REAL: REPAIRABILITY-AWARE LATENT ALIGNMENT FOR LOW-RESOURCE LANGUAGES

Anonymous authors
Paper under double-blind review

ABSTRACT

Multilingual language models remain less accurate in low-resource languages. Translation-based approaches change the input language, and fine-tuning updates model weights; activation steering preserves both but can also damage correct answers. We introduce REAL, an inference-time method that proposes cross-lingual latent interventions and selects among their executed outcomes using predicted repair and damage. REAL improves low-resource answers without changing the backbone, translating the test question, or querying paired high-resource states. On all 14,042 official MMMLU examples in each of four Yoruba/Swahili model–language settings, REAL improves Qwen3-8B and Llama-3.1-8B-Instruct by 6.65–11.12 percentage points. Across 56,168 predictions, accuracy rises by 9.04 points while preserving 98.56% of initially correct answers. With the same candidate library, verification adds 2.01 points over base-only selection and 1.51 over confidence gating. Permuting exact residual correspondence retains the gain under separately calibrated field-specific verification, whereas low-resource-only residuals trail by 4.53 points. On a full-test, automatically translated Tibetan extension, REAL gains 3.55 and 2.51 points across the two models. On AfriQA free generation, the library contains 3.38–8.06 F1 points of preserving-oracle headroom across four models, although the tested selectors do not deliver reliable gains. Together, these results establish repair-aware selection as an effective way to turn cross-lingual interventions into accuracy gains for low-resource choice scoring.

1 INTRODUCTION

Multilingual language models serve many languages, but their accuracy remains uneven across them (Hu et al., 2020; OpenAI, 2024). Translation pivots address low-resource questions by changing their input language (NLLB Team, 2022), while task-specific cross-lingual alignment can update model weights (Liu & Niehues, 2025). Inference-time activation steering keeps the original question and frozen backbone (Sundar et al., 2025; Maraia et al., 2026), making it attractive when translation or weight updates are unavailable.

Its difficulty is deciding which intervention to trust. A cross-lingual transport moves a low-resource hidden state using patterns learned from separate high- and low-resource examples, but a state that looks more aligned need not produce a better answer. In our Swahili diagnostics, a fixed transport improves semantic retrieval by 15.85 points for Qwen3-8B and 36.78 points for Llama-3.1-8B, while reducing answer accuracy by 5.35 and 0.88 points. The same intervention library contains thousands of correct alternatives to base errors. The problem is therefore to identify which candidate changes the answer for the better.

We address this problem through *candidate-specific repairability*: whether an executed intervention repairs a base error, damages a correct prediction, or leaves correctness unchanged. REAL separates two decisions. A cross-lingual residual field and a base-side proposal rule determine which transports to execute. A second decision, made after execution, uses the resulting evidence to predict repair and damage and choose between the candidates and identity. The verifier is supervised offline on a partition disjoint from both anchor construction and evaluation. At inference, it observes neither gold labels nor the test question's high-resource counterpart.

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.

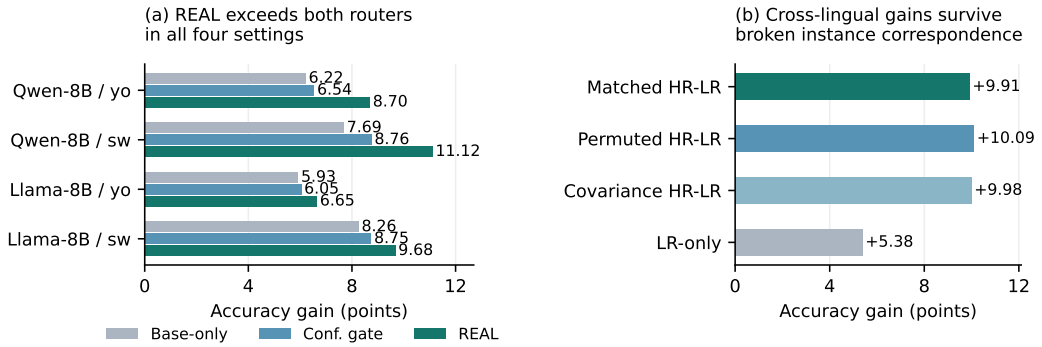


Figure 1: **Repairability selection improves every setting; useful residual structure survives disrupted correspondence.** Left: REAL outperforms both base-only selection and confidence gating in all four MMMLU settings ($N = 14{,}042$ each). Right: matched, permuted, and covariance-sampled HR–LR fields all outperform LR-only residuals on Qwen3-8B ($N = 28{,}084$ per field), with a separate verifier fitted for each field. These panels use different protocols: the left uses the main $k = 2$ policy; the right evaluates the full library.

REAL makes candidate-specific repair and damage the targets of selective steering. Conditional steering and calibrated abstention already determine when to intervene (Lee et al., 2025; Hedström et al., 2025). REAL instead evaluates *which executed cross-lingual candidate should replace the original prediction*, using joint base/candidate evidence and separate repair/damage targets. The distinction is operational: on the same intervention library and test examples, REAL gains 2.01 points over base-feature-only selection and 1.51 points over confidence/margin gating. Post-intervention selection adds value beyond identifying difficult inputs before intervention.

The results also refine what makes the candidate field useful. Matched, correspondence-permuted, and covariance-sampled high-resource residuals all yield approximately ten-point improvements under field-specific verification, whereas low-resource-only residuals yield 5.38 points. In these field-specific systems, successful repair does not require exact instance correspondence. The evidence favors reusable cross-lingual residual structure coupled with a readout-specific selector over a mechanism that depends on retrieving the uniquely correct semantic match.

Across both 8B model families and both official languages, REAL improves every full-test setting by 6.65–11.12 points. It recovers 5,357 of the preserving oracle's 6,585 available repairs while damaging 282 of 19,583 base-correct predictions. Figure 1 summarizes the two central findings: selection improves over strong routers across all four settings, and useful residual structure survives removal of exact correspondence. Our contributions are:

- **A repairability-based decision rule.** We formulate latent intervention selection using competing repair and damage events, with identity as an explicit alternative.
- **An effective cross-lingual intervention framework.** REAL combines a reusable transport field with post-intervention verification, achieving a 9.04-point pooled gain with 98.56% base-right preservation.
- **Cross-lingual repair without exact residual pairing.** Matched, permuted, and covariance-sampled fields retain approximately ten-point gains under field-specific verification. Selector controls independently establish the value of post-intervention evidence.

2 Related Work

Cross-lingual representations and intervention. Multilingual encoders learn shared representations through masked language modeling and cross-lingual objectives (Devlin et al., 2019; Pires et al., 2019; Conneau & Lample, 2019; Artetxe & Schwenk, 2019), yet multilingual benchmarks expose persistent language gaps (Hu et al., 2020; OpenAI, 2024). Recent work studies middle-layer alignment for transfer, instance-level alignment and patching, and shared concept spaces (Liu &

Model	Lang	N	Base	REAL	Δ [95% CI]	Pres.	W2R	R2W
Qwen3-8B	yo	14,042	31.78	40.47	+8.70 [8.20, 9.19]	97.94	1,313	92
Qwen3-8B	sw	14,042	37.28	48.40	+11.12 [10.60, 11.65]	99.81	1,571	10
Llama-3.1-8B-Instruct	yo	14,042	31.07	37.72	+6.65 [6.17, 7.12]	97.16	1,058	124
Llama-3.1-8B-Instruct	sw	14,042	39.33	49.01	+9.68 [9.17, 10.21]	98.99	1,415	56
Pooled	yo+sw	56,168	34.87	43.90	+9.04 [8.78, 9.29]	98.56	5,357	282

Table 1: Strict full-test MMMLU results. Each model–language cell calibrates a $k = 2$ policy on 1,991 non-test IDs and evaluates all 14,042 official test examples exactly once. Scores and intervals are percentage points; Pres. is base-right preservation.

Niehues, 2025; Ravisankar et al., 2026; Körner et al., 2026). Activation patching by Dumas et al. (2025) disentangles concept and language information using controlled translation and definition prompts. REAL uses a bank built from separate examples to propose interventions for unseen questions; its evaluation concerns answer repair rather than identifying concept representations. Building on activation addition and contrastive activation addition (Turner et al., 2023; Rimsky et al., 2024), cross-lingual steering changes multilingual computation without weight updates (Sundar et al., 2025; Maraia et al., 2026; Mahmoud et al., 2026). Language-neuron amplification can improve target-language behavior while degrading cross-language transfer (Rahmanisa et al., 2025); ReCoVeR instead preserves task performance while controlling answer language (Sterz et al., 2025). CLAS rescales shared and language-specific MLP activations and reports that task gains need not track proximity to English (Pokharel et al., 2026). REAL targets a different decision: which executed candidate repairs the answer without damaging a correct one.

Selective steering and repair. CAST gates behavior vectors using activation-based conditions (Lee et al., 2025). MERA learns error probes, derives adaptive steering magnitudes, and calibrates thresholds with an abstention criterion (Hedström et al., 2025). Its online rule uses an error estimate at the intervention state; calibration evaluates intervention outcomes. REAL instead observes executed candidate trajectories and predicts their joint repair/damage events before selecting an output. Both exploit supervision and selective intervention, but differ in their online decision variables and objectives. Our base-only and confidence controls test the value of the decision stage within REAL's intervention library. Selective prediction and model routing provide broader context (Geifman & El-Yaniv, 2017; 2019; Chen et al., 2023). Appendix O compares these decision rules and situates REAL relative to translation and knowledge editing.

3 Problem and Empirical Premise

For a low-resource prompt x_ℓ, answer set $\mathcal{C}$, and frozen model scores $s_0(c \mid x_\ell)$, let $\mathcal{I}$ denote the identity operator. Its unmodified trajectory predicts

$$\hat{y}_0 = \arg\max_{c \in \mathcal{C}} s_0(c \mid x_\ell). \tag{1}$$

A parameterized transport $\mathcal{T}_\eta$, with specification η, induces a counterfactual prediction $\hat{y}_\eta$; REAL returns $\hat{y}_{\mathrm{REAL}}$. We optimize accuracy while protecting examples already correct under $\mathcal{I}$. Base-right preservation is

$$\mathrm{Pres} = \frac{|\{i : \hat{y}_{0,i} = y_i \wedge \hat{y}_{\mathrm{REAL},i} = y_i\}|}{|\{i : \hat{y}_{0,i} = y_i\}|}. \tag{2}$$

For each non-identity transport, define repair and damage events

$$G^+(x, \eta) = \mathbb{1}[\hat{y}_0 \neq y, \hat{y}_\eta = y], \tag{3}$$

$$G^-(x, \eta) = \mathbb{1}[\hat{y}_0 = y, \hat{y}_\eta \neq y]. \tag{4}$$

These labels are available only on non-test calibration data. They distinguish recovering a base error from destroying a correct prediction. For $\lambda = 1$, $G^+ - G^-$ is exactly the change in the correctness indicator, so Eq. 5 estimates conditional accuracy improvement. Larger λ assigns additional cost to damage; the derivation is given in Appendix P.

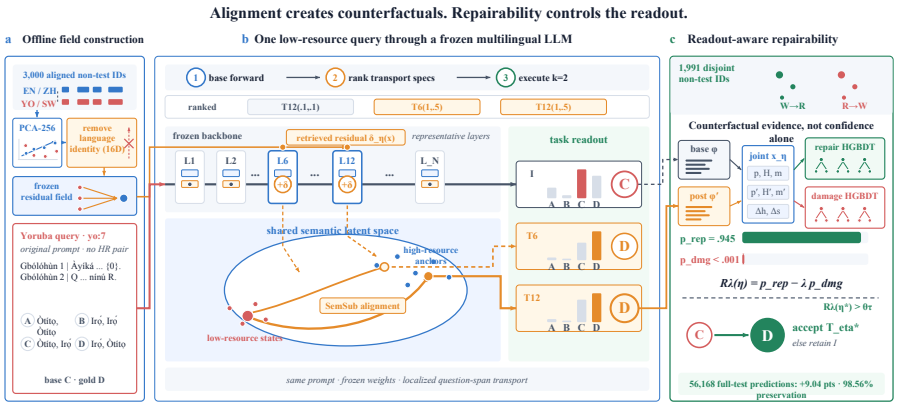


Figure 2: REAL makes aligned counterfactuals reachable and verifies their readout. **(a)** Disjoint non-test partitions construct a PCA-256 residual field after removing a 16-dimensional language-identity subspace (3,000 IDs), and fit/calibrate the verifier (1,991 IDs). **(b)** In an audited Yoruba trace, a compact frozen-backbone ribbon marks L6/L12 as intervention sites and elides other layers through L_N. Below it, orange SemSub trajectories move low-resource states toward high-resource semantic anchors; three aligned readout cards show identity C and two counterfactual D outcomes. **(c)** Joint base/post features feed separate histogram-gradient-boosted repair/damage estimators; a candidate replaces identity only if $p_{\text{rep}} - \lambda p_{\text{dmg}} > \theta_\tau$.

The need for selective repair is empirical. Pooled fixed transports $\mathcal{T}_{12}^{.1,.1}$, $\mathcal{T}_{6}^{1,.5}$, and $\mathcal{T}_{12}^{1,.5}$ change accuracy by $-0.28/-2.76/-2.54$ points and preserve only 93.48/71.19/76.28% of base-right examples. Yet they can strongly improve semantic retrieval. For example, Qwen3-8B and Llama-3.1-8B on Swahili gain 15.85 and 36.78 semantic Top-1 points under $\mathcal{T}_{12}^{1,.5}$ while losing 5.35 and 0.88 accuracy points. Alignment creates reachable alternatives, but the task readout decides whether they are repairs.

We model this decision through damage-adjusted repair utility:

$$\mathcal{R}_\lambda(x,\eta) = \Pr(W{\to}R \mid \phi(x,\eta)) - \lambda \Pr(R{\to}W \mid \phi(x,\eta)), \tag{5}$$

where $\phi(x,\eta)$ contains observable base and counterfactual-trajectory features. For retained set $\mathcal{C}_k(x)$, let $\eta^* = \arg\max_{\eta\in\mathcal{C}_k(x)} \widehat{\mathcal{R}}_\lambda(x,\eta)$. REAL returns $\hat{y}_{\eta^*}$ only when its utility clears θ_τ; otherwise it returns $\hat{y}_0$ from $\mathcal{I}$. The threshold is selected on non-test validation data for target preservation τ. This is an empirical validation constraint, not a finite-sample guarantee of test preservation; calibrated thresholds may be negative.

The preserving oracle selects a correct transport only for a base-wrong example and otherwise retains $\mathcal{I}$. Its gain measures *candidate headroom*; the difference between oracle and REAL is *selection regret*. This separates failure to generate a useful trajectory from failure to recognize one. The oracle is diagnostic only and is never used to fit or run REAL.

4 REAL

Design principle. REAL separates transport proposal from output acceptance (Figure 2). One non-test partition parameterizes *where a state could move*; a disjoint partition learns *whether an executed move should replace identity*. PCA, prototypes, and residual retrieval instantiate the candidate field, while the verifier implements the readout-specific decision.

Cross-lingual candidate field. From 3,000 aligned non-test semantic IDs, REAL mean-pools question-span states

$$\bar{h}_L(x) = |Q(x)|^{-1} \sum_{t\in Q(x)} h_{L,t}(x). \tag{6}$$

English/Chinese states define a 256-dimensional PCA basis, after removing 16 English–Chinese language directions. The bank stores normalized low-resource prototypes and projected English-minus-low-resource residuals at the intervention layers. Prototypes are grouped by 64-way spherical k-means; retrieval uses the top 50 normalized neighbors within the predicted cluster when available.

Parameterized SemSub transport. A transport specification is $\eta = (L, \alpha, \rho)$: intervention layer, residual scale, and relative-norm bound. For projected state z and prototype neighbors $\mathcal{N}_{50}(x)$, it retrieves

$$r_\eta(x) = \sum_{i \in \mathcal{N}_{50}(x)} \mathrm{softmax}_i(\mathrm{sim}(z, z_i)/0.05) P(r_i). \tag{7}$$

The transport operator adds

$$\delta_\eta(x) = \mathrm{clip}_{\rho \|\bar{h}_L(x)\|_2}(\alpha r_\eta(x)) \tag{8}$$

to question-span tokens and leaves all other states unchanged; $\mathcal{I}(h) = h$ is the identity operator. We abbreviate $\mathcal{T}_{(L,\alpha,\rho)}$ as $\mathcal{T}_L^{\alpha,\rho}$. The frozen intervention library shared by both 8B model families is

$$\mathcal{L}_{\mathrm{REAL}} = \{\mathcal{T}_{12}^{.1,.1}, \mathcal{T}_6^{1,.5}, \mathcal{T}_{12}^{1,.5}\}. \tag{9}$$

Readout-specific verifier. The disjoint 1,991-ID non-test partition supplies candidate-level repair labels $W{\to}R$ and damage labels $R{\to}W$. Separate histogram-gradient-boosted estimators predict the two terms in Eq. 5. Base features include option probabilities, entropy, margin, prompt length, token counts, and answer-format confidence. Post-transport features include candidate entropy, entropy change, relative hidden-state change, semantic-gain proxies, prediction change, and damage/safe projection scalars. Gold labels, correctness, oracle transports, observed outcomes, and same-question high-resource states are forbidden at inference.

Both estimators are fitted on all non-identity training rows. Their binary targets are the joint repair and damage events in Eq. 4; rows outside the respective event receive target zero. Both event probabilities are estimated without knowing base correctness at inference. The risk weight λ exposes an explicit preservation tradeoff. The verifier predicts transport utility from observable features, with $\mathcal{I}$ as the fallback. Auxiliary predictors are trained offline while the backbone stays frozen.

Algorithm 1: Repairability-constrained transport
1. Base evaluation. Run $\mathcal{I}$ once; record the base prediction, hidden states, and base-side features.
2. Transport proposal. Rank $\eta \in \mathcal{L}_{\mathrm{REAL}}$ using the base-side prior and retain $\mathcal{C}_k(x)$ with $k = 2$.
3. Counterfactual execution. Execute only $\mathcal{T}_\eta$ for $\eta \in \mathcal{C}_k(x)$; compute label-free post-intervention features.
4. Verify or retain. Score the executed transports with $\widehat{\mathcal{R}}_\lambda$. Return $\hat{y}_{\eta^*}$ if its utility clears θ_τ; otherwise return $\hat{y}_0$.

Reproducibility contract. The 1,991 calibration IDs are assigned by a stable hash of subject and sample ID to 75% train and 25% validation partitions. Numeric features are zero-imputed with missingness indicators and scaled without mean centering; categorical features are missing-value imputed and one-hot encoded with unknown categories ignored. Both proposal and post-intervention estimators use histogram gradient boosting with 90 iterations, 31 leaves, learning rate 0.06, $\ell_2 = 0.05$, and seed 20260613. The proposal prior ranks $\eta \in \mathcal{L}_{\mathrm{REAL}}$ with $\hat{p}^+ - 2\hat{p}^-$ before retained transports are rescored. Constant empirical probabilities replace a classifier only for single-class calibration labels.

For $\lambda \in \{1, 2, 3, 5\}$ and $\tau \in \{.90, .95\}$, the risk weight and threshold are selected on the train/validation split of the 1,991 IDs and frozen before official evaluation. The main policy uses $k = 2$ and $\tau = .95$. Its primary cost is candidate count: one identity forward plus k intervention forwards. Table 5 reports the accuracy–compute tradeoff across candidate budgets.

Component	Main protocol
Anchor construction	3,000 non-test IDs; PCA 256; remove 16 language dims; 64 clusters; top 50
Verifier calibration	Disjoint 1,991 non-test IDs; $\lambda \in \{1, 2, 3, 5\}$; $\tau \in \{.90, .95\}$
Verifier model	75/25 hash split; HGB, 90 iterations, 31 leaves, lr .06, $\ell_2 = .05$
Candidate budget	$k = 2$ from $\mathcal{L}_{\mathrm{REAL}}$ before intervention; $\mathcal{I}$ always available
Evaluation	14,042 official test IDs/language; deterministic full test; no overlap
Uncertainty	5,000 paired bootstrap samples; analysis seed 20260715

Table 2: Self-contained protocol for the strict 8B results.

5 EXPERIMENTS

Models, languages, and task. We evaluate Qwen3-8B and Llama-3.1-8B-Instruct on all official MMMLU Yoruba and Swahili test examples, using deterministic next-token likelihood over answer labels Hendrycks et al. (2021); OpenAI (2024). Each model–language cell contains 14,042 examples. Transfer analyses cover automatically translated Tibetan MMMLU and official AfriQA gold-passage free generation Ogundepo et al. (2023). They test language transfer and the transition from choice scoring to generated answers.

Data roles. The intervention library was selected using non-test development data and frozen before official MMMLU evaluation. The 3,000-ID anchor partition and 1,991-ID verifier partition are disjoint subsets of multilingualized MMLU auxiliary data. Aligned English, Chinese, Yoruba, and Swahili questions may be used offline; official MMMLU test rows may not. The anchor partition builds geometry but supplies no verifier labels. The verifier partition supplies G^+/G^- labels but cannot alter the anchor bank or intervention library. No official test examples or outcomes inform library selection, anchor construction, verifier fitting, or policy calibration. Final evaluation is deterministic and full-test, after each model–language policy is frozen.

Comparisons. We compare $\mathcal{I}$ and each fixed transport in $\mathcal{L}_{\mathrm{REAL}}$ with base-feature-only selection, a validation-tuned confidence/margin gate, entropy-min and semantic-gain-max rerankers, a preserving oracle, and full-library residual-field replacements. The field controls are matched English-minus-low-resource residuals, a correspondence-permuted version of that bank, samples from its covariance with matched per-prototype norms, and low-resource-only deltas. Every learned field receives its own non-test-calibrated verifier. We additionally report full-test NLLB (NLLB Team, 2022) translation pivots for Qwen3-8B Yoruba/Swahili; translation changes the input protocol, whereas REAL keeps it fixed.

Metrics and leakage controls. We report accuracy, paired accuracy delta, 95% paired-bootstrap intervals, W2R, R2W, preservation, edit rate, oracle capture, and candidate count. Anchor, verifier, and official test IDs are mutually disjoint. All transition fields are recomputed against the canonical identity trajectory; selector inputs are audited against gold, correctness, oracle, and outcome leakage. A leave-language-out verifier audit tests transfer without target-language verifier labels, while retaining the target-language anchor bank. Reproduction requires model weights, hidden-state hooks, MMMLU, the translated non-test auxiliary corpus, and the released partitions and policies; a black-box API is insufficient.

All confidence intervals use 5,000 paired example-level bootstrap samples with analysis seed 20260715. The code-and-data supplement includes aggregate results and analysis scripts. The supplement provides the prompt format, feature exclusions, intervention-library provenance, sensitivity analyses, and task-transfer diagnostics.

6 RESULTS AND ANALYSIS

REAL improves every full-test cell. Table 1 shows gains from 6.65 to 11.12 points across both model families and languages. Pooled over 56,168 predictions, REAL raises accuracy from 34.87

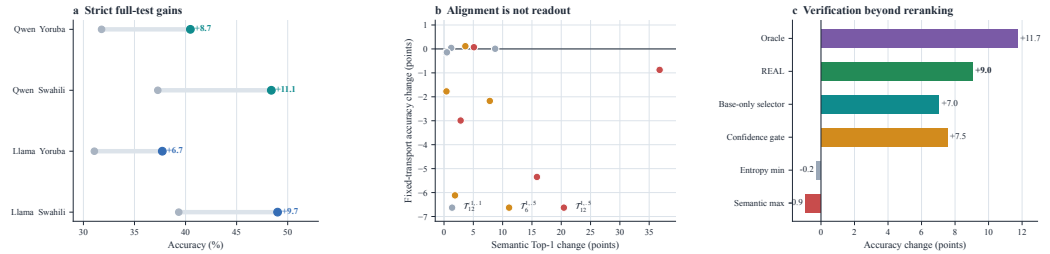


Figure 3: Evidence for selective alignment. (a) REAL improves every strict full-test model–language cell. (b) Fixed SemSub transports can improve semantic retrieval while reducing answer accuracy. (c) Ordinary reranking fails, base-side routing recovers only part of REAL's gain, and the preserving oracle measures remaining candidate-space headroom.

Policy	Δ	Pres.	W2R/R2W
Fixed $\mathcal{T}_{12}^{.1,.1}$	-0.28	93.48	1,118/1,277
Fixed $\mathcal{T}_{6}^{1,.5}$	-2.76	71.19	4,091/5,642
Fixed $\mathcal{T}_{12}^{1,.5}$	-2.54	76.28	3,218/4,645
Entropy minimum	-0.24	90.95	1,637/1,772
Semantic gain maximum	-0.90	84.28	2,572/3,079
Base-feature-only	+7.02	98.70	4,200/255
Confidence/margin	+7.52	100.00	4,226/0
REAL $k = 2$	**+9.04**	**98.56**	**5,357/282**
Preserving oracle	+11.72	100.00	6,585/0

Table 3: Pooled selector controls on the same 56,168 predictions. Delta and preservation are percentage points.

to 43.90, repairs 5,357 base errors, and damages 282 base-correct answers. The resulting 9.04-point gain has a 95% CI of [8.78,9.29] and 98.56% preservation. All four cells share the frozen library $\mathcal{L}_{\text{REAL}}$. Verifier fitting and policy calibration use disjoint non-test data.

Post-intervention selection adds value beyond routing. On the same candidate library and 56,168 test predictions, entropy-min and semantic-gain-max reranking yield -0.24 and -0.90 points, while base-feature-only selection and confidence/margin gating gain +7.02 and +7.52. REAL adds 2.01 paired points over base-only selection (95% CI [1.86,2.16]) and 1.51 over confidence gating [1.35,1.67]. Its advantage over base-only selection is positive in all four cells, ranging from 0.73 to 3.43 points (Figure 1). The gain comes from additional repairs: relative to confidence gating, REAL adds 1,131 repairs and 282 damages, for 849 net additional correct answers. REAL gains these 849 correct answers while preserving 98.56% of base-correct predictions; confidence gating preserves 100%. A preserving oracle repairs 6,585 examples with zero damage; REAL recovers 81.4% of these repairs.

Useful cross-lingual structure survives disrupted correspondence. Under Qwen3-8B full-library evaluation, matched, correspondence-permuted, and covariance-sampled HR–LR fields yield +9.91, +10.09, and +9.98 points. Matched-minus-permuted and matched-minus-covariance differences are -0.19 [-0.42, 0.04] and -0.08 [-0.32, 0.16]. LR-only deltas yield +5.38, trailing matched HR–LR by 4.53 points [4.18,4.88]. These results show that exact residual pairing is dispensable for the observed improvement. Cross-lingual residual structure supplies useful candidates even after exact correspondence is removed. All four fields receive their own verifier, so the comparison concerns the complete field-plus-selector system. The five-seed cross-language audit reaches the same conclusion: matched/permuted/covariance/LR-only fields yield +9.49/+9.91/+9.72/+5.67 points.

Transfer, cost, and translation. The independently calibrated $k = 1/2/3$ frontier yields +7.23/+9.04/+9.47 points with 2/3/4 total forwards, so $k = 2$ captures most available gain. NLLB

Residual field	Δ	Pres.	Oracle	Capture
Matched HR–LR	+9.91	98.95	+12.99	79.06
Permuted HR–LR	+10.09	98.43	+12.99	81.90
Covariance HR–LR	+9.98	98.37	+13.01	81.06
LR-only deltas	+5.38	98.14	+7.52	79.98

Table 4: Qwen3-8B full-library residual-field audit ($N = 28{,}084$ per field). Every field receives an independently calibrated verifier.

Budget	$k = 1$	$k = 2$	$k = 3$
Accuracy Δ	+7.23	+9.04	+9.47
Preservation	99.37	98.56	98.98
Total forwards	2	3	4
Estimated time (ms)	94	161	229

Table 5: Pooled strict transport-budget frontier. Times are reconstructed from saved evaluation-log throughput; they are not direct online latency measurements and exclude offline bank construction.

improves Qwen Yoruba/Swahili by +2.46/+3.92 under its own translation-pivot scoring pipeline. The pivot changes the input language; REAL obtains its gains while retaining the original low-resource prompt.

7 Transfer and Calibration

Verifier transfer across languages. In leave-language-out calibration, the proposal policy, repair/damage estimators, and threshold are fit on one language and frozen for the other; language identity is removed. Qwen yo→sw and sw→yo retain +11.52 and +8.21 points, while Llama retains +9.49 and +6.37. A stricter residual-field transfer audit averages five seeds and gives matched SemSub +9.49 points bidirectionally. The verifier transfers in both directions and both model families without target-language calibration labels. The target-language anchor bank remains available to construct candidates.

Calibration efficiency. With five fixed seeds, Qwen3-8B matched-field gains at 50/100/250/500/1,991 calibration IDs are +6.70/+8.76/+8.82/+9.63/+9.91 points. Using 100 IDs already yields +8.76 points, compared with +9.91 using all 1,991 IDs. The appendix reports geometry sensitivity and a 50-example 14B diagnostic.

Tibetan extends the end-to-end result. On 14,042 automatically translated Tibetan questions, both 8B models improve with candidate selection. Qwen rises from 29.20% to 32.75% (+3.55 points; 94.46% preservation); Llama rises from 28.93% to 31.43% (+2.51 points; 97.66% preservation). In contrast, Llama's fixed interventions yield 0.00, -0.20, and -0.07 points across the three actions. Its repair-aware selector recovers +2.51 points from this same library. Tibetan therefore repeats the central pattern on a third language: selecting interventions matters more than applying one transport to every question. The translated test has no human quality audit (Appendix L).

Free-generation readout transfer. On AfriQA, preserving-oracle F1 gains remain positive across Qwen-8B/4B and Llama-8B/3B (+8.06/+6.50/+4.00/+3.38; $N = 548$ each), but the corresponding protocol-primary selectors yield only +0.00/+0.51/+0.86/+0.13 and none has a positive 95% lower bound. This distinguishes transfer of useful candidates from transfer of their verification. Full results and selector definitions appear in Appendix K.

8 Discussion

Candidate reachability and decision quality are separate resources. Unconditional transports reduce pooled accuracy, yet their candidate set admits 6,585 preserving repairs. REAL realizes 5,357

Model	Cal.→test	Δ [95% CI]	Pres.
Qwen3-8B	yo→sw	+11.52 [10.97,12.06]	98.97
Qwen3-8B	sw→yo	+8.21 [7.74,8.68]	99.19
Llama-3.1-8B	yo→sw	+9.49 [9.00,9.99]	99.44
Llama-3.1-8B	sw→yo	+6.37 [5.96,6.79]	99.56

Table 6: Language-blind leave-language-out calibration. No target-language verifier label is used.

of those repairs while incurring 282 damages. These counts explain why evaluating only alignment quality or only unconditional intervention accuracy misses the useful operating regime. Verification converts available correct alternatives into selected answers by estimating their repair and damage probabilities.

Residual structure matters more than exact pairing in these experiments. The strongest residual controls preserve cross-lingual statistics while destroying instance correspondence. Their success shifts the explanation from precise semantic matching to the availability of useful perturbations within a cross-lingual residual field. A candidate generator can therefore exploit reusable distributional structure, while a separate predictor resolves its task-specific utility. The finding locates the useful structure at the level of the residual field and its task-specific selection.

Post-intervention evidence improves the decision. A difficult input does not determine whether a particular transport will repair it. Candidate-specific changes in entropy, prediction, and hidden-state geometry provide additional evidence about that transport's outcome. Across all four MMMLU settings, post-intervention selection outperforms base-only routing, establishing the value of this decision stage.

9 Limitations and Responsible Use

The complete official MMMLU evaluation spans two 8B model families and Yoruba/Swahili. Tibetan adds a full-test translated language extension, and the 14B diagnostic covers 50 examples. Broader model scales and human-audited Tibetan data are natural next tests.

REAL requires an offline anchor bank and labeled verifier calibration. Cross-language transfer removes target-language verifier labels while retaining target-language anchors. The main policy uses three backbone forwards. Its preservation threshold is calibrated empirically; reliability under severe feature degradation or distribution shift remains unmeasured. Language-specific deployment also requires checks of factuality, cultural validity, and translation quality.

Routing ablations isolate decision-stage effects within the frozen intervention library, while the MERA/CLAS comparison concerns algorithmic design. Residual-field comparisons evaluate independently calibrated field-plus-selector systems. Fine-tuning and translation involve different training or input protocols from the frozen-backbone setting studied here.

10 Conclusion

REAL converts cross-lingual intervention headroom into measured prediction gains by learning which executed candidates repair rather than damage an answer. Across four full-test settings, it improves a frozen backbone by 6.65–11.12 points and exceeds strong base-only and confidence-based selection. Residual controls further show that useful cross-lingual structure need not preserve exact instance correspondence. These findings establish the value of separating cross-lingual candidate generation from repair-aware selection. A reusable residual field supplies alternatives; candidate-specific verification converts them into net accuracy gains.

AI Use Statement

Generative AI assisted with manuscript revision and comparison of related literature.

REFERENCES

Mikel Artetxe and Holger Schwenk. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. In *Transactions of the Association for Computational Linguistics*, 2019.

Lingjiao Chen, Matei Zaharia, and James Zou. FrugalGPT: How to use large language models while reducing cost and improving performance. *arXiv preprint arXiv:2305.05176*, 2023.

Alexis Conneau and Guillaume Lample. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, 2019.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, 2019.

Clément Dumas, Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. Separating tongue from thought: Activation patching reveals language-agnostic concept representations in transformers. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 31822–31841. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.1536. URL https://aclanthology.org/2025.acl-long.1536/.

Junfeng Fang, Houcheng Jiang, Kun Wang, Yunshan Ma, and Xiang Wang. AlphaEdit: Null-space constrained knowledge editing for language models. *arXiv preprint arXiv:2410.02355*, 2024.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, 2017.

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. In *Proceedings of ICML*, 2019.

Anna Hedström, Salim I. Amoukou, Tom Bewley, Saumitra Mishra, and Manuela Veloso. To steer or not to steer? mechanistic error reduction with abstention for language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 22924–22945. PMLR, 2025. URL https://proceedings.mlr.press/v267/hedstrom25a.html.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *Proceedings of ICLR*, 2021.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. In *Proceedings of ICML*, 2020.

Felicia Körner, Max Müller-Eberstein, Anna Korhonen, and Barbara Plank. When meanings meet: Investigating the emergence and quality of shared concept spaces during multilingual language model training. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3149–3169. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.eacl-long.145. URL https://aclanthology.org/2026.eacl-long.145/.

Bruce W. Lee, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Erik Miehling, Pierre Dognin, Manish Nagireddy, and Amit Dhurandhar. Programming refusal with conditional activation steering. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Oi47wc10sm.

Danni Liu and Jan Niehues. Middle-layer representation alignment for cross-lingual transfer in fine-tuned LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15979–15996. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.778. URL https://aclanthology.org/2025.acl-long.778/.

Omar Mohamed Mahmoud, Buddhika Laknath Semage, Thommen George Karimpanal, and Santu Rana. Improving multilingual language models by aligning representations through steering. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference*, pp. 2090–2103, 2026. doi: 10.63317/244dsoue8zu2. URL https://aclanthology.org/2026.lrec-1.164/.

Gabriele Maraia, Leonardo Ranaldi, Marco Valentino, and Fabio Massimo Zanzotto. Can activation steering generalize across languages? a study on syllogistic reasoning in language models. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2739–2753. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.eacl-long.125. URL https://aclanthology.org/2026.eacl-long.125/.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, 2022.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. In *Proceedings of ICLR*, 2023.

NLLB Team. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*, 2022.

Odunayo Ogundepo et al. AfriQA: Cross-lingual open-retrieval question answering for African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 14957–14972. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-emnlp.997. URL https://aclanthology.org/2023.findings-emnlp.997/.

OpenAI. MMMLU: Multilingual massive multitask language understanding. https://huggingface.co/datasets/openai/MMMLU, 2024.

Telmo Pires, Eva Schlinger, and Dan Garrette. How multilingual is multilingual BERT? In *Proceedings of ACL*, 2019.

Rhitabrat Pokharel, Ameeta Agrawal, and Tanay Nagar. Cross-lingual activation steering for multilingual language models. *arXiv preprint arXiv:2601.16390*, 2026. URL https://arxiv.org/abs/2601.16390.

Inaya Rahmanisa, Lyzander Marciano Andrylie, Mahardika Krisna Ihsani, Alfan Farizki Wicaksono, Haryo Akbarianto Wibowo, and Alham Fikri Aji. Unveiling the influence of amplifying language-specific neurons. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 919–968, 2025. doi: 10.18653/v1/2025.findings-ijcnlp.55. URL https://aclanthology.org/2025.findings-ijcnlp.55/.

Kartik Ravisankar, HyoJung Han, Sarah Wiegreffe, and Marine Carpuat. Can you map it to English? the role of cross-lingual alignment in the multilingual performance of LLMs. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4854–4872. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.eacl-long.225. URL https://aclanthology.org/2026.eacl-long.225/.

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

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. Language models are multilingual chain-of-thought reasoners. *arXiv preprint arXiv:2210.03057*, 2023.

Hannah Sterz, Fabian David Schmidt, Goran Glavaš, and Ivan Vulić. ReCoVeR the target language: Language steering without sacrificing task performance. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 19390–19405, 2025. doi: 10.18653/v1/2025.findings-emnlp.1056. URL https://aclanthology.org/2025.findings-emnlp.1056/.

Anirudh Sundar, Sinead Williamson, Katherine Metcalf, Barry-John Theobald, Skyler Seto, and Masha Fedzechkina. Steering into new embedding spaces: Analyzing cross-lingual alignment induced by model interventions in multilingual language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2375–2401. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.118. URL https://aclanthology.org/2025.acl-long.118/.

Alexander Matt Turner, Lisa Thiergart, David Udell, Gavin Leech, Ulisse Mini, and Monte MacDiarmid. Activation addition: Steering language models without optimization. *arXiv preprint arXiv:2308.10248*, 2023.

A STRICT REPRODUCIBILITY PROTOCOL

The intervention library was chosen using non-test development data and frozen before official MMMLU evaluation. The main result then uses three mutually disjoint data roles. A 3,000-ID multilingualized MMLU-aux partition constructs the SemSub anchor bank. A separate 1,991-ID non-test MMLU-aux partition fits the verifier and selects its risk weight and threshold. Evaluation uses all 14,042 official MMMLU test examples for each model–language cell. Neither official test examples nor their outcomes were used to choose the library, construct anchors, fit the verifier, or calibrate its risk weight and threshold. Artifact aliases A0, A1, A2, and A3 map respectively to $\mathcal{I}$, $\mathcal{T}_{12}^{.1,.1}$, $\mathcal{T}_{6}^{1,.5}$, and $\mathcal{T}_{12}^{1,.5}$; these IDs are provenance labels, not method primitives.

Data role	Size	Permitted use
Anchor partition	3,000 IDs	PCA, language removal, prototypes, and residual bank
Verifier partition	1,991 IDs	Selector fitting and validation-only (λ, θ_τ) choice
Official test	14,042 per language	One deterministic final evaluation after policy freezing

Table 7: Data separation for every strict 8B result. The anchor and verifier partitions are non-test MMLU-aux data and are mutually disjoint.

A.1 SEMSUB TRANSPORT CONSTRUCTION

Let $Q(x)$ be the question-token span and $h_{L,t}(x)$ the hidden state at layer L. The pooled low-resource state is

$$\bar{h}_L(x) = |Q(x)|^{-1} \sum_{t \in Q(x)} h_{L,t}(x). \tag{10}$$

The anchor partition provides aligned low-resource and English/Chinese states. We fit a 256-dimensional PCA basis, estimate and remove a 16-dimensional English–Chinese language subspace, and store normalized low-resource prototypes together with projected high-minus-low residuals. Prototypes are grouped by 64-way spherical k-means. Retrieval uses the 50 nearest normalized prototypes, restricted to the predicted cluster when at least five members are available. With temperature $T = 0.05$ and specification $\eta = (L, \alpha, \rho)$, the retrieved residual is

$$r_\eta(x) = \sum_{i \in \mathcal{N}_{50}(x)} \frac{\exp(\mathrm{sim}(z, z_i)/T)}{\sum_j \exp(\mathrm{sim}(z, z_j)/T)} P(r_i), \tag{11}$$

where z is the projected pooled state and P maps a residual through the retained semantic basis. The transport update is

$$\delta_\eta(x) = \alpha r_\eta(x) \min\left(1, \frac{\rho \|\bar{h}_L(x)\|_2}{\alpha \|r_\eta(x)\|_2 + \epsilon}\right), \tag{12}$$

and the same $\delta_\eta(x)$ is added only to question-span tokens. Let $\mathcal{I}(h) = h$ and abbreviate $\mathcal{T}_{(L,\alpha,\rho)}$ as $\mathcal{T}_L^{\alpha,\rho}$. The deployed library is $\mathcal{L}_{\mathrm{REAL}} = \{\mathcal{T}_{12}^{.1,.1}, \mathcal{T}_{6}^{1,.5}, \mathcal{T}_{12}^{1,.5}\}$.

A.2 VERIFIER AND CALIBRATION CONTRACT

Candidate-level supervision contains two non-test labels: repair $y_\eta^+ = \mathbb{1}[\mathcal{I}\text{ wrong}, \mathcal{T}_\eta\text{ correct}]$ and damage $y_\eta^- = \mathbb{1}[\mathcal{I}\text{ correct}, \mathcal{T}_\eta\text{ wrong}]$. Separate estimators produce $\hat{p}_\eta^+$ and $\hat{p}_\eta^-$, and the validation objective is

$$\widehat{\mathcal{R}}_\lambda(x, \eta) = \hat{p}_\eta^+ - \lambda \hat{p}_\eta^-, \qquad \lambda \in \{1, 2, 3, 5\}. \tag{13}$$

Base-side features include option probabilities, entropy, margin, prompt length, token counts, fertility, and answer-format confidence. Post-intervention features include candidate entropy, entropy change, relative hidden-state change, semantic-gain proxies, prediction change, and available damage/safe projection scalars. Per-example gold labels, correctness, oracle choices, and repair/damage outcomes are blacklisted at inference.

Model	Lang	N	Base	REAL	Δ [95% CI]	Pres.	W2R	R2W
Qwen3-8B	yo	14,042	31.78	40.47	+8.70 [8.20,9.19]	97.94	1,313	92
Qwen3-8B	sw	14,042	37.28	48.40	+11.12 [10.60,11.65]	99.81	1,571	10
Llama-3.1-8B	yo	14,042	31.07	37.72	+6.65 [6.17,7.12]	97.16	1,058	124
Llama-3.1-8B	sw	14,042	39.33	49.01	+9.68 [9.17,10.21]	98.99	1,415	56
Pooled	yo+sw	56,168	34.87	43.90	+9.04 [8.78,9.29]	98.56	5,357	282

Table 8: Strict $k = 2$ headline results. Percentages are computed over all official test IDs. W2R/R2W are transitions relative to the canonical identity trajectory.

For each model–language cell, the base-side prior retains the top k transports from $\mathcal{L}_{\mathrm{REAL}}$. The verifier selects the maximum-utility transport only when its utility exceeds θ_τ; otherwise it retains $\mathcal{I}$. Both λ and θ_τ are selected on the validation portion of the 1,991 IDs for $\tau \in \{0.90, 0.95\}$ and frozen before official test evaluation. The main paper uses $k = 2$ and $\tau = 0.95$.

The 1,991 IDs are assigned by a stable hash of subject and sample ID to 75% train and 25% validation partitions. Numeric features are zero-imputed with missingness indicators and scaled without mean centering; categorical features are missing-value imputed and one-hot encoded. Both proposal and verifier estimators use histogram gradient boosting with 90 iterations, 31 leaves, learning rate 0.06, $\ell_2 = 0.05$, and seed 20260613. The proposal prior ranks transports by $\hat{p}^+ - 2\hat{p}^-$ before the post-intervention models rescore the retained set.

B Strict Full-Test Results

Exact selected-row hashes are retained for each cell. The Qwen3-8B rows independently replay the strict P25 full-library audit, and the $k = 2$ rows also replay exactly inside the P29 cost analysis. Pooled REAL repairs 5,357 examples and damages 282. The preserving oracle repairs 6,585 with zero base-right damage, so REAL captures 81.4% of candidate-space repair opportunities and leaves 2.69 points of oracle regret.

C Selector and Reranking Controls

Every control below uses the same 56,168 official test predictions and canonical identity definition. Learned or tuned controls use only the non-test verifier split and validation-only policy selection.

Policy	Δ	Pres.	W2R	R2W
Fixed $\mathcal{T}_{12}^{.1,.1}$	-0.28	93.48	1,118	1,277
Fixed $\mathcal{T}_{6}^{1,.5}$	-2.76	71.19	4,091	5,642
Fixed $\mathcal{T}_{12}^{1,.5}$	-2.54	76.28	3,218	4,645
Base-feature-only	+7.02	98.70	4,200	255
Confidence/margin	+7.52	100.00	4,226	0
Entropy-min reranker	-0.24	90.95	1,637	1,772
Semantic-gain-max	-0.90	84.28	2,572	3,079
REAL $k = 2$	**+9.04**	98.56	5,357	282
Preserving oracle	+11.72	100.00	6,585	0

Table 9: Pooled strict selector controls ($N = 56{,}168$). Scores are percentage points.

REAL exceeds base-feature-only selection by +2.01 points (95% paired-bootstrap CI [1.86,2.16]) and confidence/margin routing by +1.51 [1.35,1.67]. It exceeds entropy-min and semantic-gain-max reranking by +9.28 and +9.94 points. This isolates post-intervention repairability evidence from generic candidate ranking, while the strong base-only controls show that low-confidence routing remains an important part of the effect.

Model	Lang	Δ	Pres.	Transport	Est. ms
Qwen3-8B	yo	6.51/8.70/9.23	99.22/97.94/98.36	95.05/98.95/98.37	105/181/257
Qwen3-8B	sw	8.07/11.12/12.04	99.41/99.81/99.18	95.43/29.55/97.73	102/173/244
Llama-8B	yo	5.99/6.65/6.87	99.54/97.16/98.95	93.58/99.39/99.02	90/156/223
Llama-8B	sw	8.34/9.68/9.74	99.31/98.99/99.33	98.46/99.19/99.31	79/135/191
Pooled	all	7.23/9.04/9.47	99.37/98.56/98.98	95.63/81.77/98.61	94/161/229

Table 11: Strict cost frontier. Entries within each cell correspond to $k = 1/2/3$, or 2/3/4 total forwards. Delta, preservation, and transport selection are percentages. The non-monotone selection rate follows independently calibrated preservation thresholds and is not a compute measure.

D Full-Library Residual-Field Ablations

P25 regenerates the complete $\mathcal{L}_{\mathrm{REAL}}$ library for each candidate field on Qwen3-8B and gives every field its own non-test-calibrated verifier. The fields isolate levels of cross-lingual structure: matched SemSub uses the retrieved English-minus-low-resource residual; the permuted field shuffles those same residuals across prototypes; the covariance field samples the English-minus-low-resource residual covariance and matches per-prototype norms; the LR-only field uses shuffled low-resource-minus-low-resource deltas projected through the same basis. The first three retain high-resource-derived transport geometry, while only the first retains exact prototype correspondence.

Field	N	Δ	Pres.	Oracle Δ	Capture
Matched HR–LR	28,084	+9.91	98.95	+12.99	79.06
Permuted HR–LR	28,084	+10.09	98.43	+12.99	81.90
Covariance HR–LR	28,084	+9.98	98.37	+13.01	81.06
LR-only deltas	28,084	+5.38	98.14	+7.52	79.98

Table 10: Strict Qwen3-8B full-library residual-field ablations. Capture is the percentage of preserving-oracle repairs recovered by the field-specific verifier. HR–LR denotes an English-minus-low-resource residual field.

The paired matched-minus-permuted estimate is -0.19 points with CI [-0.42,0.04], and matched-minus-covariance is -0.08 [-0.32,0.16]. Exact prototype correspondence therefore adds no detected accuracy advantage after field-specific calibration. In contrast, the matched high-to-low field exceeds LR-only deltas by +4.53 [4.18,4.88]. This localizes the supported alignment claim: high-resource-derived residual geometry is effective, while the current task gain is distributional rather than dependent on instance-level semantic matching. Because the competitive controls inherit the high-to-low residual bank or its covariance, they are not arbitrary-noise baselines.

E Strict Transport-Budget Frontier

Each k below receives an independent validation-only choice of (λ, θ) under the same strict protocol. Executed transports and intervention forwards equal k; total forwards include the identity evaluation. Estimated milliseconds are reconstructed from saved evaluation-log throughput and are not online deployment measurements.

Moving from $k = 1$ to $k = 2$ adds 1.81 pooled points for one extra intervention forward. Moving from $k = 2$ to $k = 3$ adds only 0.43 points. Non-identity transports are selected on 95.63/81.77/98.61% of examples, but the final answer label changes on only 21.79/26.98/28.07%. This is why $k = 2$ is the main transport-budget setting. "Transport-budgeted" describes candidate-count limitation only; REAL is not claimed to be latency-optimized.

F Translation-Pivot Baseline

NLLB-200-distilled-600M translates the full Qwen3-8B Yoruba and Swahili test inputs to English before option scoring. It improves Yoruba from 25.98 to 28.44 (+2.46) and Swahili from 27.35 to

31.27 (+3.92). Strict REAL improves its own low-resource base by +8.70 and +11.12. These values are informative but not a unified system comparison: NLLB changes the input and scoring pipeline, whereas REAL keeps the low-resource prompt fixed. Llama full-test NLLB, commercial MT, and a fully unified direct/NLLB/REAL scoring implementation are not included.

G Calibration and Transfer Sensitivity

P26 varies the non-test verifier sample size over $\{50, 100, 250, 500, 1000, 1991\}$ with five fixed seeds. Each residual source receives its own train/validation policy, which is then frozen before full-test evaluation. The 1,991-ID rows exactly replay P25.

Audit	Matched	Permuted	Covariance	LR-only
$n_{cal} = 50$	+6.70	+6.98	+6.45	+1.90
$n_{cal} = 100$	+8.76	+9.15	+9.09	+4.21
$n_{cal} = 250$	+8.82	+9.48	+9.54	+5.42
$n_{cal} = 500$	+9.63	+9.70	+9.81	+5.13
$n_{cal} = 1991$	+9.91	+10.09	+9.98	+5.38
Cross-language	+9.49	+9.91	+9.72	+5.67

Table 12: P26 pooled Qwen3-8B accuracy deltas for the same residual-field definitions as Table 10. Calibration rows average five seeds over Yoruba and Swahili; cross-language trains on one language and tests the other without target-language verifier labels.

The verifier transfers across languages: matched SemSub retains +9.49 points under bidirectional leave-language-out evaluation. Across the preregistered 50–500 low-shot aggregate, matched minus permuted is -0.35 points (95% CI $[-0.48, -0.22]$) and matched minus covariance is -0.24 $[-0.39, -0.11]$. Under cross-language transfer, the differences are -0.42 $[-0.60, -0.24]$ and -0.23 $[-0.41, -0.04]$. Both preregistered instance-matching-advantage gates therefore fail. This does not negate cross-lingual alignment: permuted and covariance fields still derive from English-minus-low-resource transport. It shows that exact semantic correspondence is not the source of the measured calibration or transfer advantage. Determining which statistics of that field are minimally sufficient is a separate mechanism question.

H Feature, Leakage, and Canonical-Base Audits

Deployable REAL never accesses a test example's English/Chinese counterpart, gold answer, correctness, oracle transport, or observed repair outcome. Same-question high-resource hidden patches are mechanism diagnostics only. The feature blacklist enforces this boundary, and aggregate prior features are computed only from non-test data. The canonical-base audit recomputes all transition and prediction-change fields against the explicit identity row before fitting any strict selector. The original run workspace retains the P26–P29 policy hashes and validation grids for exact replay.

Reproduction requires model weights, hidden-state hooks, official MMMLU data, the translated non-test MMLU-aux anchor/calibration corpus, and the frozen 3,000/1,991 ID partition. A black-box API is insufficient because SemSub requires question-span pooling and layer-level intervention. The released aggregate artifacts do not redistribute upstream datasets, model weights, hidden-state tensors, or large per-example prediction dumps.

I Prompt and Evaluation

All MMMLU examples use the following shape:

```
Question: {question}
A. {choice_0}
B. {choice_1}
C. {choice_2}
D. {choice_3}
Answer:
```

Answers are scored by deterministic next-token likelihood over A–D. Base accuracy, repaired accuracy, W2R/R2W, preservation, and edit rate are computed from the same canonical identity predictions. Confidence intervals use paired example-level bootstrap resampling.

J Auxiliary Evidence and Scope Boundaries

Semantic movement versus readout. The 8B retrieval diagnostic is used only to establish representation–readout mismatch. Qwen3-8B Swahili $\mathcal{T}_{12}^{1,.5}$ improves semantic Top-1 by 15.85 points while reducing answer accuracy by 5.35; Llama-3.1-8B Swahili $\mathcal{T}_{12}^{1,.5}$ improves Top-1 by 36.78 while reducing accuracy by 0.88. These rows motivate verification and are not headline accuracy results.

Qwen3-4B. The 4B ladder documents method evolution and mechanism ablations, not the source of the 8B claim. At the 0.95 preservation target, base-only, family-first, and $k = 2$ verification improve accuracy by +1.12, +1.38, and +1.90 points, respectively, with 95.45/96.51/97.47% preservation.

SemSub hyperparameter sensitivity. An auxiliary Qwen3-4B mechanism grid contains 576 configurations over PCA dimensions $\{64, 128, 256\}$, removed language dimensions $\{0, 16, 32\}$, prototype counts $\{10, 50\}$, layers $\{6, 8, 10, 12\}$, and full/semantic neighbor spaces. Averaged semantic Top-1 gains rise from 6.47/9.92/9.11 points at 64 dimensions with 0/16/32 removed directions, to 11.48/16.86/16.96 at 256 dimensions. Prototype counts 10 and 50 are nearly matched (11.83 versus 11.70 points), and layer averages range from 10.62 to 12.67 points. This post-hoc 4B representation-level grid is reported only as a robustness diagnostic; it was not used to choose the strict 8B policy, and it is not an 8B selector-level hyperparameter sweep. The 64-cluster choice remains fixed and is not covered by this sensitivity audit.

Model scale. A Qwen3-14B-AWQ $N = 50$ sanity check reproduces semantic/readout mismatch but is not a full REAL experiment. The positive headline claim is limited to two 8B model families.

K Open-Generation Candidate Transfer

We evaluate whether the frozen transport library remains useful when the answer is no longer one of four fixed labels. AfriQA gold-passage QA supplies a Yoruba or Swahili question together with an English evidence passage; the model greedily generates an unconstrained English answer, evaluated by token F1 Ogundepo et al. (2023). The test split contains 253 Yoruba and 295 Swahili questions ($N = 548$ per model). Candidate construction remains fixed, but the preserving oracle now selects the highest-F1 transport only when it improves over identity. It is diagnostic and never available at inference.

Model	Base F1	Oracle Δ [95% CI]	N
Qwen3-8B	44.17	+8.06 [5.95,10.25]	548
Qwen3-4B	32.42	+6.50 [4.78,8.39]	548
Llama-3.1-8B	52.80	+4.00 [2.60,5.58]	548
Llama-3.2-3B	37.91	+3.38 [2.21,4.67]	548

Table 13: AfriQA free-generation candidate headroom. Positive intervals across four model sizes/families show that the latent transport library contains useful generated-answer counterfactuals beyond A–D scoring.

We separately fit sequence-level, label-free-at-test selectors using generation likelihood/entropy trajectories, grounding and copying features, base-to-candidate contrast, and candidate consensus. Training uses only 247 Yoruba training questions with out-of-fold calibration; test policies are then frozen for Yoruba and Swahili. The preregistered pooled gains are +0.00, +0.51, +0.86, and +0.13 F1 points for Qwen3-8B, Qwen3-4B, Llama-3.1-8B, and Llama-3.2-3B, respectively; all corresponding confidence intervals include zero. A post-hoc no-consensus Qwen3-4B ablation reaches

+1.16 [0.26,2.19], but its significance does not replicate on the two Llama models. Table 13 supports *candidate-transfer generalization*; learning a stable cross-model sequence-level repairability boundary remains open.

Protocol-primary selectors. For Qwen3-8B and Qwen3-4B, the primary policy is the full-feature task-native logistic selector. For the subsequent Llama confirmation, the primary policy is the no-consensus logistic variant, fixed before the Llama outputs were generated. The Qwen3-4B no-consensus ablation (+1.16 F1) motivated that confirmation and is not substituted for the original Qwen primary result. Across the four protocol-primary policies, the gains and 95% intervals are +0.00 [0.00,0.00], +0.51 [-0.37,1.41], +0.86 [-0.58,2.37], and +0.13 [-1.04,1.27]. The two Llama primary gains are smaller than their simple-reranker gains (+1.33 and +0.91).

L TIBETAN EXTENSION

The Tibetan setting contains 14,042 automatically translated English MMMLU questions per model. Candidate selection improves both 8B models on the full set. The translation is unofficial and lacks a human quality audit; the result tests an additional input language rather than establishing an official Tibetan benchmark score.

Model	Setting	Accuracy	Δ	Pres.
Qwen3-8B	Identity baseline	29.20	0.00	100.00
Qwen3-8B	REAL, pres95	30.59	+1.40	98.76
Qwen3-8B	REAL, best pres90	32.75	+3.55	94.46
Llama-3.1-8B	Identity baseline	28.93	0.00	100.00
Llama-3.1-8B	Direct $\mathcal{T}_{12}^{.1,.1}$	28.93	0.00	96.70
Llama-3.1-8B	REAL (A1/A2/A3)	31.43	+2.51	97.66

Table 14: Full-test Tibetan extension ($N = 14{,}042$ per model). Deltas compare each model with its own identity baseline; the Llama fixed action and selector share the intervention library.

For Qwen, the preservation-calibrated pres95 policy yields +1.40 points (98.76% preservation); a separately selected pres90 policy yields +3.55 points (94.46% preservation). For Llama, fixed A1/A2/A3 actions change accuracy by 0.00/-0.20/-0.07 points. The A1/A2/A3 selector instead makes 447 wrong-to-right changes and 95 right-to-wrong changes, yielding 352 net correct answers and +2.51 points. These contrasts isolate the value of selecting among unstable interventions on this translated language set.

M RESPONSIBLE USE

REAL measures answer repair and preservation in white-box benchmark evaluation. Deployment requires separate assessment of factuality, fairness, cultural validity, and performance under distribution shift. High-stakes applications require task-specific validation, community review, and monitoring beyond MMMLU choice scoring.

N ADDITIONAL CALIBRATION AND SCALE DIAGNOSTICS

Geometry sensitivity and model scale. The canonical PCA-256/remove-16/top-50/cluster-64 configuration is an operating point, not a claimed optimum. A 576-configuration Qwen3-4B diagnostic varies PCA dimension, removed language dimensions, prototype count, layer, and neighbor space. Average semantic Top-1 gains increase from 9.64 to 17.70 points as PCA dimension moves from 64 to 256; removing 0/16/32 language directions gives 12.71/17.70/17.98; and 10 versus 50 prototypes gives 17.29/17.70. A separate temperature sweep is stable around the canonical value. These are representation-level diagnostics, not 8B policy tuning. On a Qwen3-14B-AWQ $N = 50$ Yoruba sanity, a mild transport improves 28% to 30%, whereas a strong layer-12 transport falls to 16%. The alignment–readout mismatch persists at 14B, but full REAL generalization beyond 8B remains open.

Calibration efficiency. The verifier does not require the full 1,991-ID split to reveal a repair signal. With independently fitted policies and five fixed seeds, pooled Qwen3-8B matched-field gains at 50/100/250/500/1,991 calibration IDs are +6.70/+8.76/+8.82/+9.63/+9.91 points. The steepest improvement occurs by 100 IDs; additional data mainly close selection regret. These curves do not replace the strict headline protocol, but they argue against an explanation based on memorizing a large fraction of the benchmark.

O ALGORITHMIC POSITIONING

Table 15 compares decision contracts, not benchmark scores. The routing controls in the main paper hold the candidate library and test examples fixed; they do not reproduce every component or compute budget of the cited methods.

Method	Intervention and online decision	Relation to REAL
Activation patching (Dumas et al., 2025)	Controlled insertion of donor-prompt states to study concept/language representations.	Mechanistic analysis of representations; REAL selects repairs for unseen questions from a non-test bank.
CLAS (Pokharel et al., 2026)	Neuron-category rescaling and blending using multilingual activation statistics.	Constructs an intervention; REAL additionally scores executed candidates for repair and damage.
CAST (Lee et al., 2025)	A condition-vector gate triggers a behavior vector.	Contextual gating precedes the behavior change; REAL's final choice uses candidate outcomes as features, without their labels.
MERA (Hedström et al., 2025)	An error probe determines adaptive displacement; held-out calibration selects a threshold or abstains.	REAL selects among executed transports via repair/damage predictions instead of steering toward a probe-error threshold.
REAL	Base-side proposal, execution of k transports, then candidate-conditioned acceptance or identity.	Learns auxiliary predictors; no backbone weight updates or paired high-resource test queries.

Table 15: Method-level distinctions. "Candidate outcomes as features" means observable predictions and activation statistics, not gold correctness.

Translation and language choice. Translation provides a complementary route to multilingual performance: NLLB expands machine translation coverage across languages (NLLB Team, 2022), while language choice in chain-of-thought prompting can affect multilingual reasoning (Shi et al., 2023). These approaches motivate distinguishing access through the input language from access through internal representations. REAL retains the target-language prompt and uses a precomputed bank rather than translating each test question. Our NLLB pivot comparison measures this alternative input protocol; it does not establish that internal intervention dominates translation in general.

Activation intervention and parameter editing. Activation addition changes a model's computation by adding a steering vector at inference (Turner et al., 2023). ROME, MEMIT, and AlphaEdit instead modify model parameters to update factual associations (Meng et al., 2022; 2023; Fang et al., 2024). REAL operates on inference-time states with a frozen backbone and learns auxiliary proposal and verification components. Its contribution is the repair-aware choice among executed candidates, including identity, rather than a new factual knowledge-editing algorithm or the general idea of activation steering.

P Repair Utility and the Limits of Semantic Similarity

An exact decision identity. Let $B = \mathbb{1}[\hat{y}_0 = y]$ and $C = \mathbb{1}[\hat{y}_\eta = y]$. For every example, $C - B = \mathbb{1}[B = 0, C = 1] - \mathbb{1}[B = 1, C = 0] = G^+ - G^-$. Taking conditional expectations gives

$$\mathbb{E}[C - B \mid \phi] = \Pr(G^+ = 1 \mid \phi) - \Pr(G^- = 1 \mid \phi). \tag{14}$$

$\mathcal{R}_1$ targets conditional accuracy gain, and $\mathcal{R}_\lambda = \mathbb{E}[C-B \mid \phi]-(\lambda-1)\Pr(G^- = 1 \mid \phi)$ additionally penalizes damage for $\lambda > 1$. This algebra motivates the two prediction targets; it does not guarantee that their learned estimates remain accurate under distribution shift. Identity has zero realized gain and zero damage, but selecting identity only when a learned score is low does not confer a pointwise safety guarantee.

Why a similarity improvement alone is insufficient. A simple counterexample separates the two objectives. Let a semantic score be $S(z) = -\|Pz - a\|^2$, with $P = (1, 0)$ and anchor $a = 1$. Let the correct class be predicted when $z_2 > 0$. Moving from $z = (0, 1)$ to $z' = (1, -1)$ improves S from -1 to 0, yet changes a correct prediction to an incorrect one. Semantic improvement need not imply readout improvement without a relationship between the semantic projection and the decision boundary. This example establishes non-implication, not a claim that these coordinates describe the evaluated models' internal circuit. The empirical retrieval/accuracy diagnostics provide the corresponding observation in the models.