512	1.79	1.81	✓				0.21
fishing	Phi-3-mini	theory persona	0.501	1.74	1.76	✓				0.23
fishing	Qwen2.5-14B	actor-reviewer	0.520	2.05	2.08	✓				0.09
fishing	Qwen2.5-14B	direct	0.553	1.93	1.96	✓				0.19
fishing	Qwen2.5-14B	theory persona	0.533	1.83	1.85	✓				0.13
fishing	Qwen2.5-32B	direct	0.585	1.83	1.85	✓				0.34
fishing	Qwen2.5-32B	specialist	0.578	1.81	1.83	✓				0.34
fishing	Qwen2.5-7B	actor-reviewer	—	—	—					—
fishing	Qwen2.5-7B	direct	0.523	1.85	1.88	✓				0.09
fishing	Qwen2.5-7B	functional delib.	—	—	—					—
fishing	Qwen2.5-7B	specialist	0.512	1.79	1.82	✓				0.11

domain	model	prompt	sign	index	corr.	L0	L1	L2	L2c	share
fishing	Qwen2.5-7B	theory persona	0.508	1.83	1.85	✓				0.09
heating	Mistral-7B	actor-reviewer	0.512	1.90	1.90	✓				0.10
heating	Mistral-7B	direct	0.602	1.80	1.80	✓				0.63
heating	Mistral-7B	functional delib.	0.553	1.84	1.84	✓				0.27
heating	Mistral-7B	specialist	0.612	1.72	1.72	✓				0.60
heating	Mistral-7B	theory persona	0.623	1.79	1.79	✓				0.60
heating	Phi-3-mini	actor-reviewer	—	—	—					—
heating	Phi-3-mini	direct	0.672	1.54	1.54	✓	✓			0.63
heating	Phi-3-mini	theory persona	0.642	1.64	1.64	✓				0.55
heating	Qwen2.5-14B	actor-reviewer	0.568	1.80	1.80	✓				0.35
heating	Qwen2.5-14B	direct	0.552	1.52	1.52	✓				0.59
heating	Qwen2.5-14B	specialist	0.588	1.42	1.42	✓				0.60
heating	Qwen2.5-14B	theory persona	0.547	1.55	1.55	✓				0.52
heating	Qwen2.5-32B	direct	0.829	1.22	1.22	✓	✓	✓	✓	0.90
heating	Qwen2.5-32B	specialist	0.841	1.04	1.04	✓	✓	✓	✓	0.94
heating	Qwen2.5-7B	actor-reviewer	0.521	1.94	1.94	✓				0.15
heating	Qwen2.5-7B	direct	0.562	1.71	1.71	✓				0.49
heating	Qwen2.5-7B	functional delib.	0.524	1.65	1.65	✓				0.39
heating	Qwen2.5-7B	specialist	0.564	1.69	1.69	✓				0.42
heating	Qwen2.5-7B	theory persona	0.513	1.73	1.73	✓				0.41
swissmetro	Mistral-7B	actor-reviewer	0.547	1.84	1.87	✓				0.29
swissmetro	Mistral-7B	direct	—	—	—					—
swissmetro	Mistral-7B	functional delib.	0.558	1.78	1.81	✓				0.32
swissmetro	Mistral-7B	specialist	0.617	1.46	1.49	✓				0.52
swissmetro	Mistral-7B	theory persona	0.648	1.62	1.65	✓				0.60
swissmetro	Phi-3-mini	actor-reviewer	0.516	2.08	2.11	✓				0.13
swissmetro	Phi-3-mini	direct	—	—	—					—
swissmetro	Phi-3-mini	theory persona	—	—	—					—
swissmetro	Qwen2.5-14B	actor-reviewer	0.554	1.89	1.92	✓				0.22
swissmetro	Qwen2.5-14B	direct	0.677	1.49	1.51	✓	✓			0.61
swissmetro	Qwen2.5-14B	specialist	0.724	1.32	1.34	✓	✓			0.78
swissmetro	Qwen2.5-14B	theory persona	0.628	1.41	1.43	✓				0.54
swissmetro	Qwen2.5-32B	actor-reviewer	0.540	1.82	1.85	✓				0.21
swissmetro	Qwen2.5-32B	direct	0.731	1.39	1.41	✓	✓			0.75
swissmetro	Qwen2.5-32B	specialist	0.781	1.24	1.26	✓	✓	✓		0.85
swissmetro	Qwen2.5-32B	theory persona	0.743	1.38	1.40	✓	✓			0.77
swissmetro	Qwen2.5-7B	actor-reviewer	—	—	—					—
swissmetro	Qwen2.5-7B	direct	0.634	1.59	1.62	✓				0.55
swissmetro	Qwen2.5-7B	functional delib.	0.546	1.77	1.79	✓				0.38
swissmetro	Qwen2.5-7B	specialist	0.648	1.37	1.40	✓				0.64
swissmetro	Qwen2.5-7B	theory persona	0.648	1.48	1.51	✓				0.62

I PROMPT ARCHITECTURES

The five architectures share one template and differ only in the instruction that opens it. Two of them also insert a block produced by an earlier call to the same model. Line breaks inside a paragraph were added to fit the page and are not in the prompt. Throughout, **blue** marks a field filled per respondent, **orange** marks what differs between architectures, **bold** marks a structural label the model sees every time, and a highlight marks the one value a perturbation changes.

Shared template

```
{INSTRUCTION} Use only supplied facts.
Which option is most likely to be selected in this decision?
A: {option A}
```

```
B: {option B}
C: {option C}
DECISION-MAKER FACTS:
{persona facts}
ALTERNATIVE FACTS:
{alternative facts}{MEMO BLOCK}
Answer with exactly one letter (A, B, C) and nothing else.
```

The letters run to the number of alternatives, four for electricity, catsup and fishing and five for heating. The option block is written under the relabelling in force, so a letter refers to a different alternative in each cyclic relabelling while the template is unchanged. The memo block is empty for three of the five architectures, and each instruction below is followed by `Use only supplied facts.`

The five instructions

```
direct
  Act as a statistical {decision} choice analyst.

theory persona
  Simulate the described decision maker under a constrained
  random-utility and planned-behavior model. Use stated alternative
  attributes for attitude and supplied household facts for constraints.
  Do not invent preferences, norms, relationships, or unavailable facts.

specialist
  Act as the specialist responsible for evaluating {attribute}.

actor-reviewer
  Make the final decision after using the supplied actor draft and
  independent reviewer critique.

functional deliberation
  Act as a neutral moderator. Reconcile the independent functional
  assessments below. The roles are analytical functions, not observed
  people.
```

{decision} is a fixed phrase per dataset, for instance `travel-mode` for swissmetro. **{attribute}** is the attribute assigned to that respondent, so the specialist is told to evaluate the attribute that is about to be perturbed. This is the one architecture whose instruction depends on the perturbation, which is why it is highlighted.

Memo block, actor-reviewer then functional deliberation

```
ACTOR DRAFT: {actor output}
REVIEWER CRITIQUE: {reviewer output}

{ROLE} MEMO: {role output}
```

Both are generated by the same model and weights that then make the choice, in a separate call before the labelled question is asked. They are produced once per respondent and reused across all relabellings, so the relabelling does not change them.

Memo prompt

```
Act as the {role} functional analyst. Compare all options using
only {focus}. Recommend one named option and give one concise
factual reason.
DECISION-MAKER FACTS:
{persona facts}
ALTERNATIVE FACTS:
{alternative facts}
```

The roles are defined by their scope (Table 8), and each dataset uses the roles that match its attributes.

Table 8: Functional roles and what each is allowed to compare.

role	scope
cost	monetary costs and affordability
time	travel times and trip context
service	headways, waiting, luggage and seating
installation	up-front installation costs
operations	annual operating costs
household fit	household and dwelling facts, no invented technical performance
price	prices and affordability
contract	contract length and rate-plan conditions
supplier	supplier characteristics
promotion	displays and feature advertising
brand	brand and package alternatives, no invented brand preferences
catch	catch rates across fishing alternatives
budget	reported income and trip affordability

Actor and reviewer prompts

```
actor
  Act as a theory-constrained decision-maker. Produce a concise
  provisional recommendation and rationale. Use only supplied facts; do
  not invent preferences or relationships.
  DECISION-MAKER FACTS:
  {persona facts}
  ALTERNATIVE FACTS:
  {alternative facts}

reviewer
  Act as an independent behavioral-validity reviewer. Identify factual
  invention, ignored constraints, or failure to compare the supplied
  attributes. Give concise revision instructions; do not invent facts.
  FACTS:
  {persona facts}
  {alternative facts}
  ACTOR DRAFT:
  {actor output}
```

One rendered prompt follows, the direct architecture on the first swissmetro respondent under the identity relabelling. The highlight is the only thing a perturbation touches.

One rendered prompt

```
Act as a statistical travel-mode choice analyst. Use only supplied
facts. Which option is most likely to be selected in this decision?
A: conventional train
B: Swissmetro
C: car
DECISION-MAKER FACTS:
Traveler: age 25-39, female, income CHF 50k-100k; trip purpose
shopping; luggage code 1; first class no; annual transit pass no.
ALTERNATIVE FACTS:
Conventional train: time 52 min, cost CHF 92, headway 60 min.
Swissmetro: time 30 min, cost CHF 133, headway 10 min, seats code
0. Car: time 52 min, cost CHF 32. All are available.
Answer with exactly one letter (A, B, C) and nothing else.
```

Under a large perturbation of cost, `cost CHF 133` becomes `cost CHF 106`, and every other character of the prompt is unchanged. The code also carries a rephrased variant, used in the robustness check of Section 3.1 and not in the main grid, which replaces the closing sentence of the instruction with `Evaluate every listed option symmetrically and make`

no assumptions beyond the record. and the question with Given this record, select the single option with the highest choice likelihood.