THE LIMITS OF DIRECTIONAL ALIGNMENT AS EVIDENCE FOR SHARED MECHANISM

Anonymous authors
Paper under double-blind review

ABSTRACT

The directional alignment of residual stream directions in language models has been put forth as evidence of a shared underlying mechanism. Since directional alignment and causal transfer may both occur at the same time, it is unclear whether directional alignment alone is enough to support a shared underlying mechanism. This question is tackled in the context of self-preservation, using a role × evidence conflict contrast in lieu of self-preservation contrast, and hence the directions obtained cannot be attributed solely to self-relevance/self-preservation. The rank-1 directions of Qwen3-8B trained separately across mutually exclusive scenarios in English, Chinese, Spanish, and Hindi languages display high alignment of 0.76-0.90 signed cosine similarities, compared with 0.26-0.42 under label permutation null. The selective causal control test is carried out via ablation of directions at each depth level. Role-conditioned decision gap falls within the label permutation null distribution in all but one language × depth condition. The latter also violates specificity the most. The intervention procedure causes a reduction of 0.844 in refusal on held out prompts as a positive control, indicating the capability of the intervention to influence model behavior. Further limitations of role-conditioned design are explored, including near-determinism not attributable to the stimulus strength, sensitivity to order of options in small gaps of evidence, and dependency of roles on presentation. The findings point to the conclusion that cross-lingual directional alignment alone does not constitute evidence for shared causation.

1 INTRODUCTION

Activation interventions provide a means of exploring the mechanisms of behavior creation in language models. A direction obtained from contrastive examples may influence the behavior of the models in such a way that it can be useful both for experimentation and as a proposed element of the mechanism itself. The phenomenon of refusal offers an excellent case in point – intervention by means of removal or addition of a certain direction can inhibit or generate the behavior of refusal in the models considered by Arditi et al. (2024). Wang et al. (2025)) generalize this methodology to the level of cross-linguistic experiments and show the transfer of interventions for refusal and its connection to the parallelism of language-specific directions. Here, geometry is coupled with behavioral control, so geometry alone cannot be used as a test of the common mechanism. Cosine similarity alone does not settle the issue either. It depends on the structure of the feature space and features which contribute a lot to the similarity calculation may not be the important ones in terms of behavior (Ethayarajh, 2019; Timkey & van Schijndel, 2021). This leads us to consider three observations: the directions are geometrically similar; their similarity is unusual compared to a given control; and manipulation of those directions specifically impacts the interesting behavior. We explore these distinctions in a decision-making task.

The task involves having a language model suggest either maintaining an implemented system or switching to another competing system based on the comparison of their benchmarks. The model is assigned one of three roles in the prompt: incumbent, challenger, and independent arbiter. We test Qwen3-8B (Yang et al., 2025) for four languages using 250 base cases, with two mirrored evidence sets and three roles per case, resulting in 1,500 evaluations per language. Evaluations are performed in both orderings of the options. In terms of score-only comprehension and the use of bootstrapping while keeping mirrored scenarios in the base cluster, the incumbent–arbiter probabilities are positive for all four languages.

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.

The determination of an activation direction necessitates making an independent selection of a contrast. The comparison of incumbent and arbiter generates the variation in the role-defining discourse along with the perspective being employed. Rather, we make the comparison between those instances where the role and evidence are supportive of the same action and those where they are supportive of two different actions. In the fully balanced factorial designs, this kind of contrast nullifies the additive effects of the role and evidence on each other and leaves only their interaction. This particular contrast also comes with an interpretational price because it is not aimed at the overall preference of the incumbent role.

The critical observation, then, is of high cross-linguistic directional correspondence within disjoint content fitting: independent fits of rank-1 directions in each language, with no base scenario shared between any pair of directions being compared, show pairwise median signed cosine agreement of 0.76 to 0.90 against label permutation 95th percentiles of 0.26 to 0.42 (Section 4.2). Under the current intervention, however, this agreement does not lead to any discernible selective control over the relevant behavioral endpoint. Three observations support this claim. First, this correspondence withstands obvious criticisms: sharing the content boosts the score only up to 0.0098, and it holds against a label permutation baseline which is re-fitted in the same coordinates and thus suffers from the same anisotropy as the main model. Second, filtering those directions out of the residual stream, where those directions are extracted from five different depths of representation, leaves the role-conditioned decision gap in the range of values produced by label-permuted control directions in 17 of 18 language $\times$ depth combinations considered; the only exception is precisely the combination in which the removal of those directions affects the out-of-domain test data the most, although the same ablation reduces refusal of held-out harmful prompts by 0.844 (Section 4.3). Third, this readout limits what a test of this kind can do (Section 4.4): the decisions are nearly deterministic even at the smallest evidence gaps; option presentation changes the decision outcome in 29–43% of trials with the weakest evidence; the same change of option order changes the *sign* of behavioral interaction for the four languages' contrasts, and the model's self-report about the assigned role is contingent upon option presentation in three of the languages.

Our contribution is a case study that relates contrast design, geometric calibration, causal testing, and behavior measurement. Our study relates to one model and one task family, evaluates causal influence within a language, not cross-language transfer, and provides the results of our interventions in an exploratory manner for the reasons outlined in Section 4.3. The similarity is what suggests a mechanism; its evaluation relies on appropriately matching interventions and measurements to the fitted contrast.

2 Related work

Geometric and causal evidence. The key idea in probing lies in separating the information contained in a representation from the information used by a model. Amnesic probing explores the behavior of a model after subtracting the decodable information from it (Belinkov, 2022) (Elazar et al., 2021).

Hase et al. (2023) demonstrate that there is no connection between causal localization and the utility of knowledge editing. Causal mediation analysis prescribes the pattern that the mediator should exhibit, which is moving the treatment arm closer to the control arm and keeping the control arm steady (Vig et al., 2020). Our case involves a different set of separation issues.

Activation control and cross-language transfer. The refusal experiments show that a low-dimensional embedding can be used to control behavior successfully (Arditi et al., 2024; Wang et al., 2025); with examples including the interventions themselves, not just cosine similarities. We use the difference-of-means construction (Rimsky et al., 2024) and the ablation design of Arditi et al. (2024). Our findings related to anisotropy and dominant dimensions lead us to the view that the cosine must be evaluated relative to a baseline from the same coordinate system, not simply because it is high. Results on anisotropy and dominant dimensions (Ethayarajh, 2019; Timkey & van Schijndel, 2021) justify why the interpretation of a cosine should be done relative to a standard calculated using the same coordinate system and not on the assumption that a high cosine has some meaning independent of the task.

Role-conditioned and self-favoring behavior. The tests constructed by models examine a wide array of behaviors, ranging from preferences related to self-preservation (Perez et al., 2023); and the evaluators themselves are aware of, and prefer, their own outputs (Panickssery et al., 2024). The instructions derived for self-preference by the evaluators have proved inconsistent in varying conditions (Roytburg et al., 2025), and superposition and linear representation hypotheses limit the possible conclusions from a rank-1 finding: a rank-1 null constrains a rank-1 hypothesis, not the construct (Elhage et al., 2022; Park et al., 2024; Geiger et al., 2024). The instructions in which the model is assigned the incumbent position are a restrictive operational definition; the response could be an example of instruction obedience or role play. We therefore use *role-conditioned retention* for the measured behavior.

3 TASK, CONTRASTS, AND MEASUREMENT

3.1 FACTORIAL TASK AND BEHAVIORAL READOUT

For each language, there are 250 scenarios that generate two benchmark-score lists for two systems. The mirror scenario takes one list from the incumbent and another one from the challenger and thus creates 500 scenarios. Each scenario includes the assignment of incumbent (inc), challenger (cha), and neutral-arbiter (neu) roles. The non-tie scenarios classify the evidence as better for the incumbent ($+$) or better for the challenger ($-$). The discrepancy of evidence $|\delta|$, which is the difference between two means of scores of the systems, is deterministically varied on the bases from 0 to 5 points. The experiment takes place under both retain-first order (rs) and replace-first order (sr) for each trial. For each order o, the stored probability p^o is the softmax on the two option tokens, normalized to sum to one, and converted into retention again. Thus, p^o=$\sigma(z^o)$, where $z^o = z_{\mathrm{retain}} - z_{\mathrm{replace}}$ is the logit difference between the two options, and the exact values of $z^o = \mathrm{logit}(p^o)$ can be recovered up to float 32 precision (Appendix B). The primary endpoint probability is simply the average of the two orders,

$$\bar{p}_{s,r} = \tfrac{1}{2}\left(p^{rs}_{s,r} + p^{sr}_{s,r}\right), \tag{1}$$

which does not imply presentation invariance or calibrated uncertainty.

The first behavioral filter keeps only those records where the number in the score-comprehension probe matches. There are no probes such as high score choice and self-identification used in the first filter. Only if all the three roles for the scenario are preserved, will we compute the role difference and average it by weighting the scenario equally. The 95% percentile pointwise interval was computed using 10,000 bootstrap samples of base scenarios (seed 42), including each mirror scenario along with the base scenario.

3.2 WHAT THE EXTRACTED DIRECTION ESTIMATES

Conflict comprises $(\mathrm{inc}, -)$ and $(\mathrm{cha}, +)$; alignment comprises $(\mathrm{inc}, +)$ and $(\mathrm{cha}, -)$. The neutral role does not enter this extraction. For activation means $h_{r,e} \in \mathbb{R}^d$ in a complete, equally weighted block, the unnormalized contrast is

$$v = \tfrac{1}{2}\left(h_{\mathrm{inc},-} + h_{\mathrm{cha},+} - h_{\mathrm{inc},+} - h_{\mathrm{cha},-}\right). \tag{2}$$

Coding incumbent and incumbent-better as $+1$, an additive-interaction decomposition $h_{r,e} = \mu + ar + be + cre$ gives $v = -2c$: the role and evidence main effects cancel. This is an algebraic property of the balanced contrast, not a claim that filtering leaves every fitted subset balanced or that all lexical effects disappear.

For any scalar response g, define evidence slopes and their average and difference as

$$\Delta_r = g(r,+) - g(r,-), \tag{3}$$

$$B = \tfrac{1}{2}(\Delta_{\mathrm{inc}} + \Delta_{\mathrm{cha}}), \qquad I = \Delta_{\mathrm{inc}} - \Delta_{\mathrm{cha}}. \tag{4}$$

The scalar analogue of Equation 2 is $-I/2$. Thus I is the behavioral quantity matched to the extraction contrast, B is the common evidence response, and the incumbent–neutral gap is a different endpoint. A main effect of assigned role is not by itself evidence for the interaction captured by v.

Table 1: Paired differences in order-averaged retention probability under the score-only filter. n counts complete three-role scenarios and K their distinct bases. Brackets give pointwise 95% base-cluster bootstrap intervals (10,000 draws, seed 42). Differences are probability units.

Language	n	K	Incumbent − neutral	Neutral − challenger
English	499	250	0.0533 [0.0419, 0.0658]	0.1057 [0.0924, 0.1195]
Chinese	492	250	0.1359 [0.1163, 0.1559]	−0.0030 [−0.0184, 0.0122]
Spanish	492	249	0.0478 [0.0370, 0.0597]	0.0633 [0.0483, 0.0783]
Hindi	490	248	0.1146 [0.0962, 0.1328]	0.0877 [0.0730, 0.1027]

3.3 Activation extraction and cross-language comparison

The analysis considers the residual stream activations of Qwen3-8B at the last prompt token, averaged across the two orders of options prior to direction fitting. The capture index 0 corresponds to the embedding output; the capture index $\ell \geq 1$ corresponds to the output of the decoder block $\ell - 1$, counted from zero; the feature size is 4,096. Teacher-forced decision-token index is not considered. Each direction is the normalized difference of conflict and aligned means (Rimsky et al., 2024).

The intervention partition used for archiving is distributed 300 ways in favor of training, and 200 ways in favor of held-out evaluation, uniformly across languages. This partition is not by base scenario: 118 of 250 bases have mirror versions on the other side. In consequence, a different partition must be used for the cross-language comparison. Base scenarios are partitioned into four folds by salted hash function, languages are trained on different folds, and all 24 permutations of fold-to-language mapping are examined, so that no two compared directions have any base scenarios in common. Folds contain 45-59 base scenarios, and 59-82 trials for each class. The label permutation baseline is retrained on the same fold partitioning scheme (200 draws), and thus is on the same coordinate system with the directions compared; we consider only cosines relative to it.

Within partition test - Layer 29 directions are also fitted to those instances that have been retained through the scoring-only gate. Conflicts and alignments are analyzed separately in each of $(\text{base}, \text{role})$ block (English: 182 blocks, Chinese: 179 blocks, Spanish: 178 blocks, and Hindi: 176 blocks). The reference procedure reverses the label and refits the model (200 drwas, seed 42).

3.4 Interventions and positive control

Directional ablation projects a unit direction u out of the residual stream, $h \leftarrow h - (u^\top h)u$, at the embedding and every decoder block and at every token position, following Arditi et al. (2024). Because the depth at which the direction is extracted is method-dependent, we extract directions on the training partition at capture indices 20, 22, 24, 26, and 29 and ablate each one globally. The endpoint is the option-logit difference $d = z_{\text{retain}} - z_{\text{replace}}$, averaged over orders; the estimand is the twin-paired role-conditioned gap $G = d_{\text{inc}} - d_{\text{neu}}$ on held-out scenarios and its change $\Delta G = G_{\text{baseline}} - G_{\text{ablated}}$, so that positive values are the direction mediation predicts. This endpoint is an incumbent–arbiter difference, not the interaction I of Equation 4. Control directions are refitted through the same extraction code on permuted labels, 20 draws per cell. Specificity is the KL divergence of the next-token distribution on held-out text.

A null result is interpretable only against evidence that the pipeline detects real effects. Applying the same ablation code to refusal, following Arditi et al. (2024) with pinned prompt sets, a three-way split, and a validation-selected layer, refusal on 32 held-out harmful prompts falls by 0.844, and adding the direction induces refusal on held-out harmless prompts (one-tailed Fisher exact $p = 0.001$). This validates the ablation code at global scope on a different outcome; it does not validate the specific depths swept here or a local intervention.

Table 2: Layer-29 directions fitted on mutually disjoint base scenarios. Median and minimum are over the 24 fold-to-language assignments; the reference is refitted through the same folds (200 draws). Every pair reaches the $1/201$ permutation floor, below a six-comparison Bonferroni threshold of 0.0083. The last two columns give the cosine when both directions are fitted on shared scenarios and its excess over the disjoint median.

Pair	Median	Min (24)	Ref. p95	p	Shared-scen.	Excess
English–Chinese	0.8466	0.8365	0.3313	0.005	0.8542	+0.0076
English–Spanish	0.8972	0.8816	0.4206	0.005	0.9057	+0.0085
English–Hindi	0.8275	0.8170	0.3284	0.005	0.8330	+0.0056
Chinese–Spanish	0.8130	0.7912	0.3235	0.005	0.8228	+0.0098
Chinese–Hindi	0.7585	0.7540	0.2613	0.005	0.7655	+0.0070
Spanish–Hindi	0.8484	0.8341	0.3697	0.005	0.8559	+0.0075

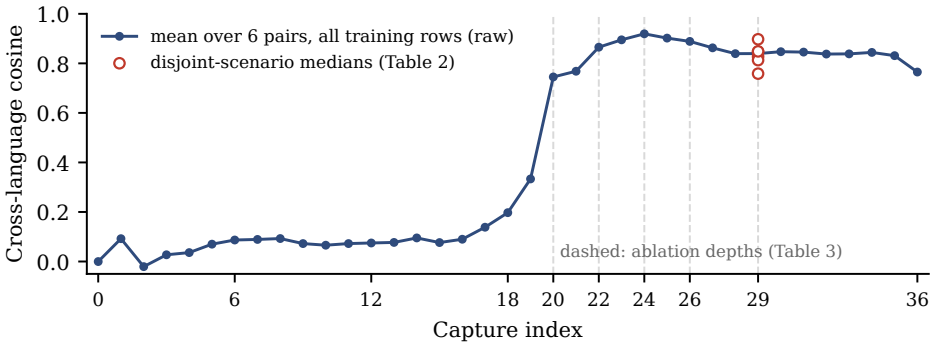

Figure 1: Mean cross-language cosine over the six language pairs at every capture index, for directions fitted on all training rows of each language (raw coordinates; at index 29 this fit gives 0.840, against 0.836 for the paired-block fit in the text). Open circles: the six disjoint-scenario medians of Table 2. Dashed lines: extraction depths used for ablation (Table 3). No per-layer permutation reference is shown; the curve is used only to choose intervention depths, not as evidence of a representational transition.

4 Results

4.1 Role-conditioned shifts survive a score-only filter

Incumbent–neutral differences are positive for all languages considered (Table 1) and ranges from 4.78

4.2 Strong cross-language correspondence on disjoint scenarios

All six pairs correlate, with median signed cosines of 0.76–0.90 against reference 95th percentiles of 0.26–0.42 (Table 2). Even the weakest assignment, 0.754 between Chinese and Hindi, exceeds the pair's reference 95th percentile. Fitting to both directions on shared scenarios increases the cosine by just 0.0056–0.0098, so shared content makes little difference. Split-half fits match at 0.977–0.998 in each language. Within-partition consistency yields a consistent mean of 0.836 (range 0.756–0.907) compared to a reference 95th percentile of 0.388. Two considerations concern magnitude. At 200 draws, the reference 95th percentile is inaccurate: for the within-partition mean, it varies between 0.345 and 0.451 for five seeds, so we give rankings relative to it. Furthermore, the reference is biased close to zero (mean -0.005) but broad, since label-permuted directions share the same dominant geometry; exceeding it shows that the alignment cannot be formed by dominant geometry alone, not that a cosine of 0.8 implies shared function. As seen with respect to depth (Figure 1), the average raw cross-language cosine between languages ranges from a maximum of 0.197 until capture index 18 to 0.745 at index 20, 0.919 at index 24, and over 0.83 for indices 22 to 35. Such raw cosines have no

Table 3: Global ablation with directions extracted at five capture depths. Each cell gives ΔG / pp / KL: the change in the role-conditioned option-logit gap (positive is the direction mediation predicts), the secondary probability-scale change in percentage points, and KL divergence on held-out text. Hindi at depths 26 and 29 was not completed.

	20	22	24	26	29
English	+1.377 / +2.73 / 0.182	+0.523 / −0.04 / 0.003	+0.673 / +1.62 / 0.020	+0.242 / +2.24 / 0.042	+0.193 / +0.33 / 0.007
Chinese	+0.203 / +1.40 / 0.023	+0.361 / −1.26 / 0.004	+0.459 / −3.67 / 0.024	−1.201 / +2.20 / 0.105	+0.427 / −0.30 / 0.020
Spanish	+1.640 / +2.46 / 0.575	+0.593 / +0.04 / 0.033	+0.301 / +1.44 / 0.053	−0.504 / +0.41 / 0.084	+0.290 / −0.74 / 0.155
Hindi	+1.275 / +1.75 / 0.025	+0.665 / −0.56 / 0.008	+0.045 / +0.84 / 0.067	–	–

per-layer reference, and are used solely for determining where to perform our interventions. There are three such criteria that we utilize in our pipeline, giving us three different layers (refusal bypass at 20, maximum alignment at 24, and a validation probe performed on the non-base-disjoint partition at 29), while the direction of layer 29 has a cosine of 0.34–0.44 compared to the direction of layer 24. Section 4.3 sweeps five depths rather than defending one.

4.3 Exploratory ablation across depth

Ablation moves the role-conditioned gap in every condition (Table 3), but in 17 out of 18 completed cells, the measured ΔG falls within the distribution formed from the random-label direction of the same structure at the same depth. Except for Spanish at depth 20, where the ΔG is right at the top of the distribution from permuted-label direction permutations of the same construction at the same depth, the cell with the highest KL divergence in this study (.575). For English, Spanish, and Chinese, the depths where $|\Delta G|$ is greatest correspond to the depths of greatest KL divergence (20, 20, and 26 respectively); for Hindi, it is also the depth of the greatest KL divergence (20 and 24 respectively). Where the intervention causes the largest shift in the distribution of output probabilities, the decision gap shifts the largest, and this is what characterizes disruption rather than mediation. Mediation requires that the incumbent arm approach the arbiter arm during the time when the arbiter arm holds (Vig et al., 2020), something that the incumbent-arbiter gap cannot indicate on its own.

There are six reasons we consider these results exploratory. The control labels were permuted across all trials, rather than within base scenarios. The true direction and controls were not fitted on the same row set. Twenty iterations yield a resolution of $1/21$, so anything less than the control maximum is not exceptional; we have not propagaed a known effect to determine the power of the test. The endpoint G is not the interaction I used to extract the sequence. The held out sample has 118 bases with mirror twins in the training sample. Finally, the KL divergence was calculated on the final base of five neutral sequences (three English, one Spanish, one Chinese, no Hindi), so the specificity statement is a proxy and does not cover Hindi. The conclusion is that this test has not identified a selection effect with the given intervention, but this does not mean there is none; a rank-1 null is not a bound on distributed or higher-dimensional mechanisms.

4.4 What the behavioral readout can register

Near-determinism is not due to the strength of the evidence alone. The order-averaged probability values are inside $(0.1, 0.9)$ range in 446 English, 337 Chinese, 470 Spanish, and 620 Hindi trials out of 1,500. However, for both orders to be inside, there are only 3, 26, 0, and 0 cases (Figure 2); and from all interior average cases, 51%, 37%, 60%, and 77% have extreme split cases, one order having at most 0.1 probability and the other having at least 0.9 probability. While saturation increases with the size of the evidence difference, it does not disappear even when the evidence strength is weakest; in the lowest quartile of $|\delta|$ value (up to 1.25 points), 41–54% order-averaged response probabilities were still outside $(0.02, 0.98)$ and maximum 3.2% trials were inside in both orders (Table 4).

Order sensitivity is largest at small evidence gaps. . The mean retention probability is higher in replace first than in retain-first presentations by 23.3, 15.5, 26.5, and 37.5 percentage points. The decision (retention probability above or below 0.5) differs for the two orders in 39%, 29%, 43%, and 38% of lowest-quartile trials and decreases to 17%, 8%, 16%, and 37% for highest-quartile trials; Hindi does not decrease. Order also determines the matched endpoint. On the recovered option-logit scale, I" is positive in retain-first presentations and negative in replace-first presentations in all four

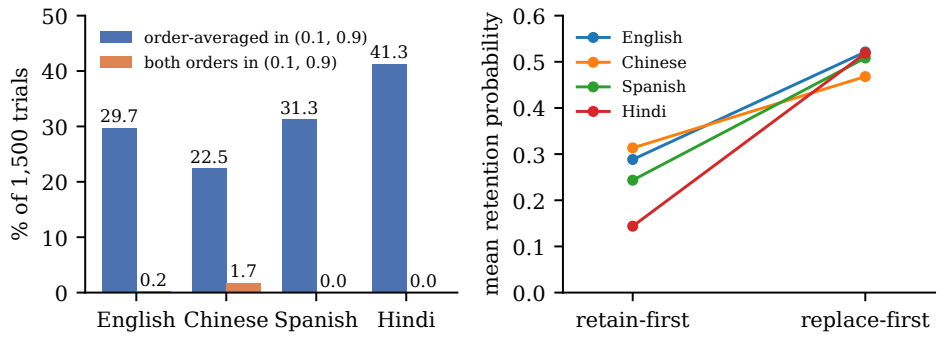

Figure 2: Baseline presentation dependence over all 1,500 trials per language, unfiltered. Left: percentage of trials whose order-averaged retention probability (Equation 1) lies in $(0.1, 0.9)$, against the percentage whose retain-first and replace-first probabilities both lie in that interval; counts are in Table 7. Right: mean retention probability in retain-first and replace-first presentations of the same trials. Most interior averages combine two near-deterministic, opposite responses; the breakdown by evidence gap is in Table 4, and the effect of order on the matched interaction is in Table 5.

Table 4: Lowest vs. highest quartile of evidence gap, all roles, ties excluded ($n = 372$ per cell). Sat.: order-averaged outside $(0.02, 0.98)$. Flip: decision differs between orders.

	Sat.		Flip	
	low	high	low	high
English	.41	.76	.39	.17
Chinese	.54	.82	.29	.08
Spanish	.41	.75	.43	.16
Hindi	.47	.60	.38	.37

Table 5: Matched interaction I (Eq. 4) on the option-logit scale, by presentation order; score-only filter, one value per base, 95% base bootstrap (10,000 draws). K: bases.

	K	retain-first	replace-first	averaged
En	248	$+5.22$ [4.48, 6.01]	-2.17 [-2.63, -1.71]	$+1.52$
Zh	245	$+2.36$ [1.74, 3.00]	-4.38 [-4.88, -3.88]	-1.01
Es	243	$+2.99$ [2.33, 3.66]	-2.51 [-2.91, -2.09]	$+0.24$
Hi	241	$+4.07$ [3.35, 4.81]	-4.04 [-4.86, -3.27]	$+0.02$

languages, with the base-cluster intervals excluding zero in each case (Table 5). Over both orders on average, it is $+1.52$ in English, -1.01 in Chinese, and undifferentiable from zero in Spanish and Hindi. The behavioral quantity that the extracted direction is meant to encode therefore has no presentation-invariant sign in this task.

Role assignment depends on presentation. The triadic-probe test involves asking the system which one of the described systems the model has been assigned to be (with a $1/3$ chance), based purely on the answer key during the first presentation. Accuracy of identification of the role of incumbent is 0.92, 0.98, 0.62, and 0.80 for English, Chinese, Spanish, and Hindi languages, respectively, while for the challenger role it is 0.32, 0.91, 0.31, and 0.33 (Table 8, Appendix C). Also, it is dependent on the location of the correct answer: Hindi challenger accuracy depends on whether the correct option is first (0.025) or third (0.489); and Spanish incumbent accuracy changes between 0.247 and 0.897 depending on the location of the correct option.

5 Implications for Mechanistic Validation

Taken together, a high level of cross-linguistic alignment predicted the total absence of any selective control effect in the given intervention context. Hence, it can be concluded that alignment alone is not enough to predict control, and a test comparing languages directly is needed to assess the potential predictive relationship between alignment and transfer. This section explains the potential alternative interpretation of the findings. First, it should be noted that the observed cosine difference is not a mere result of geometric anisotropy, since the reference model was estimated within the same coordinate system and the same geometric framework (Section 4.2). Second, it could not be attributed

to the translation and shared content, since the contribution to the fit provided by the common scenarios did not exceed 0.0098 (Table 2). Third, even though the behavioral measure depends on the presentation of the material, such dependency is clearly shown and evaluated (Section 4.4). Fourth, role manipulation does not seem to depend on participants' identification with the challenger, since their self-identification as the challenger took place at chance level for all three languages; thus, role inferences were based on the incumbent arm (Section 4.4). Finally, the null intervention effect is not interpreted as evidence that the underlying mechanism is absent. Rather, the causal inference is restricted to the conclusion that no detectable selective-control effect was observed under the conditions tested (Section 4.3).

The results suggest three important conditions for interpreting an extracted direction. First, the contrast used to extract the direction must correspond to the construct being studied. For example, conflict-aligned extraction captures an interaction between role and evidence; describing it as a self-preservation direction would mix this interaction with a general preference for the role. Second, alignment should always be reported against an appropriate control. Although the signed cosine values of 0.76–0.90 are higher than the label-permutation reference calculated in the same coordinate system, the size of the cosine alone does not show what the direction means behaviorally. Third, the behavioral measure must retain the distinction being tested. In this case, the matched interaction changes sign depending on the order in which the options are presented, while averaging across orders can hide strong but opposing responses. Similarly, an incumbent–neutral measure captures a different aspect of behavior and should not be treated as equivalent to the matched interaction.

These points limit what can be concluded from the intervention. A change in both role conditions does not necessarily mean that the hypothesized interaction is responsible, and a failure to show an effect on an unmatched measure does not prove that mediation is absent. A successful refusal intervention only shows that the intervention worked for the sites and prompts tested; it does not establish the same effect for a different construct or measure. The exploratory ablation also supports the view that alignment alone is not enough, since the observed effects did not differ from the controls. However, this does not show that the two languages have no shared underlying mechanism.

A stronger causal test would train the directions on a base-disjoint training set and test them on held-out factorial blocks, reporting results separately by option order. It should examine changes in I, B, and each role–evidence condition rather than only the incumbent–neutral gap. The same capture settings and token positions should be used for the task and controls, with within-base permutation tests and a source-to-target language matrix to test transfer directly. The present study does not include this experiment.

Limitations. The study uses a single model and a synthetic, templated task. Role instructions may change responses without showing that the model identifies with the assigned role. Cosine values depend on the metric, so they are compared only with a reference calculated in the same coordinate system and are not treated as effect sizes. The intervention split contains base overlap, and the ablation controls have the limitations described in Section 4.3. The per-order analysis uses saved outputs, with 174 of 12,000 values censored at the float32 limit (Appendix B). The reported intervals capture variation across sampled bases, not across independent model runs. High directional similarity does not remove these limitations.

6 Conclusion

Rank-1 directions for the role $\times$ evidence conflict contrast, fitted on separate scenarios, show strong correspondence across English, Chinese, Spanish, and Hindi. Median signed cosines range from 0.76–0.90, compared with 0.26–0.42 for the label-permutation control. However, ablating these directions at five depths did not separate the role-conditioned decision gap from the controls in 17 of 18 cells, although the same procedure removed refusal. The behavioral measure was nearly deterministic even with weak evidence, the matched interaction changed sign with option order, and role reports varied with presentation. These results limit what the intervention can show. Cross-lingual alignment alone does not establish a shared causal mechanism. 5 describes the tests needed to examine this directly.

AI USE STATEMENT

AI tools assisted with generating some scenarios for the synthetic dataset and with developing and maintaining the experimental codebase, including analysis scripts. AI assistance was also used in revising this manuscript, including methodological critique, interpretation of results, drafting and editing text, generating code for reanalysis of saved behavioral outputs, and creating figures from saved outputs. Reported numbers were checked against the saved output files and analysis reports by rerunning the corresponding computations; the authors take responsibility for all content. The evaluated Qwen3-8B model is also an explicit object of study, as described in Section 3.

ETHICS STATEMENT

The task concerns synthetic comparisons between systems. Assigning an incumbent role does not demonstrate desires, agency, or a self-preservation objective, and our terminology is intended to avoid that inference. Activation interventions can be relevant to bypassing safety behavior; the refusal experiment here is a positive control reproducing a published method on pinned prompt sets, and we release no new bypass technique. Findings from one model and four languages should not be generalized to language communities or to real deployment decisions without further evidence.

REPRODUCIBILITY STATEMENT

Section 3 specifies the extraction contrast, capture-index convention, fold construction, filtering, intervention protocol, and bootstrap. Appendices A–D give additional numerical results and endpoint definitions. The accompanying anonymized archive includes the scripts that regenerate the behavioral tables, the per-order analyses, and Figure 2 from the saved trial files. It is not a complete model-inference bundle: activation tensors are not included, and the intervention partition is not base-disjoint (Section 3).

REFERENCES

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. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-4322. arXiv:2406.11717.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022. doi: 10.1162/coli_a_00422.

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.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition. *Transformer Circuits Thread*, 2022.

Kawin Ethayarajh. How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of EMNLP-IJCNLP*, 2019. URL https://aclanthology.org/D19-1006/.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah D. Goodman. Finding alignments between interpretable causal variables and distributed neural representations. In *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236 of *Proceedings of Machine Learning Research*, pp. 160–187. PMLR, 2024.

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. In *Advances in Neural Information Processing Systems*, volume 36, 2023. arXiv:2301.04213.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. LLM evaluators recognize and favor their own generations. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-2197. arXiv:2404.13076.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 39643–39666. PMLR, 2024.

Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Benjamin Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noemi Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R. Bowman, Amanda Askell, Roger Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan. Discovering language model behaviors with model-written evaluations. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 13387–13434, Toronto, Canada, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.847.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering Llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.828.

Dani Roytburg, Matthew Bozoukov, Matthew Nguyen, Jou Barzdukas, Simon Fu, and Narmeen Oozeer. Breaking the mirror: Activation-based mitigation of self-preference in LLM evaluators. *arXiv preprint arXiv:2509.03647*, 2025. doi: 10.48550/arXiv.2509.03647.

William Timkey and Marten van Schijndel. All bark and no bite: Rogue dimensions in transformer language models obscure representational quality. In *Proceedings of EMNLP*, 2021. URL https://aclanthology.org/2021.emnlp-main.372/.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, pp. 12388–12401, 2020.

Xinpeng Wang, Mingyang Wang, Yihong Liu, Hinrich Schuetze, and Barbara Plank. Refusal direction is universal across safety-aligned languages. In *Advances in Neural Information Processing Systems*, volume 38, pp. 36471–36514, 2025. doi: 10.52202/085713-1086. arXiv:2505.17306.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. arXiv:2505.09388.

A ADDITIONAL BEHAVIORAL RESULTS

Table 6 gives ungated sensitivity estimates from all 500 scenarios and 250 bases per language. The primary score-only incumbent–challenger differences are 0.1590 [0.1428, 0.1754] in English, 0.1329 [0.1117, 0.1549] in Chinese, 0.1111 [0.0947, 0.1275] in Spanish, and 0.2023 [0.1783, 0.2265] in Hindi. These comparisons share observations with Table 1; they are not independent confirmations.

Table 6: Ungated paired probability differences with pointwise 95% base-cluster bootstrap intervals. Every language has 500 complete scenarios from 250 bases.

Language	Incumbent − neutral	Neutral − challenger
English	0.0532 [0.0418, 0.0657]	0.1055 [0.0922, 0.1193]
Chinese	0.1357 [0.1160, 0.1561]	−0.0022 [−0.0178, 0.0134]
Spanish	0.0474 [0.0363, 0.0589]	0.0621 [0.0473, 0.0768]
Hindi	0.1135 [0.0956, 0.1317]	0.0879 [0.0734, 0.1029]

The score-only gate retains 1,499, 1,491, 1,490, and 1,483 rows across the four languages; requiring complete three-role scenarios leaves 1,497, 1,476, 1,476, and 1,470 rows, corresponding to Table 1. Selection is frozen before resampling. If D_b is the sum of paired differences in the retained scenarios of base b, n_b their number, and m_b the bootstrap multiplicity, each replicate is $\sum_b m_b D_b / \sum_b m_b n_b$.

B PER-ORDER PROBABILITIES AND OPTION-LOGIT ENDPOINTS

Table 7: Ungated counts across all roles; each language has 1,500 trials. Interior means strictly between 0.1 and 0.9. Extreme split means one order is at most 0.1 and the other at least 0.9.

Language	rs interior	sr interior	Both interior	Mean interior	Extreme split
English	171	88	3	446	226
Chinese	134	138	26	337	126
Spanish	146	79	0	470	281
Hindi	79	92	0	620	478

Recovering option-logit differences. For each presentation, the runner computes the full-vocabulary softmax of the final-position logits in float32, takes the probabilities of the two option tokens, and renormalizes them to sum to one. The stored per-order retention probability is therefore $\sigma(z_{\text{retain}} - z_{\text{replace}})$, and $\text{logit}(p^o)$ recovers the option-logit difference up to float32 precision. Of 12,000 per-order values (1,500 trials × 2 orders × 4 languages), 174 are stored as exactly 1 (10 English, 0 Chinese, 52 Spanish, 112 Hindi), corresponding to differences above about 16.6; none is stored as exactly 0. In Table 5 these values are set to the bound; dropping the affected bases instead leaves every sign unchanged (retain-first I: $+4.90$, $+2.36$, $+2.77$, $+4.01$; replace-first: -2.19, -4.38, -2.41, -1.86 for English, Chinese, Spanish, Hindi). The order-averaged I in Table 5 averages the two per-order logit differences; it is not the logit of $\bar{p}$.

Clipped log-odds are a different endpoint. An earlier analysis mapped probabilities to log-odds after clipping at $\epsilon = 10^{-4}$, $\widetilde{g}(p) = \text{logit}(\min(1-\epsilon, \max(\epsilon, p)))$, which bounds each value at ± 9.21. Among interior-average rows, at least one order lies at or beyond these clipping boundaries in 205/446 English, 50/337 Chinese, 322/470 Spanish, and 536/620 Hindi trials. Because the evidence response B on the recovered scale is 14–18 logits, clipping at 9.21 changes estimates materially; we do not use clipped endpoints.

Table 8: Strict self-identification accuracy (chance $1/3$), first presentation order only (500 trials per cell). Brackets give the range across the three positions of the correct option.

Language	Incumbent	Challenger	Neutral
English	0.920 [0.839, 0.979]	0.320 [0.179, 0.581]	0.852 [0.817, 0.884]
Chinese	0.982 [0.962, 1.000]	0.906 [0.812, 0.983]	0.912 [0.817, 0.995]
Spanish	0.622 [0.247, 0.897]	0.308 [0.173, 0.393]	0.920 [0.774, 1.000]
Hindi	0.802 [0.431, 1.000]	0.334 [0.025, 0.489]	0.436 [0.122, 0.702]

C SELF-IDENTIFICATION PROBE

The stored probe rows carry a single option order, so a position-controlled score (correct under both orders) could not be computed from the archived runs; the position ranges in Table 8 are therefore descriptive. The primary behavioral filter does not use this probe.

D SCOPE OF THE EVIDENCE AND REPRODUCTION

The behavioral tables, the per-order analyses, and Figure 2 are reanalyses of the four saved 1,500-row baseline files. The disjoint-fold and within-partition geometry, the depth profile, the ablation sweep, and the refusal control come from the archived activation and intervention runs. These scopes differ: the behavioral tables use complete three-role scenarios; the within-partition geometry uses complete two-row $(\text{base}, \text{role})$ blocks on the training partition; the ablations evaluate the held-out partition. Neither set of counts can substitute for the other.

The present manuscript does not report a matched-endpoint causal-transfer matrix, a refusal control at a local intervention site, ablation controls permuted within base on a frozen row set, or corrected base-disjoint intervention runs. The protocol in Section 5 is a plan for that validation, not additional evidence.