FAIR FACT-CHECKING: CLOSING THE CROSS-LINGUAL GAP IN LLM FACTUAL JUDGEMENT WITH ROSH

Anonymous authors
Paper under double-blind review

ABSTRACT

Misinformation on social media remains a critical problem, and more and more people settle it by asking a language model instead of a fact checker, about anything from a claim circulating on a platform to a plain question of fact. Whether models judge such claims reliably is still debated; whether they judge them equally well in every language people ask in has gone almost unasked. We test eight models from five families, 3B to 70B, on 1,500 claims that mix fact-checked news with encyclopedic facts and exist in identical form in eight languages. On every model English is judged better than the other languages, and the gap is widest on the smallest models, where Llama-3B on Arabic is no better than guessing. Existing remedies either retrain the model on more multilingual data or fit an unconstrained map between language representations, and neither asks whether the model already holds the answer and simply fails to say it. It largely does: a linear probe recovers the truth from the very activations the model fails to express. We propose RoSh, a per-language shift and rotation of the residual stream, computed in closed form at three layers, with no training and no weight modified. It improves every model and closes 75% of the gap on average, and it helps most where the model was worst: the two smallest models end close to English, Arabic on Llama-3B goes from chance to nearly the English level, and a fifth fewer of the claims answered correctly in English are lost in translation. What remains is no longer a read-out failure: afterwards the head recovers at least as much of what is encoded outside English as it does in English. An unconstrained map fitted on the same pairs falls below the untouched baseline, so the orthogonality constraint is doing the work, and every model clears a scrambled-correspondence control and ten further controls. On the two benchmarks of the closest inference-time method, latent-space intervention, run with its own data and metric code, RoSh's gains are five to thirteen times larger.

1 INTRODUCTION

Social media platforms have been criticised for years for letting false stories spread unchecked, and the damage does not stay online. It has been traced into elections (Allcott & Gentzkow, 2017) and, during the pandemic, into public health. Because false stories also diffuse faster and reach wider audiences than true ones, fact-checks typically arrive long after public attention has shifted (Vosoughi et al., 2018). In response, platforms partnered with independent fact-checking organisations such as Snopes, PolitiFact, and Full Fact to suppress flagged content in user feeds. However, this approach assumes that users actively seek out verification, and increasingly they do not. Across 48 markets, 10% of adults now use an AI chatbot for news every week, up from 7% a year earlier, and 16% among under-35s (Newman et al., 2026).

A chatbot asked about a claim becomes the fact checker, and unlike a fact-checking desk it answers in whatever language it is asked: one model serves all 48 of those markets, in many languages. That makes a new question pressing: is the model's verdict on a claim the same whichever language the claim arrives in? To find out, we evaluated eight language models on a straightforward task: given a factual claim with a known truth value, we queried the same model in English and in another target language to test whether its verdict remained consistent. For Llama-3B, consistency broke

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.

down sharply. The model achieved an AUROC of 0.857 in English but dropped to 0.523 in Arabic, marginally better than random guessing. Similarly, Mistral-7B scored 0.834 in English compared to 0.646 in Portuguese. Across an evaluation of eight models spanning five architecture families (3B to 70B parameters) and $1{,}500$ claims held identical across eight languages, English consistently outperforms all other languages by an average of $+0.046$ AUROC, a gap that widens to $+0.116$ and $+0.130$ on the two smallest models. Because every claim is evaluated symmetrically across all languages, this disparity cannot be attributed to differences in prompt difficulty. The quality of the verdict depends on the language you happen to speak.

Two things make that worse than a benchmark gap. The users it fails have the fewest alternatives, since where the major services are restricted or unaffordable people fall back on regional systems trained on far less data, and nothing in the answer reveals the problem. The judge being replaced was no more reliable: given stories fact-checkers had already settled, ordinary readers produced verdicts tracking their politics rather than the evidence (Babaei et al., 2021). A model may have a perception of its own, and ours depends on language. It is also asked about everything, so our claims mix fact-checked news with encyclopedic facts, and the gap is as wide on one half as on the other.

Prior literature extensively documents this disparity across cross-lingual verification and dialectal fairness (Lucas et al., 2026b;a; Qi et al., 2023; Han et al., 2026; Liu et al., 2025). However, existing mitigations either require resource-intensive retraining that is impractical for downstream deployment, or rely on unconstrained inference-time mappings (Ghorbanpour et al., 2026; Wang et al., 2025) that are free to rescale and project, and so to discard information as well as re-orient it, which makes a change in score hard to read as evidence of what the underlying model knew. Crucially, both approaches overlook a fundamental question: before concluding that a model lacks factual knowledge in a target language, one must first determine whether it possesses the internal knowledge yet fails to express it.

Research question. *When a model judges a claim worse outside English, how much of the gap is knowledge it lacks, and how much is knowledge it holds but fails to read out? And can the second part be recovered at inference time, without changing a single weight?*

Both parts are real, and of similar size. Linear probing on intermediate activations recovers the truth in every language we test, including those where the model's own verdict falls to chance, and at those layers the probe scores far above the model's own output head. Comparing probe against head within each language splits the gap in two: 0.026 is a shortfall in what the model encodes, and 0.021 is the head reading non-English less efficiently than English (Section 5.1). Only the second part can be repaired without changing the pre-trained parameters. We propose ROSH, a per-language shift and rotation applied to the residual stream at selected intermediate layers, computed as the closed-form orthogonal Procrustes solution (Schönemann, 1966) on unlabelled parallel claims and applied at inference, with no gradient taken and no weight modified. It closes 75% of the gap on average and removes it entirely on Qwen-32B and Llama-70B, lifting accuracy from 0.727 to 0.757 and cutting the share of English-correct claims lost in translation by a fifth, from 15.7% to 12.3%. The largest gains land where the gap was worst: Arabic on Llama-3B recovers from 0.523 to 0.841, Portuguese on Mistral-7B from 0.646 to 0.896. What remains is no longer a read-out failure: afterwards the head recovers at least as much of what is encoded outside English as it does in English.

The gains carry beyond our dataset: on the two benchmarks of the closest inference-time method, latent-space intervention (Ghorbanpour et al., 2026), run with its own data and evaluation scripts, ROSH's gains are five to thirteen times larger, and neither benchmark contains news claims (Table 11).

Contributions.

- **The verdict depends on the language.** English leads by $+0.046$ AUROC on average and by 0.334 on Arabic for Llama-3B, over $1{,}500$ claims that appear in every language.
- **The gap has two parts, and only one is read-out.** A probe splits it into 0.026 of representational shortfall and 0.021 of read-out loss.
- **ROSH removes the read-out part entirely.** A closed-form, label-free shift and rotation at three layers closes 75% of the gap, and eleven controls, including an unconstrained map that falls below the untouched baseline, rule out generic perturbation.

2 Related Work

Who receives a good judgement. Unequal judgements predate language models: readers of fact-checked news trusted claims that matched their politics and distrusted the rest (Babaei et al., 2021); we ask the same question with a model as the judge and the input language in place of politics. Translating claims into English does not settle it, since translation-based pipelines carry disparities of their own (Singhal et al., 2024). The disparity itself is well measured: BMLAMA and RankC (Qi et al., 2023) and MuBench, across 61 languages (Han et al., 2026), test whether a model answers the same fact alike across languages, Liu et al. (2025) tie consistency to entity alignment and improve it by English subject substitution, and Piratla et al. (2025) trace part of the gap to response variance and reduce it by inference-time ensembling.

Where knowledge sits. Llama-2's intermediate layers favour English equivalents of the intended output (Wendler et al., 2024), cross-lingual factual failures arise both in recall and in converting the answer to the target language (Lu et al., 2025), and factual associations sit in middle-layer MLP modules that can be edited directly (Meng et al., 2022). Truth in particular can be read from hidden states, without supervision by enforcing consistency across paired answers (Burns et al., 2023) or with a small supervised classifier (Azaria & Mitchell, 2023); true and false statements separate geometrically, though probe directions of similar accuracy can differ when used to intervene (Marks & Tegmark, 2023), and a general truth direction can be told apart from a polarity-sensitive one that reverses under negation (Bürger et al., 2024). Our probes fit this picture: the truth is recoverable at mid-depth in every language, and what varies is whether it reaches the output head.

Aligning representations to close the gap. The closest work intervenes on representations at inference. INCLINE fits layer-wise alignment matrices by least squares on parallel sentences (Wang et al., 2025); latent-space intervention trains layer-wise autoencoders that align languages in a shared latent space, improving cross-lingual consistency while largely preserving accuracy (Ghorbanpour et al., 2026); steering vectors shift target-language activations toward English at a single layer (Mahmoud, 2026); and Agarwal et al. (2025) instead fine-tune on batches of semantically equivalent multilingual examples.

What is new. Every intervention above fits an unconstrained map or vector, free to rescale and project, and so to discard information as well as re-orient it; a change in score then mixes the two effects. RoSh is an affine isometry, distance-preserving and invertible, so it provably loses nothing (Proposition E.1). The estimator itself is old: Schönemann (1966) solved the orthogonal Procrustes problem in closed form, Smith et al. (2017) showed that the orthogonal constraint improves bilingual word-embedding alignment, and Conneau et al. (2018) made it unsupervised. What is new is that the same relationship holds inside a running model, between languages, in the residual stream, at inference time.

3 Method: Rotation–Shift Alignment (RoSh)

A transformer's verdict on a factual claim is read from a single direction in its residual stream. We observe that the truth is linearly readable in every language we test, while the read-out head recovers much less of it outside English. Table 1 shows it for two models. On Llama-3B in Arabic the model's own verdict is at chance, 0.523 AUROC, yet a linear probe on the same activations reaches 0.871; in every language the probe stays within 0.06 of its English value, while the read-out falls up to 0.334 behind. Across all eight models the probe reaches 0.910 outside English where the head reaches 0.823 (Table 10), and Section 5.1 splits the gap into what the model fails to encode and what it fails to read out. Figure 2 shows the same claims from the read-out's own perspective, before and after the correction. RoSh acts on that directly, applying a closed-form orthogonal rotation and a mean shift to the hidden states, per language and per layer, without changing any model weight. Figure 1 gives the picture; Appendix J shows how the lookup table behind it is built.

3.1 Problem Setup

Let $\mathcal{M}$ be a frozen transformer of depth L and residual width d. For a claim x in language ℓ, write $\mathbf{h}_l^\ell(x) \in \mathbb{R}^d$ for the residual-stream activation at layer l and the answer-token position. We are given a parallel corpus $\mathcal{D} = \{(x_i^{\text{en}}, x_i^\ell)\}_{i=1}^n$ of n claims expressed in both English and a target language.

Table 1: The truth is encoded even where the verdict fails: linear probe against the model's own read-out, held out, at the first selected layer (for English, the layer most often selected first). Read-out values are those of Table 3.

Model		EN	PT	PL	DE	AR	ZH	RU	TR
Llama-3B	probe	0.922	0.908	0.906	0.908	0.871	0.875	0.888	0.884
	read-out	0.857	0.787	0.794	0.851	0.523	0.772	0.787	0.669
Gemma-9B	probe	0.935	0.928	0.936	0.937	0.912	0.899	0.927	0.920
	read-out	0.878	0.859	0.862	0.862	0.844	0.762	0.857	0.820

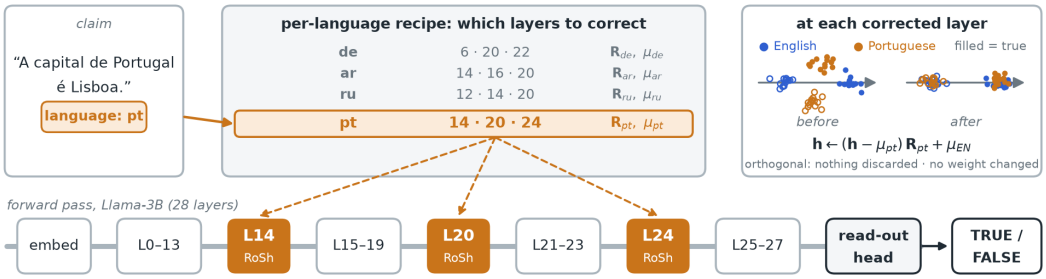

Figure 1: **Language-conditioned correction at inference.** The language of the incoming claim selects a recipe: the layers to correct and the rotation $\mathbf{R}_\ell$ and shift to apply at each. Shown is Llama-3B with the Portuguese recipe. Only residual-stream activations at the listed layers are transformed; the model's weights are shared across all languages and never modified. The inset is a schematic of what the correction does at one layer, not a measurement.

The hypothesis is that the model internally represents the truth value of x_i in every language, but the subspace carrying this signal in language ℓ is rotated relative to the English subspace that the read-out head was implicitly trained to decode. Formally, there exists an orthogonal matrix $\mathbf{R}_l^\ell \in \mathbb{R}^{d\times d}$ such that, after the affine map

$$\boxed{\mathbf{h}_l^\ell \longleftarrow \left(\mathbf{h}_l^\ell - \boldsymbol{\mu}_l^\ell\right)\mathbf{R}_l^\ell + \boldsymbol{\mu}_l^{\text{en}}} \tag{1}$$

the transformed activations are approximately aligned with their English counterparts, and the model's own read-out head can now decode the correct factual judgement.

A toy example makes Eq. 1 concrete. Take $d = 3$ and a read-out that scores a claim by its first coordinate. In English two true claims sit at $(2, \pm 1, 0)$ and two false ones at $(-2, \pm 1, 0)$, so they score $+2$ and -2 and separate perfectly. In Arabic the same claims sit at $(4, 7, 0)$ and $(6, 7, 0)$ (true) and at $(4, 3, 0)$ and $(6, 3, 0)$ (false). Truth is still perfectly separated, but along the second coordinate, which the read-out ignores, so true and false claims both score 4 or 6 and the verdict is a coin toss. Eq. 1 subtracts the Arabic mean $(5, 5, 0)$, rotates by $-90°$ and adds the English mean $(0, 0, 0)$, taking $(4, 7, 0)$ to $(-1, 2, 0)$ and then to $(2, 1, 0)$, exactly its English counterpart. The other three claims land on theirs, and the scores return to ± 2. Figure 2 shows the same effect on a larger synthetic cloud, and Appendix G derives $\mathbf{R}$ for this example.

3.2 Fitting the Transform via Orthogonal Procrustes

Given matched English and target-language activations, we aim to find the rotation, $\mathbf{R}_l^\ell$, that makes the target-language representation look as similar as possible to the English one. In particular, we propose to recover $\mathbf{R}_l^\ell$ in closed form using the *orthogonal Procrustes problem* (Schönemann, 1966). Given the parallel activations, we first centre each language cloud:

$$\boldsymbol{\mu}_l^\ell = \frac{1}{n}\sum_{i=1}^{n} \mathbf{h}_l^\ell(x_i^\ell), \qquad \boldsymbol{\mu}_l^{\text{en}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{h}_l^{\text{en}}(x_i^{\text{en}}), \tag{2}$$

$$\mathbf{A} = \left[\mathbf{h}_l^\ell(x_i^\ell) - \boldsymbol{\mu}_l^\ell\right]_{i=1}^n \in \mathbb{R}^{n\times d}, \qquad \mathbf{B} = \left[\mathbf{h}_l^{\text{en}}(x_i^{\text{en}}) - \boldsymbol{\mu}_l^{\text{en}}\right]_{i=1}^n \in \mathbb{R}^{n\times d}. \tag{3}$$

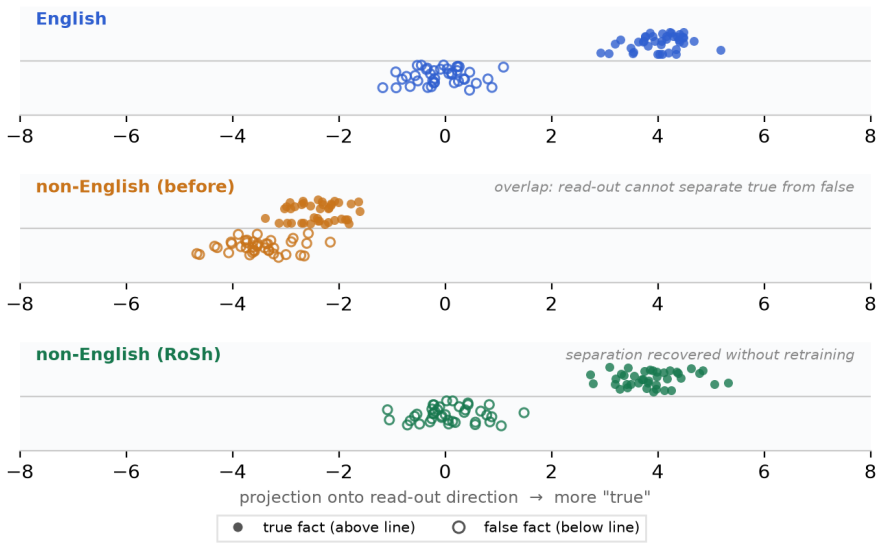

Figure 2: **What the read-out head sees.** Each dot is one claim projected onto the read-out direction, true above the centre line and false below. English (top) separates cleanly; non-English before correction (middle) overlaps almost completely; after RoSh (bottom) the separation matches English, without retraining. Same synthetic geometry as the inset of Figure 1.

The rotation that best maps $\mathbf{A}$ onto $\mathbf{B}$ is the one minimising the Frobenius norm of the residual:

$$\mathbf{R}_l^\ell = \underset{\mathbf{R}\,:\,\mathbf{R}^\top\mathbf{R}=\mathbf{I}}{\arg\min} \left\|\mathbf{A}\,\mathbf{R} - \mathbf{B}\right\|_F. \tag{4}$$

A solution is given by the SVD of the cross-covariance matrix:

$$\mathbf{M} = \mathbf{A}^\top\mathbf{B} \;=\; \mathbf{U}\,\mathbf{\Sigma}\,\mathbf{V}^\top, \qquad \mathbf{R}_l^\ell = \mathbf{U}\,\mathbf{V}^\top. \tag{5}$$

The solution is not unique, and this does not matter. With $n = 750$ claims and $d \geq 3{,}072$, $\mathbf{M}$ has rank below d, so $\mathbf{R}_l^\ell$ is determined only on the directions the fitting claims span and is free on the rest (Proposition C.1). Those directions carry at most 1.7% of a held-out activation, and switching to the canonical solution, the optimal $\mathbf{R}$ closest to the identity, moves held-out AUROC by less than 0.001 (Appendix C); the tables use the standard SVD solution.

The full map in Eq. 1 is an affine isometry, distance-preserving and invertible, so it keeps all the information an activation carries about the claim's truth (Proposition E.1): an improvement cannot come from discarding a distracting part of the representation, only from a change of frame.

3.3 The Two Components

Equation 1 composes a rotation about the non-English centroid with a translation onto the English one, and each can be run alone:

$$\text{shift only: } \mathbf{h}_l^\ell \leftarrow \mathbf{h}_l^\ell - \boldsymbol{\mu}_l^\ell + \boldsymbol{\mu}_l^{\text{en}}, \qquad \text{rotation only: } \mathbf{h}_l^\ell \leftarrow \left(\mathbf{h}_l^\ell - \boldsymbol{\mu}_l^\ell\right)\mathbf{R}_l^\ell + \boldsymbol{\mu}_l^\ell. \tag{6}$$

A translation cannot change an orientation, so under a linear read-out a pure shift would leave any ranking metric untouched; the corrected state, however, passes through further non-linear blocks, so the shift does contribute: alone it recovers 43% of the full gain against 74% for the rotation (the two are not additive).

3.4 Greedy Layer Selection

Not every layer carries a misalignment worth correcting, and the useful ones are not independent: once a layer is corrected, the activations arriving at every later layer are different, so the best second

layer depends on which one was chosen first. Searching all combinations is prohibitive, so we use greedy forward selection on the fit split alone. That is, we greedily build a small sequence of intervention layers, repeatedly choosing the layer that gives the largest additional gain after the previously selected corrections are already in place.

For each language we sweep every s-th layer between 15% and 97% of depth, with $s = 2$, or 3 for the models of 64 to 95 layers, giving 12 to 27 candidates. We start at 15% because early runs showed that the lowest layers do not yet carry enough truth for a correction to help, while later layers do: in the lowest 30% of depth a probe reads the truth outside English at 0.76 AUROC against about 0.90 higher up, and only 2 of the 56 best single-layer corrections fall there. The fit split is halved into two inner folds of 375 claims: at each candidate we fit $\mathbf{R}_l^\ell$ on one and score the corrected model's AUROC on the other. The best layer is selected and its transform refitted on all 750 fitting claims. The sweep is then repeated with that correction in place, up to $D = 3$ layers, each rotation fitted on activations as they leave the previous correction. Greedy search keeps this affordable: on Mistral-7B it takes $14 + 13 + 12 = 39$ fits, where every three-layer combination would take $\binom{14}{3} = 364$. Ranking by AUROC uses labels on the fitting split; the transform itself does not. The output is a per-language *recipe*: at most three layers, each with its rotation $\mathbf{R}_{l_k}^\ell$ and centroid pair $(\boldsymbol{\mu}_{l_k}^\ell, \boldsymbol{\mu}_{l_k}^{\mathrm{en}})$.

Inference. At test time no additional data is needed: the language selects its recipe, the forward pass runs as normal with the residual stream at the answer position transformed by Eq. 1 at each listed layer, and the verdict is read from the unmodified output head. The cost is one $d \times d$ multiply at up to three layers, on a single token. Algorithm 1 in Appendix K gives the procedure in full.

3.5 Controls

Eleven controls test that the transform does what we claim. Each runs on identical data at the same layers and breaks exactly one assumption: whether each component is needed, whether the claim-level pairing matters, whether any perturbation of matched size would serve, and whether the map is specific to its layer and direction. A control passes when it scores below the full method (Appendix L).

4 Data and Experimental Setup

Claims. We evaluate on 1,500 factual claims[1] with known truth values, balanced between true and false, each in identical form in eight languages: English, Portuguese, Polish, German, Arabic, Chinese, Russian and Turkish, spanning four scripts and four families. The set mixes claims of the kind fact-checkers adjudicate with encyclopedic facts, and the non-English versions are taken as released from BMLAMA-53 (Qi et al., 2023); Appendix B lists the claim types and gives examples. The claims are split once into 750 for fitting and 750 held out: every transform, layer choice and threshold is estimated on the fit half, the held-out half is scored exactly once, and a claim falls on the same side of the split in every language (Appendix A).

Models. We use eight instruction-tuned models from five families, from 3B to 70B parameters and with residual widths from 3,072 to 8,192 (Table 4). All are frozen and none is fine-tuned. They run in half precision (bf16 for the two Gemma models), except Llama-70B and DeepSeek-67B, which run in 4-bit. Every prompt is round-tripped through its tokeniser before any activation is collected (Appendix M).

Scoring. Each claim is presented in its own language with a fixed template that asks for a single verdict token (Appendix B). The score of a claim is the logit difference between the TRUE and FALSE tokens at the answer position,

$$s(x) = \mathbf{u}^\top \mathbf{h}_L(x), \qquad \mathbf{u} = \mathbf{w}_{\mathrm{TRUE}} - \mathbf{w}_{\mathrm{FALSE}}, \tag{7}$$

read from the unmodified head after RoSh has changed $\mathbf{h}_l$ at the selected layers.

[1] Code and seeds: https://anonymous.4open.science/r/fair-factchecking-40E8.

Metrics. We report held-out AUROC of $s(x)$ as the primary metric, the language gap (English minus the mean of the other seven languages) and the share of it closed, accuracy, directional consistency, and a best-layer linear probe; Appendix A defines each.

External benchmarks. To compare with latent-space intervention (Ghorbanpour et al., 2026) on its own ground, we also run RoSH on its two benchmarks, KLAR and mParaRel, with three 8B models (Qwen3-8B, Llama-3.1-8B and Aya-8B), using their data and metric code (Appendix H).

5 Results

Every transform is fitted on 750 claims and scored on the 750 it never saw. Accuracy and consistency are read at each language's median operating point, fitted on the training half.

Table 2: Non-English mean, held out. Gap is English minus non-English, and gap closed is $(\text{gap}_{\text{before}} - \text{gap}_{\text{after}})/\text{gap}_{\text{before}}$ (Appendix A). Directional consistency is the share of claims the model gets right in English that it also gets right in the other language. The Mean row averages the eight model rows; its gap closed is the mean of the per-model values, with values above 100% counted as 100%.

		non-EN		gap		gap		directional
Model	EN	before	after	before	after	closed	accuracy	consistency
Llama-3B	0.857	0.740	0.829	+0.116	+0.028	76%	0.671 → 0.736	0.750 → 0.841
Mistral-7B	0.834	0.704	0.809	+0.130	+0.025	80%	0.649 → 0.726	0.715 → 0.806
Gemma-9B	0.878	0.838	0.865	+0.040	+0.012	69%	0.740 → 0.759	0.879 → 0.905
Mistral-24B	0.888	0.866	0.878	+0.022	+0.011	53%	0.756 → 0.767	0.883 → 0.898
Gemma-27B	0.860	0.833	0.844	+0.027	+0.016	42%	0.729 → 0.746	0.839 → 0.851
Qwen-32B	0.885	0.879	0.889	+0.006	−0.004	>100%	0.770 → 0.778	0.903 → 0.911
Llama-70B	0.879	0.865	0.883	+0.014	−0.003	>100%	0.752 → 0.780	0.895 → 0.915
DeepSeek-67B	0.875	0.861	0.872	+0.014	+0.003	80%	0.748 → 0.762	0.882 → 0.892
Mean	0.870	0.823	0.859	+0.046	+0.011	75%	0.727 → 0.757	0.843 → 0.877

The gap closes on every model, by 75% on average, and Qwen-32B and Llama-70B end with non-English slightly ahead of English. Where the gap was widest the correction is largest: Llama-3B and Mistral-7B start 0.116 and 0.130 behind and end within 0.03. Directional consistency rises from 0.843 to 0.877 (of the claims a model gets right in English, the share it gets wrong in translation falls from 15.7% to 12.3%), so the claims English handles correctly survive translation more often than before. Every tokeniser was verified in all eight languages before any run (Appendix M).

Table 3: Baseline and corrected AUROC per language, held out. Shading shows the change: green for a gain, darker the larger it is; red for a drop; grey for no change.

Model	PT	PL	DE	AR	ZH	RU	TR
Llama-3B	0.787 → 0.840	0.794 → 0.829	0.851 → 0.856	0.523 → 0.841	0.772 → 0.798	0.787 → 0.838	0.669 → 0.798
Mistral-7B	0.646 → 0.896	0.802 → 0.878	0.807 → 0.847	0.585 → 0.674	0.659 → 0.702	0.791 → 0.869	0.639 → 0.795
Gemma-9B	0.859 → 0.872	0.862 → 0.884	0.862 → 0.871	0.844 → 0.853	0.762 → 0.845	0.857 → 0.868	0.820 → 0.865
Mistral-24B	0.874 → 0.874	0.875 → 0.890	0.893 → 0.897	0.859 → 0.892	0.824 → 0.828	0.889 → 0.902	0.846 → 0.860
Gemma-27B	0.856 → 0.855	0.840 → 0.860	0.847 → 0.863	0.826 → 0.841	0.790 → 0.787	0.830 → 0.863	0.842 → 0.841
Qwen-32B	0.879 → 0.882	0.888 → 0.898	0.891 → 0.900	0.877 → 0.893	0.862 → 0.875	0.892 → 0.908	0.866 → 0.868
Llama-70B	0.882 → 0.892	0.884 → 0.907	0.885 → 0.896	0.851 → 0.862	0.820 → 0.854	0.872 → 0.904	0.861 → 0.866
DeepSeek-67B	0.865 → 0.870	0.880 → 0.896	0.871 → 0.888	0.642 → 0.658	0.855 → 0.865	0.876 → 0.886	0.820 → 0.815

Arabic on Llama-3B moves from 0.523, which is a coin toss, to 0.841, and Portuguese on Mistral-7B from 0.646 to 0.896 (Table 3). Of the 56 model-language cells, 51 improve, one is unchanged and the largest regression is 0.005. Most languages that stay flat were already close to English, so there was little left to close; Gemma-27B Chinese (0.790 to 0.787) is the exception.

Beyond our data. On the two benchmarks of the closest inference-time method, latent-space intervention (LSI), run with its own data and metric code on three 8B models, RoSH raises KLAR accuracy by 5.63 points and agreement with English by 5.48, against the 0.44 and 1.02 that Ghorbanpour et al. (2026) report for LSI, and raises mParaRel agreement on Aya by 2.24 against 0.27

(Table 11). Neither benchmark contains news claims, and on mParaRel no translation system its templates come from shows a significant loss (Appendix I), so the correction is not specific to our data.

5.1 What RoSH fixes, and what remains

Comparing the model's read-out with a probe on the same activations splits the gap in two (Appendix F, Table 10). The non-English representation is genuinely weaker, since a probe reaches 0.910 against 0.935 in English, a shortfall of 0.026. On top of that the head reads non-English less efficiently, losing 0.086 AUROC relative to the probe outside English against 0.066 within it, a further 0.021. The two account for the $+0.046$ we measure.

RoSH acts on the second component and removes it. After correction the head loses 0.051 relative to the probe in non-English, below the 0.066 it loses in English on English's own representation, and it is at or below the English level on every model except Mistral-24B, where it exceeds it by 0.002. So the residual gap of 0.011 is smaller than the representational shortfall of 0.026: the corrected head now reads non-English *more* efficiently than English, by 0.015, which offsets part of what the representation lacks.

Our analysis also turns up a finding that has nothing to do with language: even in English the head leaves 0.066 AUROC on the table relative to a probe on its own intermediate activations. Read-out inefficiency therefore appears to be a general property of the model, while non-English languages suffer an additional read-out penalty that RoSH specifically removes.

5.2 Selected layers

The search settles on three layers for every model, mostly in the upper middle of the network but often reaching early layers, and the stacks differ by language within a model: on Llama-3B, Portuguese takes 14, 20 and 24 and Arabic 16, 20 and 14. The recipe is therefore fitted per language (Appendix N).

5.3 Controls

A gain on held-out data does not establish that the transform does what we claim, so we ran the eleven arms of Section 3.5, each fitted and applied exactly as the method is. Averaged over eight models and seven languages, RoSH reaches 0.853 against an untouched baseline of 0.820. The eight arms scored on non-English all land below 0.853, and the three scored on English land below the English baseline of 0.870, which is the comparison they call for. The informative ones are close together in construction and far apart in score. Scrambling the item pairing before fitting, which preserves both clouds and destroys only which claim matches which, falls to 0.621; a random rotation of matched size falls to 0.658; an unconstrained linear map on the same pairs reaches 0.786, below doing nothing at all. Fitting English against itself returns 0.870 and moves the score by 0.000, which is the null the isometry argument requires. Per model the fitted map beats all three scrambled runs, with z between $+1.83$ and $+8.13$. An earlier six-model version of this study used a single language-direction steering vector, which failed this same control on four of the six, and was abandoned for that reason. Tables 14 and 15 in Appendix O give every arm.

6 Discussion

The gap is partly a read-out failure. A linear probe separates true from false in every language, including where the model's own verdict is near chance, so the content survives in the residual stream. Of the 0.046 gap, 0.021 is the head failing to reach that content, which RoSH removes, and 0.026 is a weaker representation, which no isometry can repair. Because the transform is invertible (Proposition E.1), an improvement can only mean the head was failing to extract something already present.

The constraint carries the result. The same pairs fitted without the orthogonality constraint give 0.786, below the untouched baseline of 0.820, against 0.853 for the constrained solution: with

750 pairs in thousands of dimensions the least-squares map is ill-conditioned and distorts held-out activations (Appendix D). Orthogonality removes that failure by construction and is what licenses the reading above.

Relation to weight-space methods. We claim no impossibility result: fine-tuning could in principle reach a model that behaves similarly. What we claim is that the correction can be *found* in closed form, from unlabelled parallel claims, in one SVD, with no gradient step and no labelled data in the target language, which is precisely the resource the affected communities lack. Nor is it a weight edit: it acts only at the answer position, differs per language and includes a shift, so it cannot be folded into the checkpoint. Two guarantees follow that have no counterpart in weight space. English is left exactly unchanged (Proposition E.4; the null is 0.000 in Table 15), and since only one language's recipe is active in any forward pass, improving one language cannot disturb another.

What this means for deployment. The transform is a per-language matrix and a per-language vector at three layers, estimated from 750 parallel claims, whose truth labels are used only to choose the layers, and applied at one token position during inference. It requires no gradient, no labelled target-language data, and no change to the shipped checkpoint, which places it within reach of whoever deploys a model rather than only whoever trained it. That matters for the specific unfairness we measure, since the languages with the largest gaps are also those with the least training data, the fewest labelled resources and the weakest locally available systems.

6.1 Limitations

Eight languages and four scripts show the effect without characterising it across the long tail, where the disparity is likely worst, and the non-English claims, taken from BMLAMA-53 (Qi et al., 2023), were translated rather than written natively. On mParaRel, the second benchmark of Ghorbanpour et al. (2026), whose templates come from five translation systems, the correction survives a split by system (Appendix I). All results use one verdict-token template, so we cannot speak to chain-of-thought or open-ended answers. The method needs claims paired across languages and does not apply where no parallel corpus exists. Truth labels do not fit the transform but do enter through layer selection, so the label-free claim covers the transform rather than the whole recipe. And the diagnosis rests on probes: we do not identify which part of the head fails, or what puts non-English representations in a different frame.

7 Conclusion

Ask a model to judge the same claim in English and then in another language and the two verdicts often disagree. Across eight models from five families and $1{,}500$ claims rendered identically in eight languages, English leads by $+0.046$ AUROC on average and by 0.334 on Arabic for Llama-3B, where the model drops to 0.523 and is guessing. Every claim appears in every language, so this is not a matter of some questions being harder.

Part of what the model needs is already there. A linear probe recovers the truth from mid-network activations even where the model's own verdict sits at chance, and separating the two shows that nearly half of the gap is the read-out failing to reach what is encoded, the rest being a genuinely weaker representation. RoSh targets the first part, with a per-language shift and rotation of the residual stream at three layers, computed in closed form from unlabelled parallel claims, no gradient taken and no weight modified. It closes 75% of the gap on average and removes it on two models, lifts accuracy from 0.727 to 0.757, and cuts the share of English-correct claims lost in translation by a fifth, from 15.7% to 12.3%. The same recipe transfers to two benchmarks built by other authors, where its gains are five to thirteen times those of the closest inference-time method, on that method's own data and metric code.

An unconstrained map on the same pairs falls below doing nothing at all. What is left is to reach the long tail of languages, where the disparity is likely largest and parallel data scarcest, and to find what puts non-English representations in a different frame in the first place.

AI USE STATEMENT

In this work, we used generative AI tools (a large language model assistant) for editing and condensing the manuscript text, drafting and checking LaTeX, writing experiment, analysis and plotting code, and cross-checking the numbers reported in the text against our result files. We have not used generative AI tools to produce any experimental result: every number in the paper is computed by our code from the models' outputs, and every AI-drafted script was run and checked by the authors. We have reviewed all AI-assisted work and take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Amit Agarwal, Hansa Meghwani, Hitesh Laxmichand Patel, Tao Sheng, Sujith Ravi, and Dan Roth. Aligning llms for multilingual consistency in enterprise applications. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 117–137, 2025.

Hunt Allcott and Matthew Gentzkow. Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2):211–236, 2017. doi: 10.1257/jep.31.2.211.

Amos Azaria and Tom Mitchell. The internal state of an LLM knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023.

Mahmoudreza Babaei, Juhi Kulshrestha, Abhijnan Chakraborty, Elissa M Redmiles, Meeyoung Cha, and Krishna P Gummadi. Analyzing biases in perception of truth in news stories and their implications for fact checking. *IEEE Transactions on Computational Social Systems*, 9(3):839–850, 2021.

Lennart Bürger et al. Truth is universal: Robust detection of lies in LLMs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *International Conference on Learning Representations (ICLR)*, 2023.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. In *International Conference on Learning Representations (ICLR)*, 2018.

Faeze Ghorbanpour, Constanza Fierro, Alexander Fraser, and Anders Søgaard. Latent-space intervention for cross-lingual factual consistency: Consistency improvements without accuracy drops. *arXiv preprint arXiv:2608.28860*, 2026. doi: 10.48550/arXiv.2608.28860. URL https://arxiv.org/abs/2608.28860. To appear at EMNLP 2026.

Wenhan Han, Yifan Zhang, Zhixun Chen, Mykola Pechenizkiy, Meng Fang, Yin Zheng, et al. Mubench: Assessment of multilingual capabilities of large language models across 61 languages. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 16163–16192, 2026.

Yihong Liu, Mingyang Wang, François Yvon, and Hinrich Schütze. On the entity-level alignment in crosslingual consistency. *arXiv preprint arXiv:2510.10280*, 2025.

Meng Lu, Ruochen Zhang, Carsten Eickhoff, and Ellie Pavlick. Paths not taken: Understanding and mending the multilingual factual recall pipeline. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 15077–15107, 2025.

Jason Lucas, Matt Murtagh-White, Ali Al-Lawati, Uchendu Uchendu, Adaku Uchendu, and Dongwon Lee. DIA-HARM: Dialectal disparities in harmful content detection across 50 english dialects. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (ACL)*, San Diego, CA, 2026a.